SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
COMPLEX NETWORKS
SCIENTIFIC DATA
COMPUTATIONAL SCIENCE
APPLIED PHYSICS

Received
20 May 2014

Accepted
9 September 2014

Published
29 September 2014

Correspondence and requests for materials should be addressed to L.Š. (lovro.subelj@fri.uni-lj.si)

Network-based statistical comparison of citation topology of bibliographic databases

Lovro Šubelj[1], Dalibor Fiala[2] & Marko Bajec[1]

[1]University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, SI-1000 Ljubljana, Slovenia, [2]University of West Bohemia, Faculty of Applied Sciences, Univerzitní 8, CZ-30614 Plzeň, Czech Republic.

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. Here we compare the topological consistency of citation networks extracted from six popular bibliographic databases including *Web of Science*, *CiteSeer* and *arXiv.org*. The networks are assessed through a rich set of local and global graph statistics. We first reveal statistically significant inconsistencies between some of the databases with respect to individual statistics. For example, the introduced field bow-tie decomposition of *DBLP Computer Science Bibliography* 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 Computer Science Bibliography* is the least consistent with the rest, while, not surprisingly, *Web of Science* is significantly more reliable from the perspective of consistency. This work can serve either as a reference for scholars in bibliometrics and scientometrics or a scientific evaluation guideline for governments and research agencies.

Bibliographic databases range from expensive hand-curated professional solutions like *Web of Science* and *Scopus* to preprint repositories[1], public servers[2] and automated services that collect freely accessible manuscripts from the Web[3,4]. These provide the basis for scientific research, where new knowledge is derived from the existing, while also the main source of its evaluation. Undoubtedly, the number of citations a paper receives is still considered to be the main indicator of its importance or relevance[5,6]. However, the probability distribution of scientific citations has been shown to follow a wide range of different forms including power-law[7], shifted power-law[8], stretched exponential[9], log-normal[10], Tsallis[11], and modified Bessel[12], to name just a few. Although some methods used in these studies might be questionable, more importantly, they are based on different bibliographic data. In fact, the content and structure of modern bibliographic databases differ substantially, while there exist only informal notions on their reliability.

One way to assess the databases is simply by the amount of literature they cover. *Web of Science* spans over 100 years and includes several dozens of millions of publication records[13,14], an extent similar to that of *Scopus*, which, however, came into existence only some ten years ago. On the other hand, the preprint repository *arXiv.org*[1] and the digital library *DBLP Computer Science Bibliography*[2] both date back to 1990s and include only millions of publications or publication records. The coverage of different bibliographic databases has else been investigated by various scholars[14–17], while others have analyzed also their temporal evolution[1,18], available features[15,19], data acquisition and maintenance methodology[14,20], and the use within a typical scientific workflow[21].

Yet, despite some notable differences, the reliability of bibliographic databases is primarily seen as the accuracy of its citation information. While citations are input by hand in the case of professional databases, services like *CiteSeer* and *Google Scholar* use information retrieval and machine learning techniques to automatically parse citations from publication manuscripts[3,4]. Expectedly, this greatly impacts bibliometric analyses[20] and standard metrics of scientific evaluation like citation counts and *h*-index[17,22]. Although networks of citations between scientific papers have been studied since the 1950s[7,13], and are also commonly used in the modern network analysis literature[23,24], there exists no statistical comparison of citation topology of different bibliographic databases.

In this study, we compare the topological consistency of citation networks extracted from six popular bibliographic databases (see Methods). The networks are assessed through local and global graph statistics by a

Table 1 | Descriptive statistics and field decompositions of citation and other networks. Respective bibliographic or online databases are given under the column denoted by ''Source''. Descriptive statistics list the number of network nodes *n* and links *m*, and the percentage of nodes in the largest weakly connected component (column labelled ''% WCC''). Columns labelled ''% In-field'', ''% Core'' and ''% Out-field'' report the percentages of nodes in each of the components of the field bow-tie decomposition (see Methods)

	Descriptive statistics			Field decomposition		
Source	# Nodes	# Links	% WCC	% In-field	% Core	% Out-field
WoS	140,362	639,110	97.0%	11.2%	51.4%	34.4%
CiteSeer	384,413	1,744,619	95.0%	10.5%	37.7%	46.8%
Cora	23,166	91,500	100.0%	8.5%	51.4%	40.1%
HistCite	4,324	41,595	98.7%	44.8%	52.2%	1.6%
DBLP	12,591	49,744	99.2%	74.5%	16.9%	7.8%
arXiv	34,546	421,534	99.6%	6.7%	74.7%	18.1%
Gnutella	62,586	147,892	100.0%	73.8%	25.7%	0.5%
Twitter	81,306	1,768,135	100.0%	13.8%	86.2%	0.0%

methodology borrowed from the machine learning literature[25]. We first reveal statistically significant inconsistencies between some of the databases with respect to individual graph statistics. For example, the introduced field bow-tie decomposition of *DBLP Computer Science Bibliography* substantially differs from the rest due to the coverage of the database or the sampling procedure, while the citation information within *arXiv.org* is proven to be the most exhaustive. Finally, we compare the consistency of databases over multiple graph statistics. The citation topology of *DBLP Computer Science Bibliography* is the least consistent with the rest, while, not surprisingly, *Web of Science* is significantly more reliable from this perspective. Note that the reliability is here seen as a deviation from the majority (see Discussion). Differences between other databases are not statistically significant. This work can serve either as a reference for scholars in bibliometrics and scientometrics or a scientific evaluation guideline for governments and research agencies.

Results

Citation networks representing bibliographic databases are compared through 21 graph statistics described in Methods. In the following, we discuss the values of statistics in the context of complex network theory. Next, we reveal some statistically significant differences in individual statistics using Student *t*-test[26]. We then select ten statistics whose independence is confirmed by Fisher *z*-test[27] and show that the databases display significant inconsistencies in the selected statistics using Friedman rank test[28,29]. Last, the databases with no significant inconsistencies are revealed by Nemenyi post-hoc test[30] and the critical difference diagram[25]. Finally, we also compare the bibliographic databases with the selected online databases to verify the predictive power of the employed statistical methodology. See Methods for further details on statistical comparison.

Graph statistics of citation networks. Table 1 shows descriptive statistics of citation networks. The networks range from thousands of nodes to millions of links, while the largest weakly connected components contain almost all the nodes. This is consistent with the occurrence of a giant connected component in random graphs[31]. Directed networks are often assessed also according to their bow-tie structure[32]. However, due to the acyclic nature of citation networks where papers can only cite papers from the past, the decomposition proves meaningless. We introduce the field bow-tie decomposition into the in-field component, which consists of papers citing no other paper, the out-field component, which consists of papers not cited by any other paper, and the field core. The out-field component thus includes the research front[7], and the in-field and core components include the knowledge or intellectual base[33]. Table 1 shows the percentage of nodes in each of the field components, while a visual representation is given in Fig. 1. Notice that, in most cases, the majority of papers is included in the core and out-field components of the citation networks. Nevertheless, the main mass of the papers shifts towards the in-component in HistCite and DBLP databases (Figure 1, panels D and E). Since the former consists of papers from merely major journals and conferences, and the latter is based on the bibliography of a single author, many of the papers in the databases cite no other. Hence, reducing a bibliographic database to only a subset of publications or authors gives notably different citation structure and also influences many common graph statistics.

Table 2 shows degree statistics of citation networks. Observe that the mean degree $\langle k \rangle$ is around 8.8 in all cases except arXiv database, which, somewhat surprisingly, coincides with the common density of real-world networks[34]. Note, however, that since $\langle k \rangle/2 = \langle k_{in} \rangle = \langle k_{out} \rangle$ for any network, the papers cite and are cited by only four other papers on average. This number becomes meaningful when one considers that far more citations come from outside the field[18,35], whereas all databases are subsets of their respective fields in some sense. Considerably higher $\langle k \rangle$ in arXiv database is most likely due to several reasons. In contrast to other databases, *arXiv.org* stores journal and conference papers, technical reports, draft manuscripts that never came to print etc. Next, the citation network studied has been released within the KDD Cup 2003 (http://www.cs.cornell.edu/projects/kddcup) and has thus presumably been cleansed appropriately. Also, the subset of *arXiv.org* considered consists of physics publications, while other databases consist of computer science publications. Regardless of the true reason, the citation information within arXiv database is notably more exhaustive, which clearly reflects in its graph structure (see field bow-tie in Fig. 1, panel F).

