reliability of **bibliographic databases** for scientometrics network analysis

Lovro Šubelj

University of Ljubljana,

Faculty of Computer and Information Science

ITIS ‘16

acknowledgements

Lovro Šubelj, **Dalibor Fiala** & **Marko Bajec**

Scientific Reports 4, 6496 (2014)

Lovro Šubelj, Marko Bajec, **Biljana M. Boshkoska**, **Andrej Kastrin** & **Zoran Levnajić**

PLoS ONE 10(5), e0127390 (2015)

Lovro Šubelj, **Nees Jan van Eck**, **Ludo Waltman**

PLoS ONE 11(4), e0154404 (2016)

study motivation

- **bibliographic databases** basis for scientific research
- main source of its **evaluation** (citations, *h*-index)
- often studied in **biblio/scientometrics** literature
- different databases give different conclusions (P(*k*))

- databases **differ substantially** between each other
- which bibliographic database is **most reliable**?

bibliographic databases

- scientific bibliographic databases
- **hand-curated** solutions — Web of Science, Scopus
- **automatic** services — Google Scholar, CiteSeer
- **preprint** repositories — arXiv, socArXiv, bioRxiv
- **field-specific** libraries — PubMed, DBLP, APS
- **national** information systems — SICRIS
- and many other

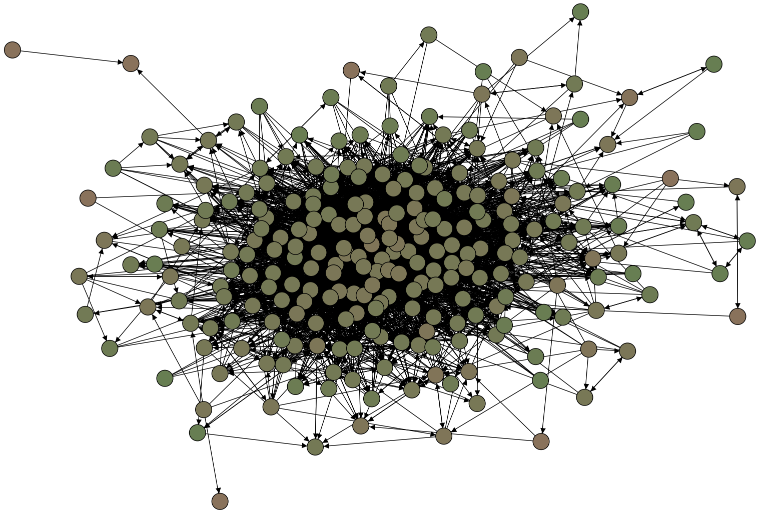
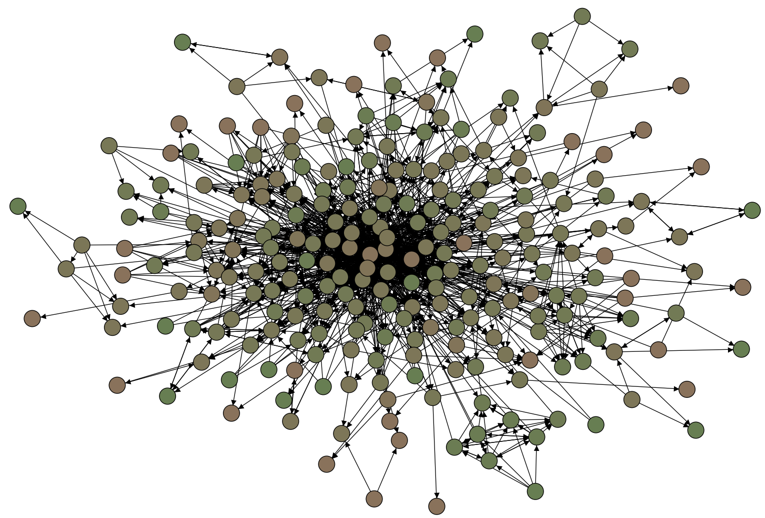
comparisons of databases

- **amount** of literature covered — WoS ≈ Scopus
- **timespan** of literature covered — WoS > Scopus
- available **features** and use in scientific workflow
- data **acquisition** and **maintenance** methodology

- content and structure **differ substantially**
- only informal notions on **reliability**

reliability of databases

- **content** — (amount of) literature covered
- **structure** — accuracy of citation information
- **networks** of **citations** between scientific papers
- **comparison** of **structure** of citation networks

