# comparison of methods for clustering citation networks

Lovro Šubelj
University of Ljubljana
Faculty of Computer and Information Science

Nees Jan van Eck
Leiden University
Centre for Science and Technology Studies

Ludo Waltman
Leiden University
Centre for Science and Technology Studies

NetSci-X '16

# study overview

## problem

grouping publications into clusters based on citation relations

## means

graph partitioning/community detection methods on citation networks

## goals

clusters of topically related publications or research areas

## wishes

experts should recognize cluster topics

small differences in cluster sizes

limited number of tiny clusters

robustness to small perturbations

reasonable computational complexity

# citation networks

## data

in-house version of Web of Science database of CWTS

## networks

citation networks represented as simple undirected graphs

| field | period | # publications | # nodes | # links |
|---|---|---|---|---|
| Scientometrics | 2009-2013 | 2,402 | 1,998 | 5,496 |
| L&IS | 1996-2013 | 43,741 | 32,628 | 131,989 |
| Physics | 2004-2013 | 1,314,458 | 1,233,542 | 9,838,008 |
| WoS | 2004-2013 | 11,780,132 | 11,063,916 | 122,148,955 |

Scientometrics — journals Journal of Informetrics, Scientometrics and JASIST

L&IS — Information Science & Library Science journal subject category

Physics — eight Physics journal subject categories and Astronomy & Astrophysics