Figure 1 plots degree distributions of citation networks, while the corresponding scale-free[36] exponents γ, γ_{in} and γ_{out} are given in Table 2. We stress that not all distributions, especially out-degree distributions, are a valid fit to a power-law form[37]. Nevertheless, the degree distributions further confirm the inconsistencies observed above. A larger number of non-citing papers results in a less steep out-degree distribution, whereas $\gamma_{out} \approx 2.6$ for HistCite and DBLP databases, while $\gamma_{out} \approx 3.8$ otherwise. On the contrary, the in-degree distribution of HistCite database is much steeper with $\gamma_{in} = 3.5$, while $\gamma_{in} \approx 2.5$ for the rest. In fact, $\gamma_{in} > \gamma_{out}$ for HistCite database, whereas $\gamma_{in} < \gamma_{out}$ for all others. Finally, the lack of low-citing papers in arXiv database prolongs the degree distributions towards the right-hand side of the scale (see Fig. 1, panel F).

Degree mixing[38] in Table 2 reveals no particularly strong correlations. Still, the in-degree and out-degree mixing coefficients $r_{(in,in)}$ and $r_{(out,out)}$ show positive correlation, while the undirected degree mixing r is negative. For comparison, $r \gg 0$ in social networks, and $r \ll 0$ for Internet and the Web[38,39]. Again, HistCite and DBLP databases deviate from common behaviour due to the reasons given above. For example, the directed degree mixing coefficient $r_{(out,in)}$

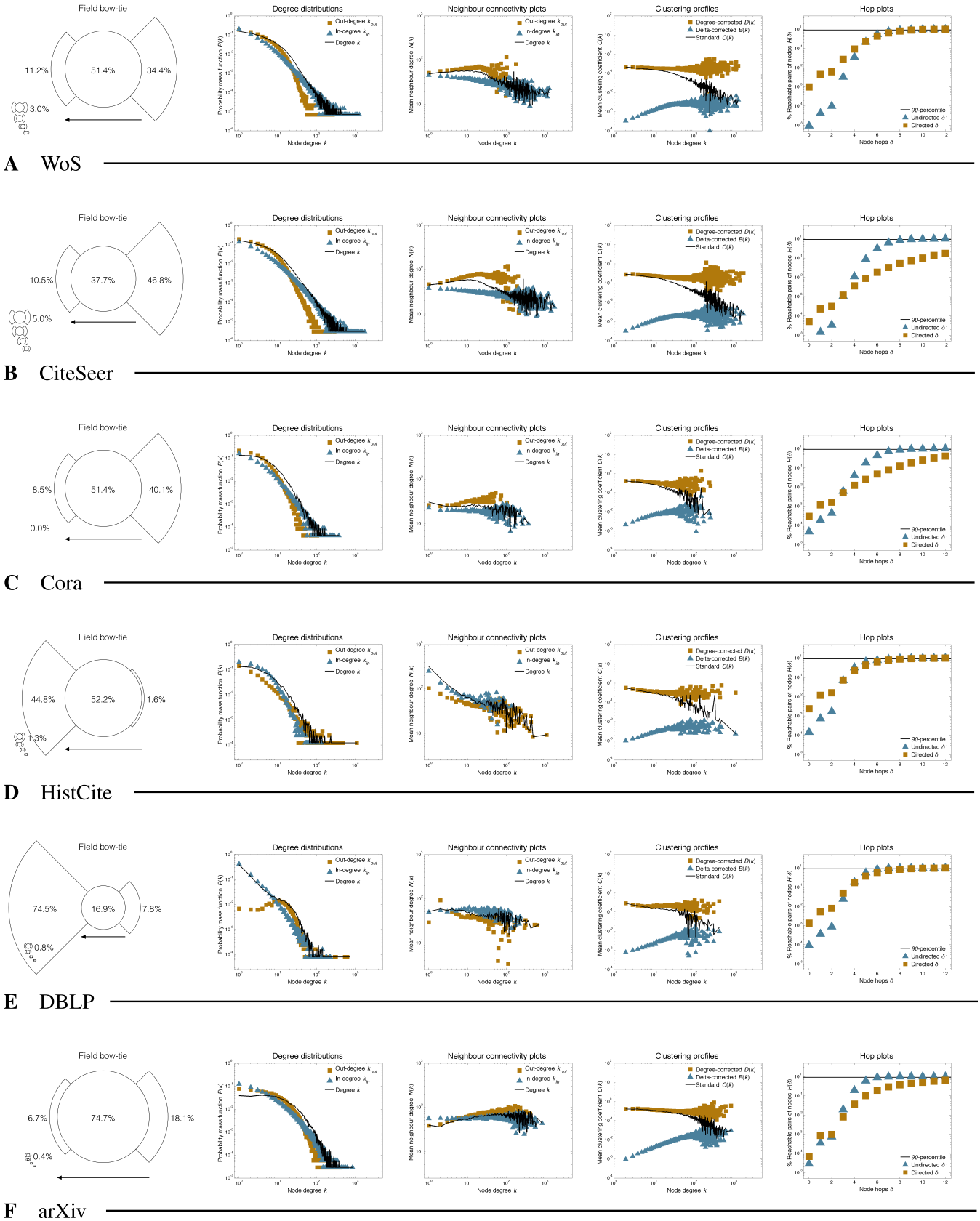


Figure 1 | Profile of citation networks extracted from bibliographic databases. Panels (A–F) show different distributions, plots and profiles of citation networks extracted from bibliographic databases. These are (from left to right): the field bow-tie decompositions, where the arrows illustrate the direction of the links and the areas of components are proportional to the number of nodes contained; the degree, in-degree and out-degree distributions $P(k)$, $P(k_{in})$ and $P(k_{out})$, respectively; the corresponding neighbour connectivity plots $N(k)$, $N(k_{in})$ and $N(k_{out})$; the clustering profiles of the standard and both unbiased coefficients $C(k)$, $B(k)$ and $D(k)$, respectively; and the hop plots for the standard and undirected diameters δ and δ', respectively (see Methods).

Table 2 | Degree distributions and mixing of citation and other networks. Respective bibliographic or online databases are given under the column denoted by "Source". Degree distributions are represented by the mean network degree $\langle k \rangle$ and the scale-free exponents of the power-law degree, in-degree and out-degree distributions (columns labelled "γ", "γ_{in}" and "γ_{out}", respectively). Degree mixing statistics list the undirected mixing coefficient r and four directed degree mixing coefficients $r_{(\alpha,\beta)}$, $\alpha, \beta \in \{in, out\}$ (see Methods)

	Degree distributions				Degree mixing				
Source	$\langle k \rangle$	γ	γ_{in}	γ_{out}	r	$r_{(in,in)}$	$r_{(in,out)}$	$r_{(out,in)}$	$r_{(out,out)}$
WoS	9.11	2.74	2.39	3.88	−0.06	0.04	−0.02	−0.03	0.09
CiteSeer	9.08	2.65	2.28	3.82	−0.06	0.05	0.00	0.00	0.12
Cora	7.90	2.88	2.60	4.00	−0.06	0.07	0.02	0.00	0.17
HistCite	9.99	2.55	3.50	2.37	−0.10	0.11	0.01	−0.13	0.00
DBLP	7.90	2.42	2.64	2.75	−0.05	0.00	−0.02	−0.05	−0.02
arXiv	24.40	2.67	2.54	3.45	−0.01	0.08	−0.04	0.00	0.11
Gnutella	4.73	6.37	7.59	4.78	−0.09	0.03	0.01	−0.01	0.00
Twitter	43.49	2.05	2.31	2.37	−0.03	0.00	0.06	−0.02	0.06

is substantially lower for HistCite database, while all directed coefficients are relatively low for DBLP database. Figure 1 plots also neighbour connectivity profiles of citation networks. Notice dichotomous degree mixing[40] that is positive for smaller out-degrees and negative for larger in-degrees, represented by increasing or decreasing trend, respectively (see, e.g., Fig. 1, panels A and B). Similar observations were recently made also in software[41] and undirected biological[40] networks. Consistent with the above, these trends are not present in HistCite and DBLP databases (see Fig. 1, panels D and E).