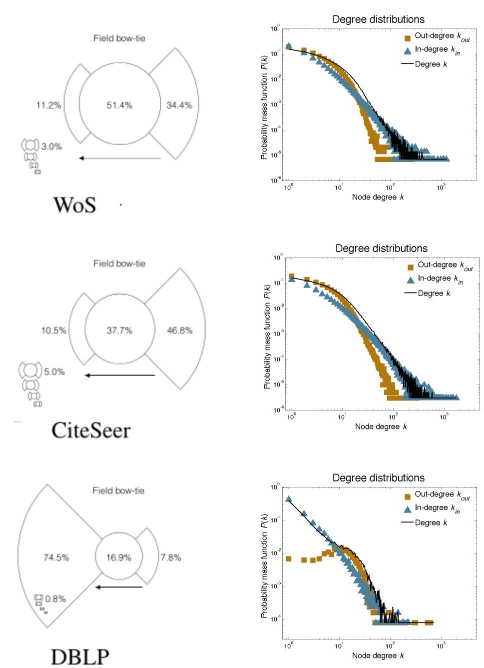
structure of citation networks

- **local/global statistics** of citation networks
- networks mostly **consistent** with **few outliers**
- outliers due to **data acquisition** in most cases

	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%

	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

	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



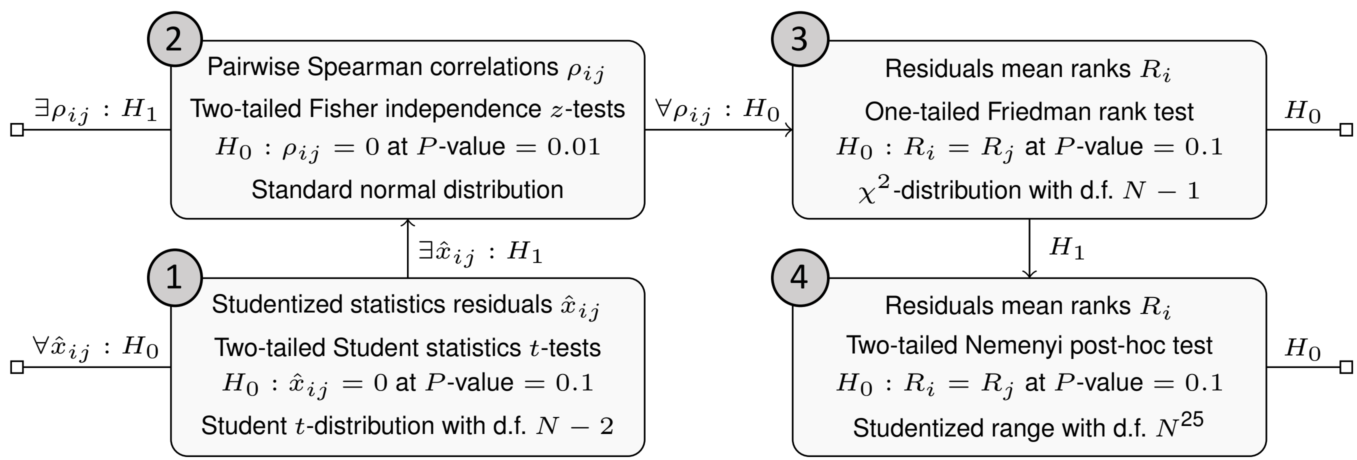
comparison of citation networks

- one can reason only about **individual statistics**
- comparison over **multiple statistics** problematic

- similar problem in machine learning community
- comparison of algorithms over **multiple data sets**
- compare **mean ranks** of algorithms over data sets
- Friedman rank test with Nemenyi post-hoc test

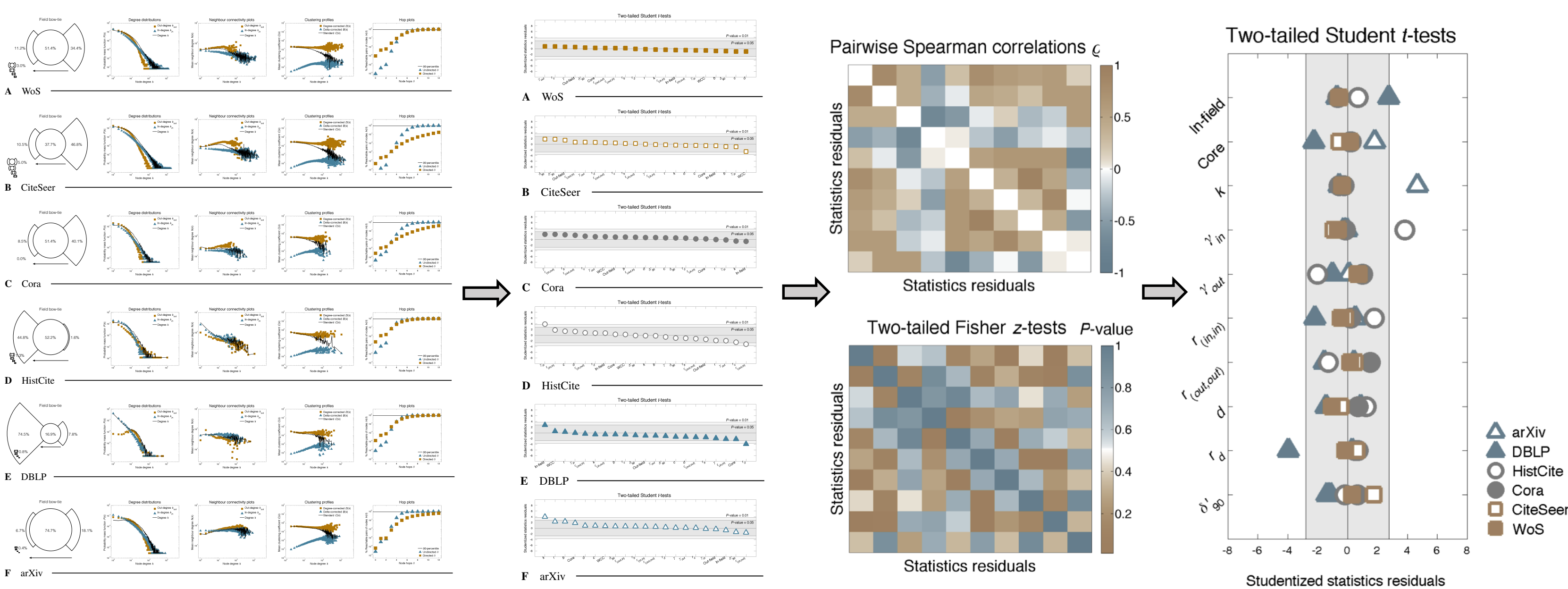
methodology of comparison

- **statistics residuals** since “true network” not known
- database **reliability** seen as **consistency** with rest
- statistics — residuals — independence — **ranks**



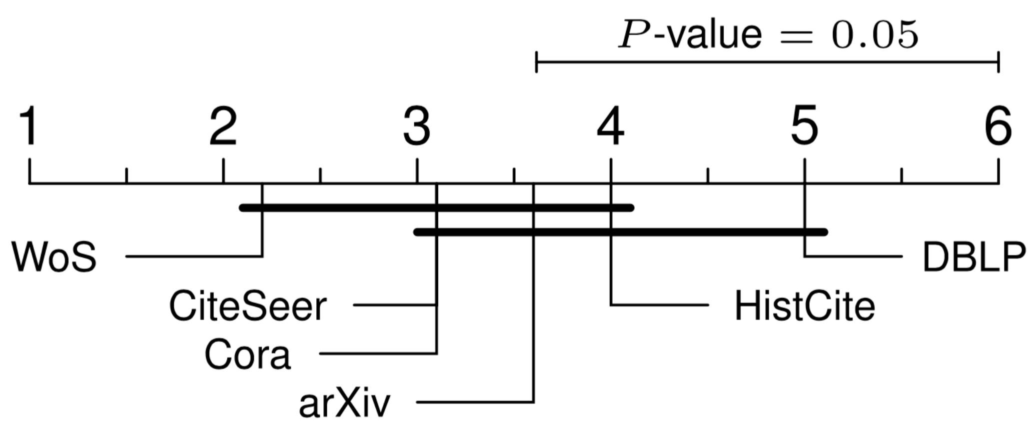
comparison of citation networks

- **statistics** — residuals — independence — **ranks**
- most statistics derived from **node distributions**