Table 3 shows clustering[42] statistics of citation networks. The mean clustering coefficients $\langle c \rangle$, $\langle b \rangle$ and $\langle d \rangle$ greatly vary across the databases, whereas $\langle c \rangle \approx 0.15$ for WoS, CiteSeer and DBLP databases, and $\langle c \rangle \approx 0.3$ in the case of Cora, HistCite and arXiv databases. This may be an artefact of the coverage or the sampling procedure used for citation extraction, while clustering can also reflect the amount of citations copied from other papers[43,44] known as indirect citation[45]. Unbiased clustering mixing coefficients r_b and r_d in Table 3 reveal strong positive correlations, similar to other real-world networks[41]. However, as before, $r_d = 0.26$ for DBLP database, while $r_d \approx 0.4$ for all others. Figure 1 plots clustering profiles of citation networks. Due to degree mixing biases[46], $C(k) \sim k^{-\alpha}$ for $\alpha \approx 1$[47], while this behaviour is absent from corrected profiles $B(k)$ and $D(k)$.

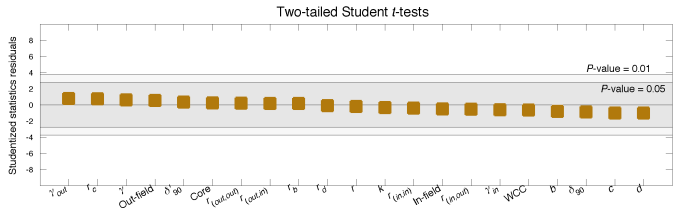
Table 3 shows also diameter statistics of citation networks. Undirected effective diameter δ'_{90} is somewhat consistent across the databases, in contrast to the directed variant δ_{90}, where $\delta_{90} \approx 8.5$ for WoS, HistCite and DBLP databases, while $\delta_{90} > 20$ for other databases. Low value of δ_{90} for HistCite and DBLP databases is due to the limited coverage discussed above, whereas the respective networks are also much smaller (see Table 1). On the other hand, low δ_{90} for WoS database is due to a rather non-intuitive phenomena that real-world networks shrink as they grow[23]. WoS database includes 50 years of literature, while the time span of, e.g., arXiv database is merely 10 years. The databases are thus not directly comparable in δ_{90} and neither is indeed inconsistent with the rest. Described can be more clearly observed in hop plots shown in Fig. 1 (see, e.g., panels A and B).

Comparison of databases by individual statistics. The above discussion was in many cases just qualitative. In the following, we reveal also statistically significant differences between some of the databases with respect to individual graph statistics. Since their values of a *true* citation network are, obviously, not known, we compute externally studentized residuals that measure the consistency of each database with the rest (Figure 2, panels A–F). Statistically significant inconsistencies in individual statistics are revealed by independent two-tailed Student t-tests (see Methods).

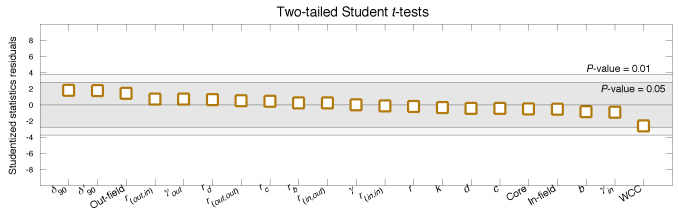
WoS, CiteSeer and Cora databases show no significant differences at P-value = 0.05. On the contrary, the scale-free in-degree exponent γ_{in} in HistCite database is significantly higher than in other databases, while the directed degree mixing coefficient $r_{(out,in)}$ is significantly lower (P-value = 0.019 and P-value = 0.033, respectively; see Table 2 and Fig. 2, panel D). This is a direct consequence of the limited coverage already noted above. For example, since the database is derived from a bibliography of a single author, highly cited papers are likely missing, which results in a much steeper citation distribution $P(k_{in})$ and thus higher γ_{in}. Next, the unbiased clustering mixing coefficient r_d is significantly lower in DBLP database (P-value = 0.017; see Table 3 and Fig. 2, panel E). Apparently, reducing the

Table 3 | Clustering and diameter statistics of citation and other networks. Respective bibliographic or online databases are given under the column denoted by "Source". Clustering distributions are represented by the means of the standard and unbiased clustering coefficients (columns labelled "$\langle c \rangle$", "$\langle b \rangle$" and "$\langle d \rangle$", respectively). Clustering mixing statistics list the corresponding mixing coefficients r_c, r_b and r_d. Diameter statistics report the means and s.e.m. of the standard and undirected effective diameters (columns labelled "δ_{90}" and "δ'_{90}", respectively)

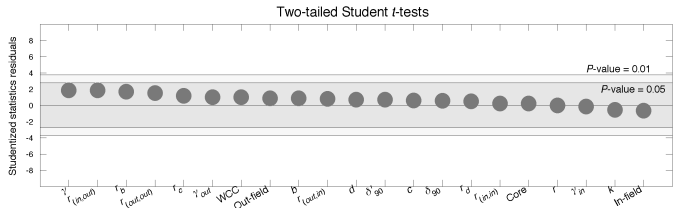
	Clustering distributions			Clustering mixing			Diameter statistics	
Source	$\langle c \rangle$	$\langle b \rangle$	$\langle d \rangle$	r_c	r_b	r_d	δ_{90}	δ'_{90}
WoS	0.14	$0.08 \cdot 10^{-2}$	0.16	0.16	0.43	0.36	8.85 ± 0.01	7.79 ± 0.03
CiteSeer	0.18	$0.07 \cdot 10^{-2}$	0.21	0.14	0.44	0.40	28.57 ± 0.23	9.01 ± 0.04
Cora	0.27	$0.46 \cdot 10^{-2}$	0.32	0.17	0.50	0.40	21.12 ± 0.16	8.17 ± 0.03
HistCite	0.31	$0.20 \cdot 10^{-2}$	0.36	0.05	0.36	0.41	7.97 ± 0.03	7.22 ± 0.04
DBLP	0.12	$0.14 \cdot 10^{-2}$	0.14	0.10	0.35	0.26	9.13 ± 0.07	6.24 ± 0.02
arXiv	0.28	$0.64 \cdot 10^{-2}$	0.33	0.13	0.46	0.39	21.71 ± 0.12	6.04 ± 0.02
Gnutella	0.01	$0.03 \cdot 10^{-2}$	0.01	0.09	0.25	0.17	12.83 ± 0.11	7.70 ± 0.01
Twitter	0.57	$0.35 \cdot 10^{-2}$	0.63	0.09	0.54	0.40	6.90 ± 0.02	5.50 ± 0.01