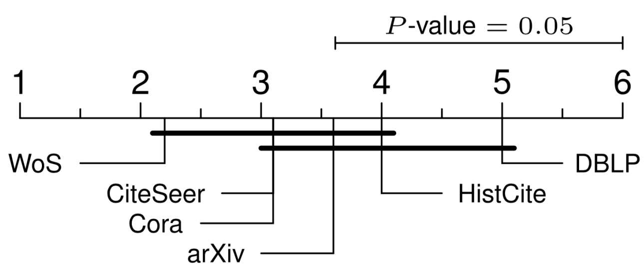
comparison of citation networks

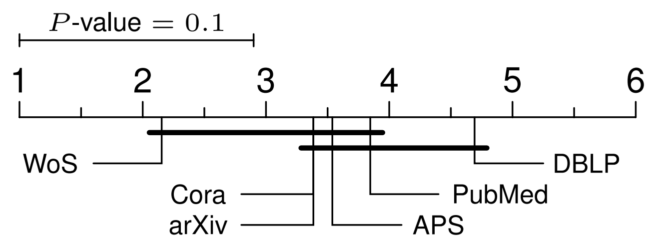
- **mean ranks** of citation networks over statistics
- connected networks are **not significantly different**
- hand-curated **WoS >** field-specific **DBLP**



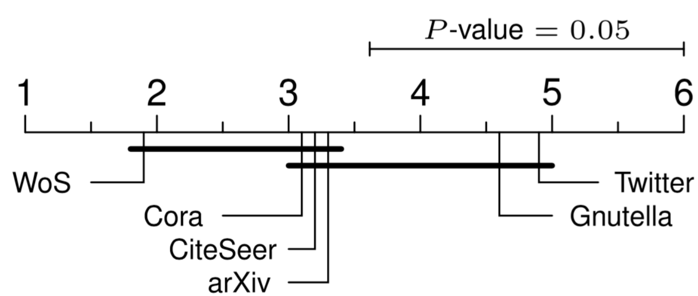
comparison with other networks

- **comparison robust** to selection of networks



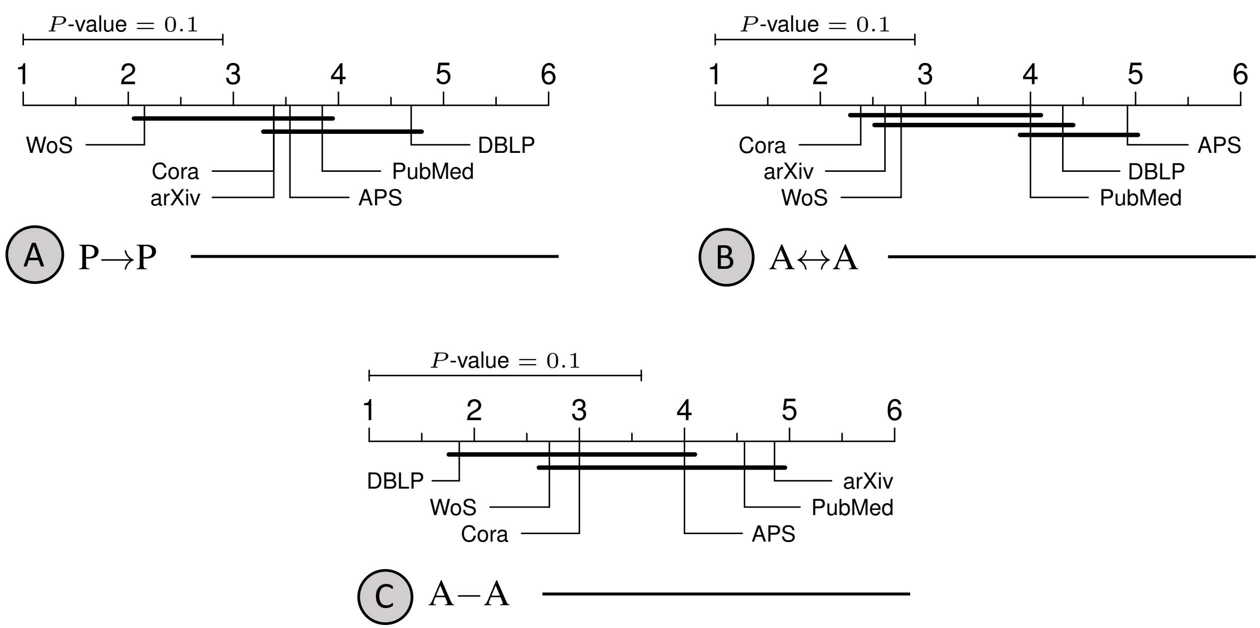


- comparison with **social networks** meaningless
- comparison with other **information networks**