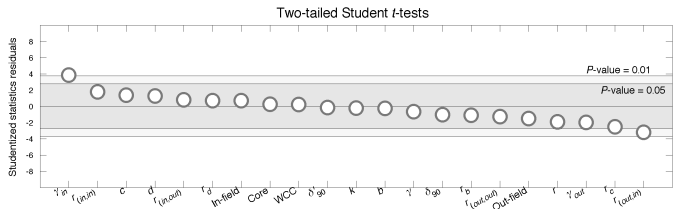
A WoS



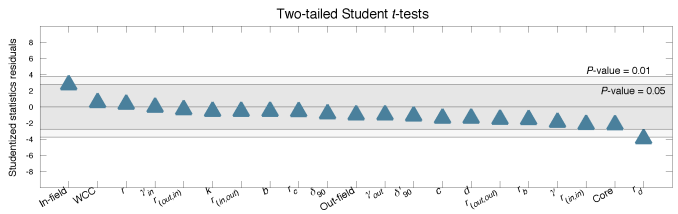
B CiteSeer



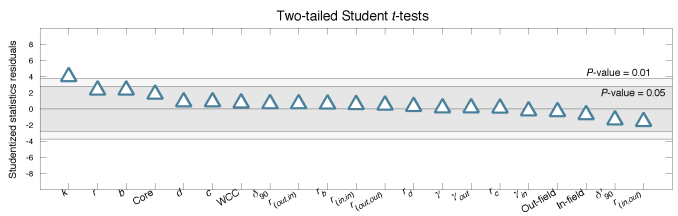
C Cora



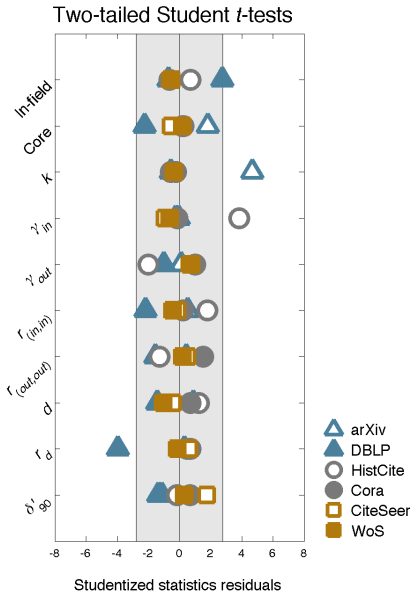
D HistCite



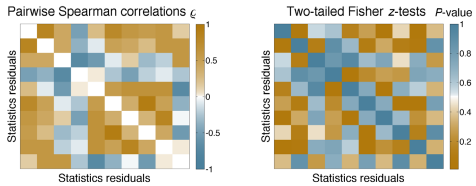
E DBLP



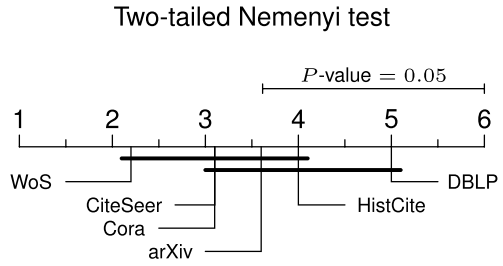
F arXiv



G Statistics



H Independence



I Comparison

Figure 2 | Comparison of bibliographic databases through statistics of citation networks. Panels (A–F) show studentized statistics residuals of citation networks extracted from bibliographic databases. The residuals 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 (see Methods).

bibliographic database to only selected publications gives a rather heterogeneous citation structure, which does not share high clustering assortativity[41], $r_d \gg 0$, of other citation networks. Note that the differences in the field bow-tie decomposition of DBLP database become statistically significant at P-value = 0.052 (see below). Finally, as thoroughly discussed above, the citation information within arXiv database is significantly more exhaustive with much higher mean degree $\langle k \rangle$ (P-value = 0.009; see Table 1 and Fig. 2, panel F). Notice that statistically significant inconsistencies between the databases are, expectedly, merely a subset of the differences exposed through the expert analysis above. Still, in summary, the results reveal that bibliographic databases with substantially different coverage have significantly different citation topology.

At P-value = 0.1, several other inconsistencies become statistically significant. For CiteSeer database, the largest weakly connected component is significantly smaller than in other databases (P-value = 0.059; see Table 1 and Fig. 2, panel B); for HistCite database, the clustering mixing coefficient r_c is lower (P-value = 0.066; see Table 3 and Fig. 2, panel D); for DBLP database, the in-field component is larger (P-value = 0.052; see Table 1 and Fig. 2, panel E), while the field core and the directed degree mixing coefficient $r_{(in,in)}$ are smaller (P-value = 0.090 and P-value = 0.095, respectively; see Table 1 and Table 2, and Fig. 2, panel E); and for arXiv database, the undirected degree mixing coefficient r and the corrected clustering coefficient $\langle b \rangle$ are higher (P-value = 0.081; see Table 2 and Table 3, and Fig. 2, panel F). Note that, due to space limitations, not all inconsistencies at P-value = 0.1 are discussed in the analysis above.

Selection of independent graph statistics. Since the adopted graph statistics of citation networks are by no means independent[42,46], one cannot simply compare the bibliographic databases over all. For this purpose, we select ten statistics listed in Fig. 2, panel G, and verify their statistical independence (see Methods). We compute Fisher transformations of the pairwise Spearman correlations between the statistics, while significant correlations are revealed by independent two-tailed z-tests (Figure 2, panel H). Notice that no correlation is statistically significant at P-value = 0.01.

The selection of independent graph statistics proceeds as follows. We first discard statistics that are sums or aggregates of the others by definition. Namely, the sizes of the largest weakly connected and out-field components (see Table 1), the scale-free degree exponent γ, the undirected degree mixing r and also both mixed directed mixing coefficients $r_{(in,out)}$ and $r_{(out,in)}$ (see Table 2). We next discard statistics whose correlations have been proven in the literature[46] or are dependent on some intrinsic characteristic of the database like the time span of the publications (see above). Namely, the standard clustering $\langle c \rangle$ and the corresponding mixing coefficient r_c, and the directed effective diameter δ_{90} (see Table 3). Finally, out of the both unbiased clustering coefficients $\langle b \rangle$ and $\langle d \rangle$, we decide for the latter, and its corresponding mixing coefficient r_d (see Table 3). We are thus left with ten statistics (Figure 2, panel G). Namely, the sizes of the in-field and core components (see Table 2), the mean degree $\langle k \rangle$, the directed scale-free exponents γ_{in} and γ_{out}, and the directed degree mixing coefficients $r_{(in,in)}$ and $r_{(out,out)}$ (see Table 2), the unbiased clustering $\langle d \rangle$ and its corresponding mixing coefficient r_d, and the undirected effective diameter δ'_{90} (see Table 3).

For some further notes on statistics independence see Discussion.

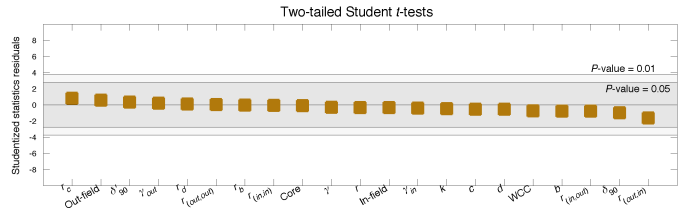
Comparison of databases over multiple statistics. In the following, we compare the bibliographic databases over independent graph statistics selected above. We rank the databases according to the studentized statistics residuals and compute their mean ranks over all statistics (see Methods). The final ranks are 2.2 for WoS database, 3.1 for both CiteSeer and Cora databases, 3.6 for arXiv database, 4.0 for HistCite database and 5.0 for DBLP database. Notice that the ranks indeed reflect the conclusions on database consistency given above. We reject the null hypothesis that the ranks of the databases are statistically equivalent by one-tailed Friedman test at P-value = 0.05 and thus compare the ranks by two-tailed Nemenyi post-hoc test (Figure 2, panel I). The databases whose ranks differ by more than a critical distance 2.38 show statistically significant inconsistencies in the selected statistics at P-value = 0.05. Hence, the citation topology of WoS database is significantly more reliable than that of DBLP database, which is the least consistent with the rest. On the other hand, the differences between other databases are not statistically significant, whereas concluding that these are consistent with *both* WoS and DBLP databases would be a statistical nonsense[25]. At P-value = 0.1, the critical distance drops to 2.17, while all conclusions still remain the same. Interestingly, neglecting the requirement for the independence of graph statistics and comparing the bibliographic databases over all 21 statistics, again gives exactly the same conclusions on their consistency. Although, the ranking changes, since arXiv database is ranked in front of Cora database.

For some further notes on database consistency see Discussion.