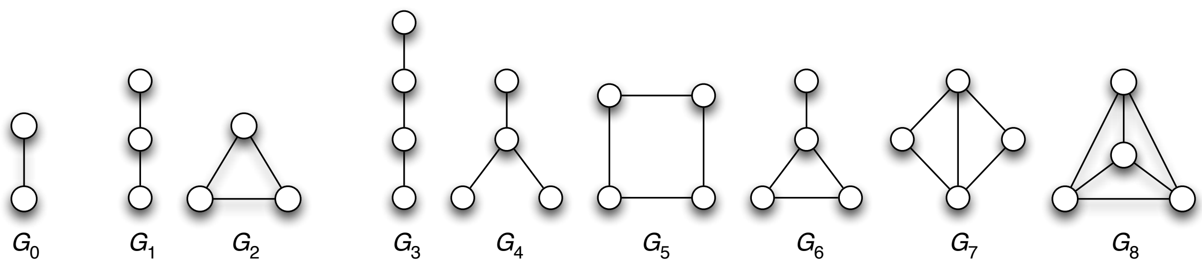
other bibliometric networks

- **A paper citation** information networks
- **C author collaboration** social networks
- **B author citation** social-information networks

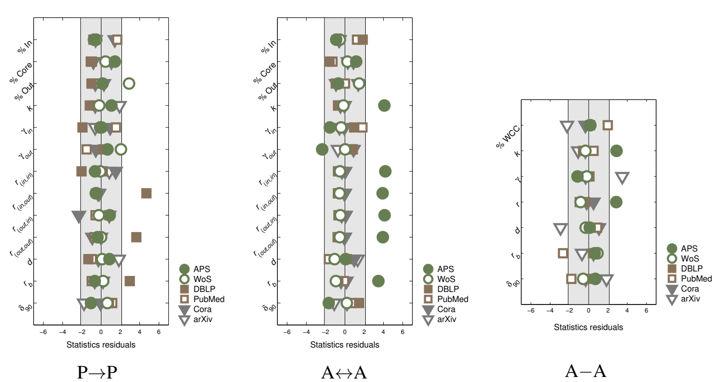


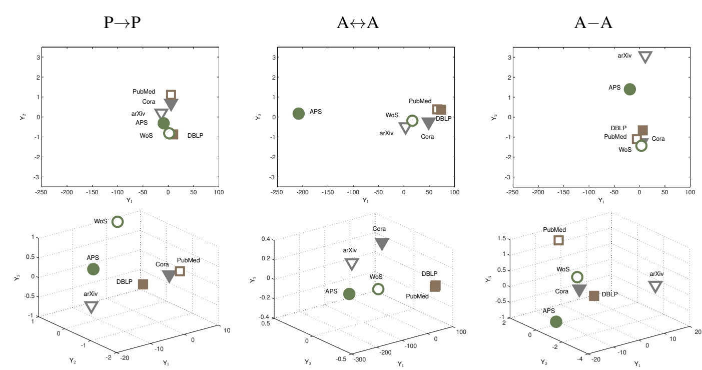
robustness of comparison

- **results robust** to selection of statistics — subgraphs



- results comparable with **other techniques** — MDS





conclusions of comparison

- notable **differences** between databases
- there is **no “best”** bibliographic database
- most appropriate depends on type of analysis
- **hand-curated** databases perform **well overall**
- **field-specific** databases perform **poorly**
- **recipes** for **future** scientometrics studies

- methodology applicable to any network data

identification of research areas

- scientific journals classified in **disciplines**, **fields**
- **research areas** of scientific papers **unknown**

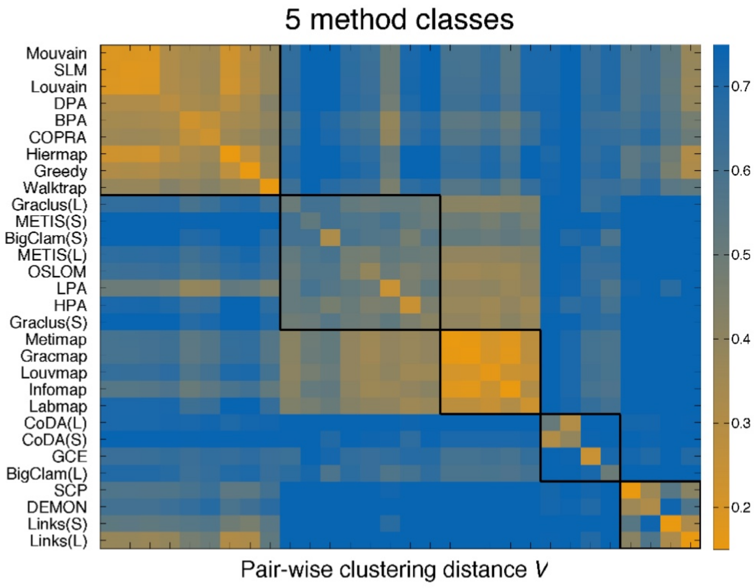
- **clustering papers** based on direct **citation relations**
- graph partitioning/community detection methods
- goal are clusters of **topically related papers**
- clusters **recognizable**, **comprehensible**, robust

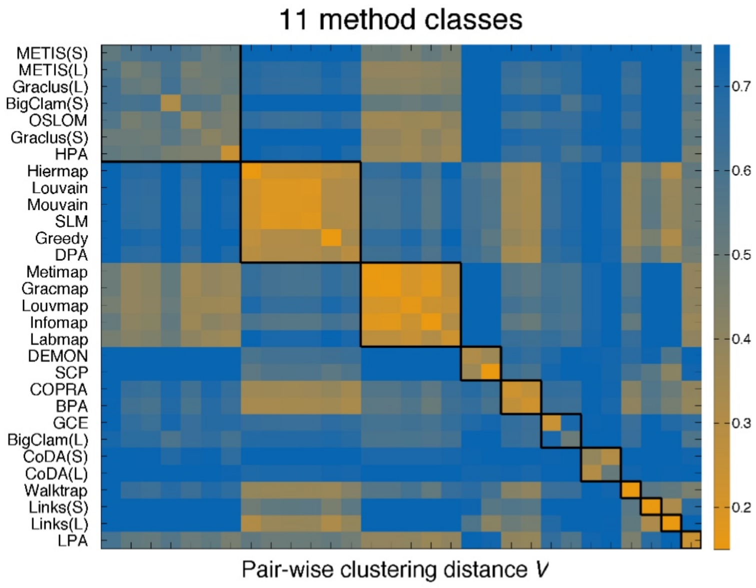
methods for clustering

class	method	description
Spectral analysis	Graclus(S|L)	k-means clustering iteration
	METIS(S|L)	multi-level k-way partitioning
Map equation	Infomap	information flows compression
	Hiermap	hierarchical flows compression
Modularity optimization	Louvain	greedy hierarchical optimization
	Mouvain	multi-level hierarchical optimization
	SLM	smart local moving optimization
Statistical methods	OSLOM	order statistics local optimization method
Label propagation	LPA	label propagation algorithm
	BPA	balanced propagation algorithm
	DPA	diffusion-propagation algorithm
	HPA	hierarchical propagation algorithm
	COPRA	community overlap propagation algorithm
Random walks	Walktrap	random walks hierarchical clustering
Link clustering	Links(S|L)	link similarity hierarchical clustering
Graph models	BigClam(S|L)	cluster affiliation matrix factorization
	CoDA(S|L)	communities through directed affiliations
Ego-networks	DEMON	democratic estimate of modular organization
Cliques	SCP	sequential clique percolation
	GCE	greedy clique expansion
2-step methods	Metilus	METIS+Graclus
	Gracmap	Graclus+Infomap
	Metimap	METIS+Infomap
	Louvmap	Louvain+Infomap
	Labmap	LPA+Infomap