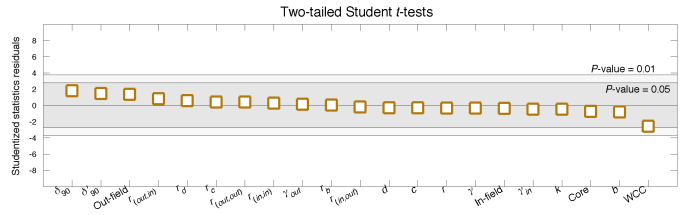
Comparison of bibliographic and online databases. To assess the power of the employed statistical methodology for quantifying the differences in network topology, we compare citation networks representing different bibliographic databases with two networks extracted from online databases. Namely, a technological network of Gnutella peer-to-peer file sharing (http://rfc-gnutella.sourceforge.net) from August 2002[23], where nodes are hosts and links are shares between them; and a social network representing Twitter social circles (http://twitter.com) crawled from public repositories[48], where nodes are users and links are follows between them. Both these networks are provided within SNAP (http://snap.stanford.edu), while their basic descriptive statistics are given in Table 1.

Note that online databases reveal knowingly different network topology than reliable bibliographic databases. For example, the majority of nodes in Gnutella database is included in the in-field component (see Methods), similarly as in DBLP database (see Table 1). Next, the mean degree $\langle k \rangle$ is considerably higher in Twitter database and lower in Gnutella database (see Table 2). Furthermore, the degree distributions of Gnutella database are not a valid fit to a power-law form[37] with higher scale-free degree exponents γ-s than in other databases (see Table 2). On the contrary, the scale-free out-degree exponent γ_{out} of Twitter database is lower, similarly as in HistCite database. Online databases also reveal notably different clustering regimes than bibliographic databases (see Table 3). The standard and unbiased clustering coefficients $\langle c \rangle$ and $\langle d \rangle$ are much higher in Twitter database, while much lower in Gnutella database. Finally, Gnutella database shows relatively heterogeneous clustering structure with very low unbiased clustering mixing coefficients r_b and r_d.

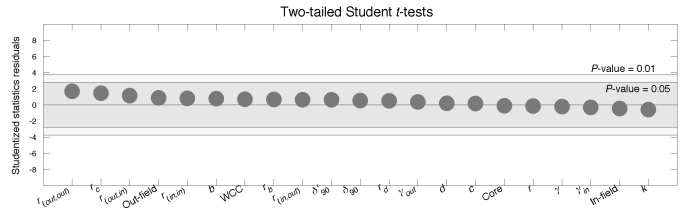
In the following, we reveal statistically significant inconsistencies between some of the databases with respect to individual graph statistics (see Methods). We consider the online databases and four most reliable bibliographic databases so that all critical values remain the same as before. Under this setting, the bibliographic databases show no inconsistencies at P-value = 0.05 (Figure 3, panels A–D). On the other hand, five most significant inconsistencies of online databases almost precisely coincide with the differences exposed through the analysis above (Figure 3, panels E and F). For Gnutella database, the in-field component is larger (P-value = 0.008), the degree and in-degree scale-free exponents γ and γ_{in} are higher (P-value = 0.011 and P-value = 0.008, respectively), and the unbiased clustering mixing coefficients r_b and r_d are lower (P-value = 0.032 and P-value = 0.011, respectively); and for Twitter database, the mean degree $\langle k \rangle$ is higher (P-value = 0.039), the out-degree scale-free exponent γ_{out} and the directed degree mixing coefficient $r_{(in,in)}$ are lower (P-value = 0.063 and P-value = 0.066, respectively), and the standard and unbiased clustering coefficients $\langle c \rangle$ and $\langle d \rangle$ are higher (P-value = 0.056 and P-value = 0.065, respectively).



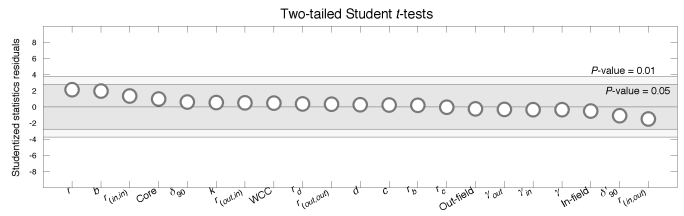
A WoS



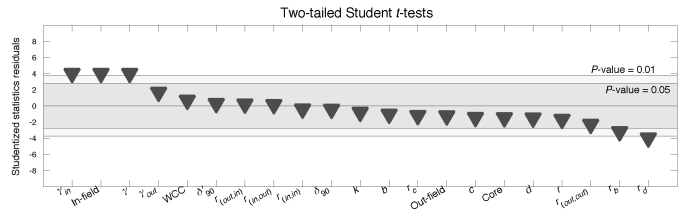
B CiteSeer



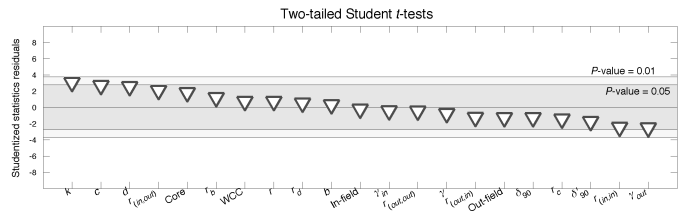
C Cora



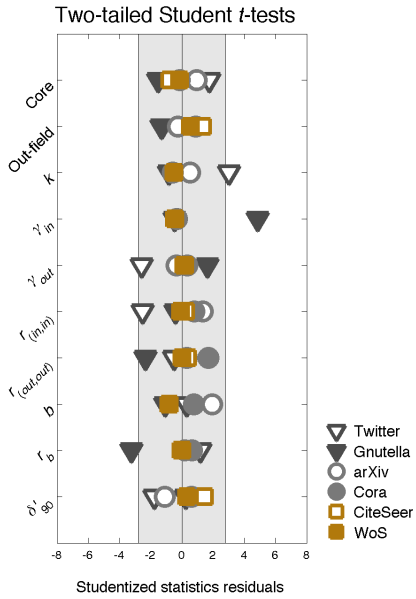
D arXiv



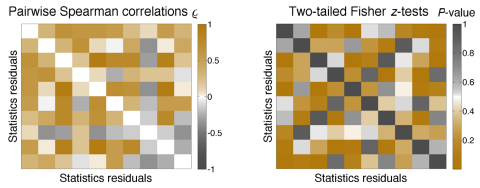
E Gnutella



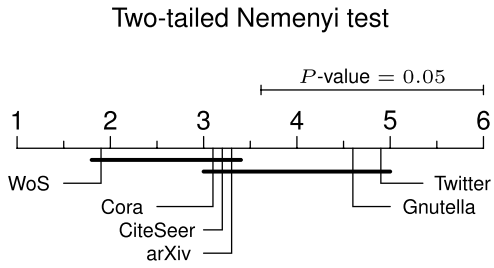
F Twitter



G Statistics



H Independence



I Comparison

Figure 3 | Comparison of bibliographic and online databases through statistics of networks. Panels (A–D) show studentized statistics residuals of citation networks extracted from bibliographic databases, while panels (E) and (F) show residuals of social and technological networks extracted from online databases. The residuals 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 (see Methods).

In the remaining, we also rank the databases over multiple graph statistics as before (see Methods). We select ten statistics listed in Fig. 3, panel G, whose pairwise independence is confirmed at P-value $= 0.001$ (Figure 3, panel H). The overall ranks of the databases are not statistically equivalent at P-value $= 0.05$ and are given in Fig. 3, panel I. Expectedly, the online databases are the least consistent with the rest, whereas the ranks are 4.6 and 4.9 for Gnutella and Twitter databases, respectively, and 1.9–3.3 for the bibliographic databases. Yet, merely WoS bibliographic database significantly differs from the online databases at P-value $= 0.05$ (see Fig. 3, panel I).

In summary, the employed statistical testing proves to be rather effective in quantifying the inconsistencies between network databases with respect to individual graph statistics. On the contrary, the comparison over multiple statistics appears to be less powerful and cannot distinguish between the online databases and all bibliographic databases considered above. Nevertheless, the statistically significant inconsistencies between WoS and DBLP bibliographic databases highlighted in the study can thus indeed be regarded as rather substantial.