classes of clustering methods

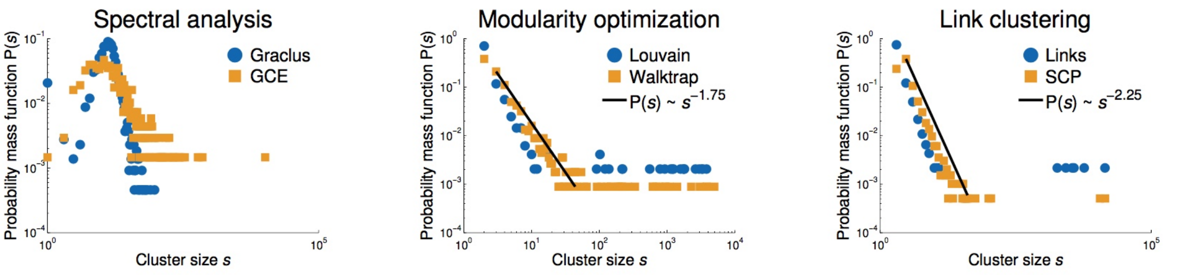
- **distances** between clusterings of methods
- smaller number of **representative methods**

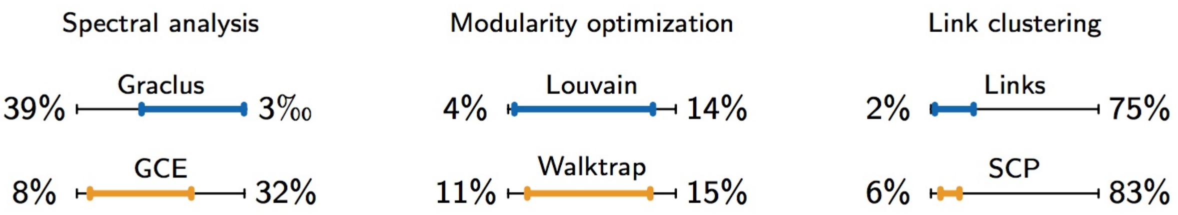




statistical comparison

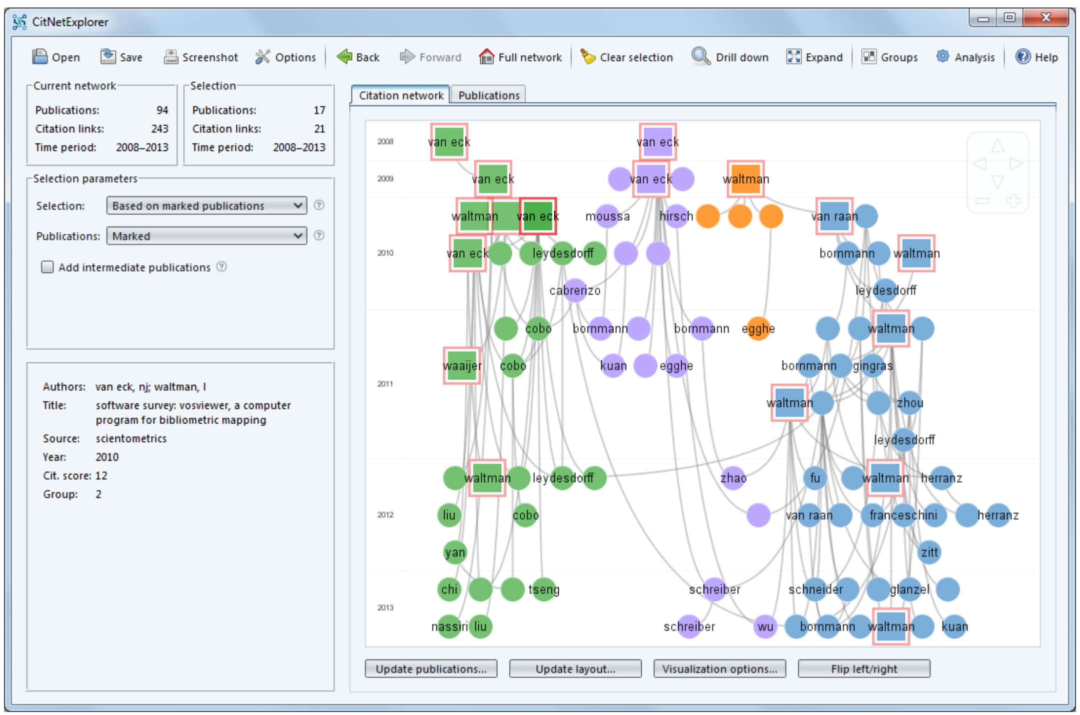
- size **distributions**, degeneracy **diagrams** etc.
- network analysis and bibliometric **metrics**