Discussion

We conduct an extensive statistical analysis of the citation information within six popular bibliographic databases. We extract citation networks and compare their topological consistency through a large number of graph statistics. We expose statistically significant inconsistencies between some of the databases with respect to individual graph statistics and compare the databases over multiple statistics. *DBLP Computer Science Bibliography* is found to be the least consistent with the rest, while *Web of Science* is significantly more reliable from this perspective. The result is somewhat surprising, since *DBLP Computer Science Bibliography* 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, although this can be avoided in the case of the latter. This work can serve either as a reference for the analyses of citation networks in bibliometrics and scientometrics literature or a guideline for scientific evaluation based on some particular bibliographic database or literature coverage policy.

We introduce the field bow-tie decomposition of a citation network (see Methods), which proves to be one of the most discriminative approaches for comparing the citation topology of bibliographic databases (see Results). We also consider 18 other local and global graph statistics. Nevertheless, we neglect some possible common patterns of nodes like motifs[49] and graphlets[50], and the occurrence of larger characteristic groups of nodes like communities[51] and modules[52]. Yet, these structures are not well understood for the specific case of citations networks and thus not easily interpretable.

In the following, we provide some further notes on the representativeness and reliability of the bibliographic databases, and the independence of the databases and adopted graph statistics.

As discussed in Methods, citation networks extracted from bibliographic databases are not necessarily representative due to citation retrieval procedure, data preprocessing techniques, size or other. It should, however, be noted that this work has been done after realizing that citation networks available from the Web provide a rather inconsistent view on the structure of bibliographic information. We have therefore collected and compared all such networks, while including also a citation network extracted from *Web of Science*. In that sense, the adopted networks are representative of the data readily available for the analyses and thus also commonly used in the literature[23,24]. Still, other citation networks could give different conclusions on the reliability of bibliographic databases. In particular because the reliability is measured through consistency of the databases. The concepts are of course not equivalent, yet the study reveals that, in most cases, only a single database deviates from a common behaviour for some particular graph statistic (see Results). Hence, the reliability can indeed be seen as a deviation from the majority to a rather good approximation.

Independence between bibliographic databases is obtained trivially, since these are either based on independent bibliographic sources or cover different literature (see Methods). On the other hand, adopted graph statistics of citation networks are by no means independent[42,46]. As this is required by several statistical tests, we reduce the statistics to a subset whose pairwise independence could be proven. Nevertheless, we only show that the statistics are not clearly dependent and we do not ensure their mutual independence. Although the conclusions of the study are exactly the same regardless of whether it is based on all or merely independent statistics (see Results), further reducing the subset of statistics would discard relevant information and no statistically significant conclusions could be made. We also stress that all results have been verified by an independent expert analysis. An alternative solution would be to transform the statistics into uncorrelated representatives using matrix factorization techniques like principal component analysis[53]. However, interpreting inconsistencies in, e.g., $0.9\gamma_{in} - 1.4r_c + 0.3\delta_{90}$ would most likely be far from trivial.

Methods

Bibliographic sources. In this study, we conduct a network-based comparison of citation topology of six bibliographic databases. These have been extracted from publicly available and commercial bibliographic sources, services, software and a preprint repository with particular focus on computer science publications. For bibliographic sources based on a similar methodology[14,15] (e.g., *Web of Science* and *Scopus*, *CiteSeer* and *Google Scholar*), a single exemplar has been selected. We have extracted a citation network from each of the selected databases. Publications neither citing nor cited by any other are discarded and any self-citations that occur due to errors in the databases are removed prior to the analysis (see below and Table 1 for details). Although the databases contain fair portions of the respective bibliographic sources, we stress that they are not all necessarily representative. Still, in most cases, these are the only examples of citation networks readily available online (due to our knowledge) and thus also often used in the network analysis literature[23,24].

WoS database. Web of Science (WoS) is informally considered as the most accurate bibliographic source in the world. It is hand-maintained by professional staff at Thomson Reuters (http://thomsonreuters.com), previously Institute for Scientific Information. It dates back to the 1950s[7,13] and contains over 45 million records of publications from all fields of science[14]. For this study, we consider all journal papers in WoS category *Computer Science, Artificial Intelligence* as of October 2013. The extracted database spans 50 years, and contains 179,510 papers from 877 journals and 639,126 citations between them. Note that 39,148 papers neither cite nor are cited by any other, while the database includes 16 self-citations.

CiteSeer database. CiteSeer or *CiteSeer*^{x} (CiteSeer) is constructed by automatically crawling the Web for freely accessible manuscripts of publications and then analyzing the latter for potential citations to other publications[3] (http://citeseer.ist.psu.edu). It became publicly available in 1998 and is maintained by Pennsylvania State University. It contains over 32 million publication records from computer and information science[14]. For this study, we consider a snapshot of the database provided within KONECT (http://konect.uni-koblenz.de) that contains 723,131 publications and 1,751,492 citations between them. Note that 338,718 publications neither cite nor are cited by any other, while the database includes 6,873 self-citations.

Cora database. Computer Science Research Paper Search Engine (Cora) is a service for automatic retrieval of publication manuscripts from the Web using machine learning techniques[4] (http://people.cs.umass.edu/∼mccallum). It contains over 50,000 publication records collected from the websites of computer science departments at major universities in August 1998. For this study, we consider a subset of the database that contains 23,166 publications and 91,500 citations between them[54] (http://lovro.lpt.fri.uni-lj.si). Note that all papers either cite or are cited by some other, while the database includes no self-citations.

HistCite database. Algorithmic Historiography (HistCite) is a software package for analysis and visualization of bibliographic databases owned by Thomson Reuters (http://www.histcite.com). It was developed in the 2000s for extracting publication records from WoS database[55]. For this study, we consider a complete bibliography of Nobel laureate Joshua Lederberg produced by HistCite in February 2008. The database contains 8,843 publications and 41,609 citations between them (http://vlado.fmf.uni-lj.si). Note that 4,519 publications neither cite nor are cited by any other, while the database includes 14 self-citations.

DBLP database. DBLP Computer Science Bibliography (DBLP) indexes major journals and proceedings from all fields of computer science[2] (http://dblp.uni-trier.de). It is freely available since 1993 and hand-maintained by University of Trier. It contains more than 2.3 million records of publications, while the citation information is extremely scarce compared to WoS and CiteSeer databases[14]. For this study, we consider a snapshot of the database provided within KONECT (http://konect.uni-koblenz.de) that contains 12,591 journal and conference papers, and 49,759 citations between them. Note that all papers either cite or are cited by some other, while the database includes 15 self-citations.

arXiv database. arXiv.org (arXiv) is a public preprint repository of publication drafts uploaded by the authors prior to an actual journal or conference submission (http://arxiv.org). It began in 1991[1] and is hosted at Cornell University. It currently contains almost one million publications from physics, mathematics, computer science and other fields. For this study, we consider all publications in arXiv category *High Energy Physics Phenomenology* as of April 2003[23] provided within SNAP (http://snap.stanford.edu). The database spans over 10 years, and contains 34,546 publications and 421,578 citations between them. Note that all publications either cite or are cited by some other, while the database includes 44 self-citations.

Citation topology. Citation networks extracted from bibliographic databases are represented with directed graphs, where papers are nodes of the graph and citations are directed links between the nodes. The topology of citation networks is assessed through a rich set of local and global graph statistics.

Descriptive and field statistics. The citation network is a simple directed graph $G(V, L)$, where V is the set of nodes, $n = |V|$, and L is the set of links, $m = |L|$. Weakly connected component (WCC) is a subset of nodes reachable from one another not considering the directions of the links. Field bow-tie is a decomposition of the largest WCC of a citation network into the in-field component, which consists of nodes with no outgoing links, the out-field component, which consists of nodes with no incoming links, and the field core.