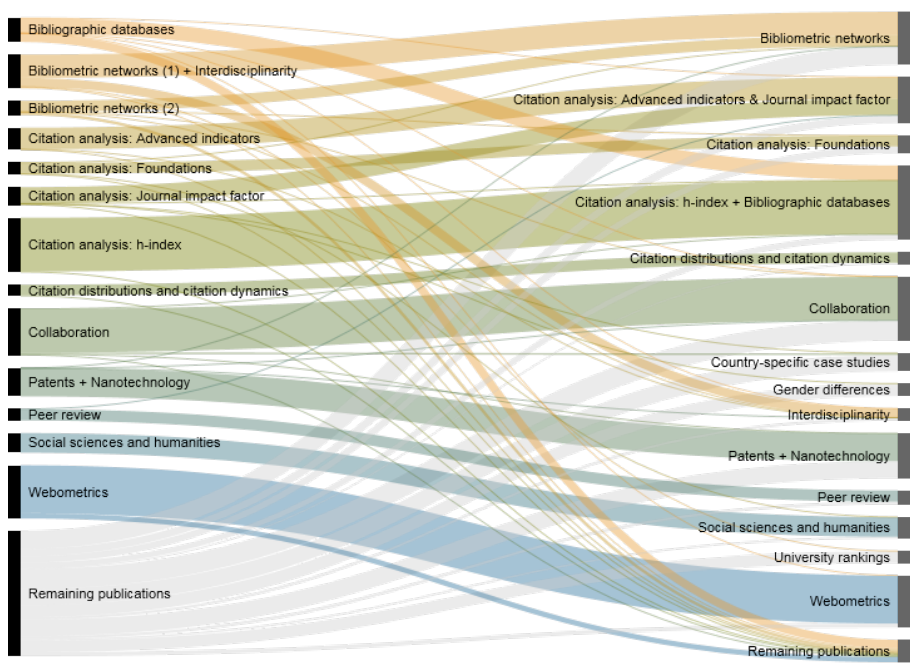
expert assessment tool

- **hands-on assessment** for scientometrics field
- **CitNetExplorer** for analyzing citation networks



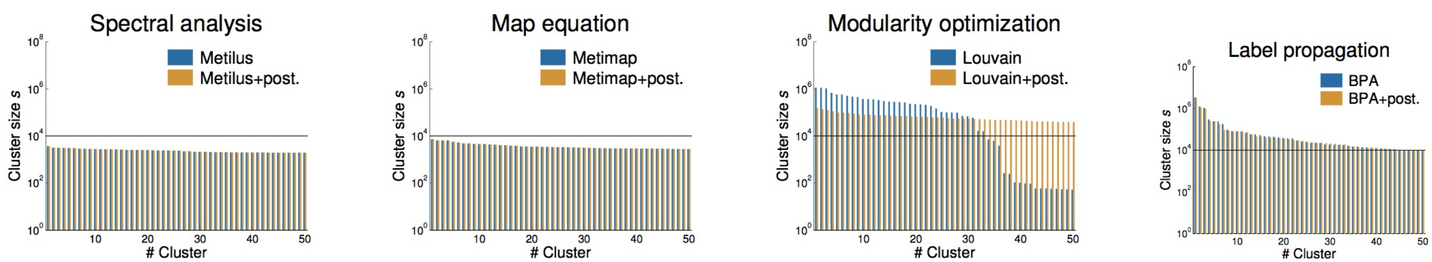
hands-on expert assessment

- **low** resolution — one cluster for **scientometrics**
- **high** resolution — four clusters for ***h*-index papers**
- **topic** resolution — limited number of methods



conclusions of identification

- methods return **substantially different** clusterings
- **no method** performs **satisfactory** by all criteria
- simple **post-processing** performs **poorly**



- **map equation methods** provide good trade-off
- entire science can be clustered in about **one hour**

references

Lovro Šubelj, Dalibor Fiala & Marko Bajec

***Scientific Reports* 4, 6496 (2014)**

Lovro Šubelj, Marko Bajec, Biljana M. Boshkoska,
Andrej Kastrin & Zoran Levnajić

***PLoS ONE* 10(5), e0127390 (2015)**

Lovro Šubelj, Nees Jan van Eck, Ludo Waltman

***PLoS ONE* 11(4), e0154404 (2016)**