Degree distributions and mixing. The in-degree k_{in} or out-degree k_{out} of a node is the number of incoming and outgoing links, respectively. k is the degree of a node, $k = k_{in} + k_{out}$, and $\langle k\rangle$ denotes the mean degree. γ is the scale-free exponent of a power-law degree distribution $P(k) \sim k^{-\gamma}$, and γ_{in} and γ_{out} are the scale-free exponents of $P(k_{in})$ and $P(k_{out})$[36]. Power-laws are fitted to the tails of the distributions by maximum-likelihood estimation, $\gamma_{\cdot} = 1 + n\left(\sum_V \ln k_{\cdot}/k_{min}\right)^{-1}$ for $k_{min} \in \{10, 25\}$. Neighbour connectivity plots show the mean neighbour degree $N(k_{\cdot})$ of nodes with degree k.[56]. The degree mixing $r_{(\alpha,\beta)}$ is the Pearson correlation coefficient of α-degrees or β-degrees at links' source and target nodes, respectively[57]:

$$r(\alpha,\beta) = \frac{1}{\sigma_{k_\alpha}\sigma_{k_\beta}} \sum_L \left(k_\alpha - \langle k_\alpha\rangle\right)\left(k_\beta - \langle k_\beta\rangle\right), \tag{1}$$

where $\langle k_{\cdot}\rangle$ and $\sigma_{k_{\cdot}}$ are the means and standard deviations, $\alpha, \beta \in \{in, out\}$. r is the mixing of degrees k[39].

Clustering distributions and mixing. Node clustering coefficient c is the density of its neighbourhood[42]:

$$c = \frac{2t}{k(k-1)}, \tag{2}$$

where t is the number of linked neighbours and $k(k-1)/2$ is the maximum possible number, $c = 0$ for $k \leq 1$. The mean $\langle c\rangle$ is denoted network clustering coefficient[42], while the clustering mixing r_c is defined as before. Clustering profile shows the mean clustering $C(k)$ of nodes with degree k[58]. Note that the denominator in equation (2) introduces biases[46], particularly when $r < 0$. Thus, delta-corrected clustering coefficient b is defined as $c \cdot k/\Delta$[59], where Δ is the maximal degree k and $b = 0$ for $k \leq 1$. Also, degree-corrected clustering coefficient d is defined as t/ω[46], where ω is the maximum number of linked neighbours with respect to their degrees k and $d = 0$ for $k \leq 1$. By definition, $b \leq c \leq d$.

Diameter statistics. Hop plot shows the percentage of reachable pairs of nodes $H(\delta)$ within δ hops[23]. The diameter is the minimal number of hops δ for which $H(\delta) = 1$, while the effective diameter δ_{90} is defined as the number of hops at which 90% of such pairs of nodes are reachable[23], $H(\delta_{90}) = 0.9$. δ' denotes the respective number of hops in a corresponding undirected graph. Hop plots are estimated over 100 realizations of the approximate neighbourhood function with 32 trials[60].

Statistical comparison. Citation networks representing bibliographic databases are compared through 21 graph statistics introduced above. These are by no means independent[42,46], neither are their values of a *true* citation network known. We thus compute externally studentized residuals of graph statistics that measure the consistency of each bibliographic database with the rest. Statistically significant inconsistencies in individual graph statistics are revealed by Student t-test[26]. We select ten graph statistics whose pairwise independence is verified using Fisher z-transformation[27]. Friedman rank test[28] confirms that bibliographic databases display significant inconsistencies in the selected statistics, while the databases with no significant differences are revealed by Nemenyi test[25,30].

Studentized statistics residuals. Denote x_{ij} to be the value of j-th graph statistic of i-th bibliographic database, where N is the number of databases, $N = 6$. Corresponding externally studentized residual $\hat{x}_{ij}$ is:

$$\hat{x}_{ij} = \frac{x_{ij} - \hat{\mu}_{ij}}{\hat{\sigma}_{ij}\sqrt{1 - 1/N}}, \tag{3}$$

where $\hat{\mu}_{ij}$ and $\hat{\sigma}_{ij}$ are the sample mean and corrected standard deviation excluding the considered i-th database, $\hat{\mu}_{ij} = \sum_{k\neq i} x_{kj}/(N-1)$ and $\hat{\sigma}^2_{ij} = \sum_{k\neq i}\left(x_{kj} - \hat{\mu}_{ij}\right)^2 \Big/ (N-2)$. Assuming that the errors in x are independent and normally distributed, the residuals $\hat{x}$ have Student t-distribution with $N - 2$ degrees of freedom. Significant differences in individual statistics x are revealed by independent two-tailed Student t-tests[26] at P-value = 0.05, rejecting the null hypothesis H_0 that x are consistent across the databases, $H_0 : \hat{x} = 0$. Notice that the absolute values of individual residuals $|\hat{x}|$ imply a ranking R over the databases, where the database with the lowest $|\hat{x}|$ has rank one, the second one has rank two and the one with the largest $|\hat{x}|$ has rank N.

Pairwise statistics independence. Denote r_{ij} to be the Pearson product-moment correlation coefficient of the residuals $\hat{x}$ for i-th and j-th graph statistics over all bibliographic databases. Spearman rank correlation coefficient ρ_{ij} is defined as the Pearson coefficient of the ranks R for i-th and j-th statistics. Under the null hypothesis of statistical independence of i-th and j-th statistics, $H_0 : \rho_{ij} = 0$, adjusted Fisher transformation[27]:

$$\frac{\sqrt{N-3}}{2} \ln \frac{1 + r_{ij}}{1 - r_{ij}} \tag{4}$$

approximately follows a standard normal distribution. Pairwise independence of the selected graph statistics is thus confirmed by independent two-tailed z-tests at P-value = 0.01.

Comparison of bibliographic databases. Significant inconsistencies between bibliographic databases are exposed using the methodology introduced for comparing classification algorithms over multiple data sets[25]. Denote R_i to be the mean rank of i-th database over the selected graph statistics, $R_i = \sum_j R_{ij}/K$, where K is the number of statistics, $K = 10$. One-tailed Friedman rank test[28,29] first verifies the null hypothesis that the databases are statistically equivalent and thus their ranks R_i should equal, $H_0 : R_i = R_j$. Under the assumption that the selected statistics are indeed independent, the Friedman testing statistic[28]:

$$\frac{12K}{N(N+1)}\left(\sum_i R_i^2 - \frac{N(N+1)^2}{4}\right) \tag{5}$$

has χ^2-distribution with $N - 1$ degrees of freedom. By rejecting the hypothesis at P-value = 0.05, we proceed with the Nemenyi post-hoc test that reveals databases whose ranks R_i differ more than the critical difference[30]:

$$q\sqrt{\frac{N(N+1)}{6K}}, \tag{6}$$

where q is the critical value based on the studentized range statistic[25], $q = 2.85$ at P-value = 0.05. A critical difference diagram plots the databases with no statistically significant inconsistencies in the selected statistics[25].

1. Ginsparg, P. ArXiv at 20. *Nature* **476**, 145–147 (2011).
2. Ley, M. The DBLP computer science bibliography: Evolution, research issues, perspectives. In *Proceedings of the International Symposium on String Processing and Information Retrieval*, 1–10 (London, UK, 2002).
3. Bollacker, K. D., Lawrence, S. & Giles, C. L. CiteSeer: an autonomous web agent for automatic retrieval and identification of interesting publications. In *Proceedings of the International International Conference on Autonomous Agents*, 116–123 (Minneapolis, MN, USA, 1998).
4. McCallum, A. K., Nigam, K., Rennie, J. & Seymore, K. Automating the construction of internet portals with machine learning. *Inform. Retrieval* **3**, 127–163 (2000).
5. Wang, D., Song, C. & Barabási, A.-L. Quantifying long-term scientific impact. *Science* **342**, 127–132 (2013).
6. Newman, M. E. J. Prediction of highly cited papers. *Europhys. Lett.* **105**, 28002 (2014).
7. Price, D. J. d. S. Networks of scientific papers. *Science* **149**, 510–515 (1965).
8. Eom, Y.-H. & Fortunato, S. Characterizing and modeling citation dynamics. *PLoS ONE* **6**, e24926 (2011).
9. Laherrère, J. & Sornette, D. Stretched exponential distributions in nature and economy: "Fat tails" with characteristic scales. *Eur. Phys. J. B* **2**, 525–539 (1998).
10. Radicchi, F., Fortunato, S. & Castellano, C. Universality of citation distributions: Toward an objective measure of scientific impact. *P. Natl. Acad. Sci. USA* **105**, 17268–17272 (2008).

11. Wallace, M. L., Larivière, V. & Gingras, Y. Modeling a century of citation distributions. *J. Infometrics* **3**, 296–303 (2009).
12. Van Raan, A. F. J. Competition amongst scientists for publication status: Toward a model of scientific publication and citation distributions. *Scientometrics* **51**, 347–357 (2001).
13. Garfield, E. Citation indexes for science: A new dimension in documentation through association of ideas. *Science* **122**, 108–111 (1955).
14. Fiala, D. Mining citation information from CiteSeer data. *Scientometrics* **86**, 553–562 (2011).
15. Falagas, M. E., Pitsouni, E. I., Malietzis, G. A. & Pappas, G. Comparison of PubMed, scopus, web of science, and google scholar: Strengths and weaknesses. *FASEB J.* **22**, 338–342 (2008).
16. Vieira, E. S. & Gomes, J. A. N. F. A comparison of scopus and web of science for a typical university. *Scientometrics* **81**, 587–600 (2009).
17. De Groote, S. L. & Raszewski, R. Coverage of google scholar, scopus, and web of science: A case study of the h-index in nursing. *Nurs. Outlook* **60**, 391–400 (2012).
18. Redner, S. Citation statistics from 110 years of physical review. *Phys. Today* **58**, 49–54 (2005).
19. Jacso, P. As we may search: Comparison of major features of the web of science, scopus, and google scholar citation-based and citation-enhanced databases. *Curr. Sci.* **89**, 1537–1547 (2005).
20. Petricek, V., Cox, I. J., Han, H., Councill, I. G. & Giles, C. L. A comparison of on-line computer science citation databases. In *Proceedings of the European Conference on Research and Advanced Technology for Digital Libraries*, 438–449 (Vienna, Austria, 2005).
21. Hull, D., Pettifer, S. R. & Kell, D. B. Defrosting the digital library: Bibliographic tools for the next generation web. *PLoS Comput. Biol.* **4**, e1000204 (2008).
22. Meho, L. I. & Yang, K. Impact of data sources on citation counts and rankings of LIS faculty: Web of science versus scopus and google scholar. *J. Am. Soc. Inf. Sci.* **58**, 2105–2125 (2007).
23. Leskovec, J., Kleinberg, J. & Faloutsos, C. Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data* **1**, 1–41 (2007).
24. Liu, Y.-Y., Slotine, J.-J. & Barabasi, A.-L. Controllability of complex networks. *Nature* **473**, 167–173 (2011).
25. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **7**, 1–30 (2006).
26. Cook, R. D. & Weisberg, S. *Residuals and Influence in Regression* (Chapman and Hall, New York, 1982).
27. Fisher, R. A. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10**, 507 (1915).
28. Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Am. Stat. Assoc.* **32**, 675–701 (1937).
29. Friedman, M. A comparison of alternative tests of significance for the problem of m rankings. *Ann. Math. Stat.* **11**, 86–92 (1940).
30. Nemenyi, P. B. *Distribution-free multiple comparisons*. PhD thesis, Princeton University (1963).
31. Erdős, P. & Rényi, A. On random graphs i. *Publ. Math. Debrecen* **6**, 290–297 (1959).
32. Broder, A. *et al.* Graph structure in the web. *Comput. Netw.* **33**, 309–320 (2000).
33. Persson, O. The intellectual base and research fronts of JASIS 1986–1990. *J. Am. Soc. Inf. Sci.* **45**, 31–38 (1994).
34. Laurienti, P. J., Joyce, K. E., Telesford, Q. K., Burdette, J. H. & Hayasaka, S. Universal fractal scaling of self-organized networks. *Physica A* **390**, 3608–3613 (2011).
35. Redner, S. Citation statistics from more than a century of physical review. *e-print arXiv:0407137v2* (2004).
36. Barabási, A. L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
37. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-law distributions in empirical data. *SIAM Rev.* **51**, 661–703 (2009).
38. Newman, M. E. J. Assortative mixing in networks. *Phys. Rev. Lett.* **89**, 208701 (2002).
39. Newman, M. E. J. Mixing patterns in networks. *Phys. Rev. E* **67**, 026126 (2003).
40. Hao, D. & Li, C. The dichotomy in degree correlation of biological networks. *PLoS ONE* **6**, e28322 (2011).
41. Šubelj, L., Žitnik, S., Blagus, N. & Bajec, M. Node mixing and group structure of complex software networks. *Adv. Complex Syst.* (2014). Accepted.
42. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
43. Simkin, M. V. & Roychowdhury, V. P. Read before you cite! *Compl. Syst.* **14**, 269–274 (2003).
44. Šubelj, L., Žitnik, S. & Bajec, M. Who reads and who cites? unveiling author citation dynamics by modeling citation networks. In *Proceedings of the International Conference on Network Science*, 1 (Berkeley, CA, USA, 2014).
45. Peterson, G. J., Pressé, S. & Dill, K. A. Nonuniversal power law scaling in the probability distribution of scientific citations. *P. Natl. Acad. Sci. USA* **107**, 16023–16027 (2010).
46. Soffer, S. N. & Vázquez, A. Network clustering coefficient without degree-correlation biases. *Phys. Rev. E* **71**, 057101 (2005).
47. Ravasz, E. & Barabási, A. L. Hierarchical organization in complex networks. *Phys. Rev. E* **67**, 026112 (2003).
48. McAuley, J. J. & Leskovec, J. Learning to discover social circles in ego networks. In *Proceedings of the Neural Information Processing Systems Conference*, 403–412 (Lake Tahoe, NV, USA, 2012).
49. Milo, R. *et al.* Network motifs: Simple building blocks of complex networks. *Science* **298**, 824–827 (2001).
50. Pržulj, N., Wigle, D. A. & Jurisica, I. Functional topology in a network of protein interactions. *Bioinformatics* **20**, 340–348 (2004).
51. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *P. Natl. Acad. Sci. USA* **99**, 7821–7826 (2002).
52. Šubelj, L. & Bajec, M. Ubiquitousness of link-density and link-pattern communities in real-world networks. *Eur. Phys. J. B* **85**, 32 (2012).
53. Pearson, K. On lines and planes of closest fit to systems of points in space. *Philos. Mag.* **2**, 559–572 (1901).
54. Šubelj, L. & Bajec, M. Model of complex networks based on citation dynamics. In *Proceedings of the WWW Workshop on Large Scale Network Analysis*, 527–530 (Rio de Janeiro, Brazil, 2013).
55. Garfield, E. Historiographic mapping of knowledge domains literature. *J. Inform. Sci.* **30**, 119–145 (2004).
56. Pastor-Satorras, R., Vázquez, A. & Vespignani, A. Dynamical and correlation properties of the internet. *Phys. Rev. Lett.* **87**, 258701 (2001).
57. Foster, J. G., Foster, D. V., Grassberger, P. & Paczuski, M. Edge direction and the structure of networks. *P. Natl. Acad. Sci. USA* **107**, 10815–10820 (2010).
58. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A. L. Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551–1555 (2002).
59. de Nooy, W., Mrvar, A. & Batagelj, V. *Exploratory Social Network Analysis with Pajek* (Cambridge University Press, Cambridge, 2005).
60. Palmer, C. R., Gibbons, P. B. & Faloutsos, C. ANF: a fast and scalable tool for data mining in massive graphs. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 81–90 (New York, NY, USA, 2002).

Acknowledgments

Authors thank J. Demšar, V. Batagelj, M. Žitnik and Z. Levnajić for comments and discussions, and Thomson Reuters for providing the access to bibliographic data. This 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.

Author contributions

L.Š. designed and performed the experiments. L.Š., D.F. and M.B. wrote the main manuscript text. All authors reviewed the manuscript. The authors have no competing financial interests.

Additional information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Šubelj, L., Fiala, D. & Bajec, M. Network-based statistical comparison of citation topology of bibliographic databases. *Sci. Rep.* **4**, 6496; DOI:10.1038/srep06496 (2014).