WoS — all journal subject categories in Web of Science

# methods

## 30 basic/derived graph partitioning/community detection methods

| 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 |

2-step — second method applied to clusters obtained by first method

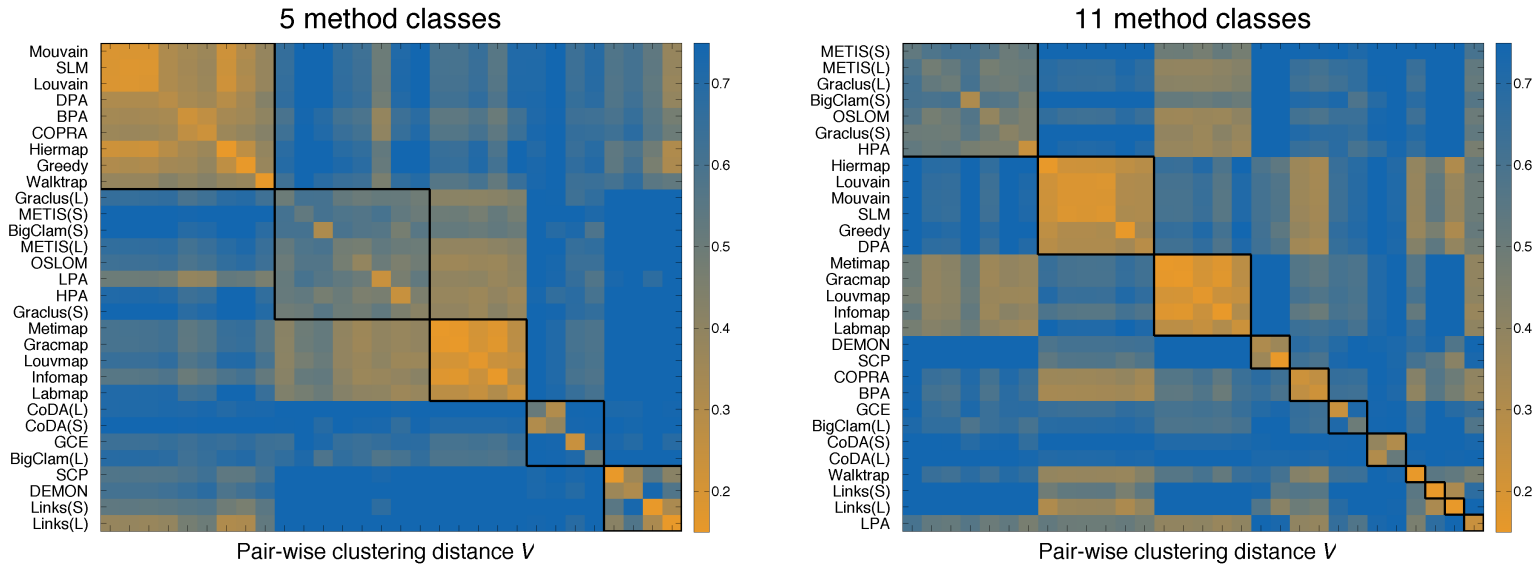
S|L — small|large clusters

# clustering distances

## clusterings

distances between clusterings by considered methods

10/15 selected representative methods



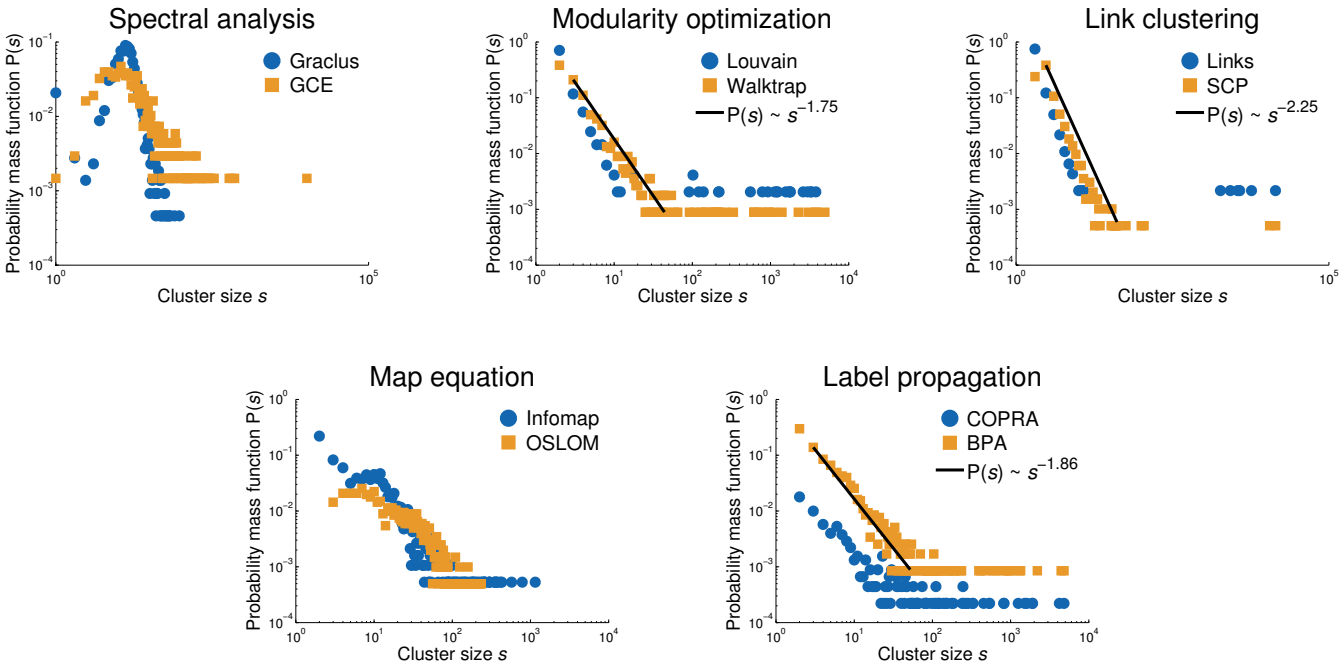

distance — normalized variation of information of clusterings

# clustering distributions

## sizes

size distributions of clusterings by representative methods

from homogeneous to inhomogeneous distributions

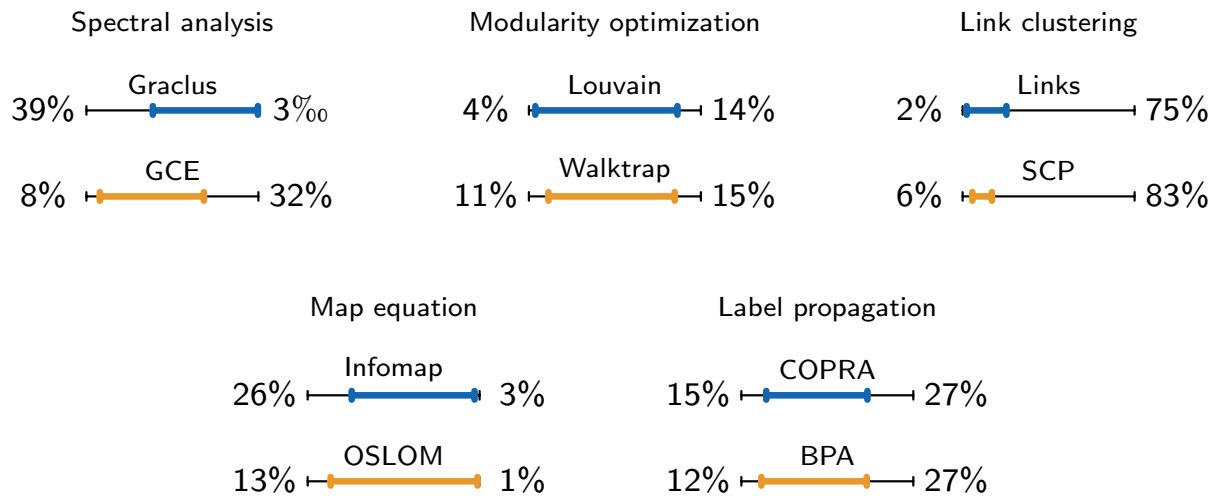


# clustering degeneracy

## ranges

degeneracy diagrams of clusterings by representative methods

narrowing effective ranges from left to right




left-hand side — % nodes in tiny clusters $< 15$ nodes

right-hand side — % nodes in largest cluster

## metrics

standard metrics of clusterings by representative methods

≈ 1500 clusters and decreasing Flake score from top/bottom

| method | # clusters | degree | expansion | Flake | modularity |
|---|---|---|---|---|---|
| Graclus | 2175 | 2.4 | 5.8 | 52% | 0.29 |
| OSLOM | 1914 | 3.8 | 4.4 | 37% | 0.45 |
| Infomap | 1871 | 5.0 | 3.2 | 19% | 0.60 |
| Louvain | 488 | 6.8 | 1.2 | 3% | 0.73 |
| Walktrap | 1127 | 6.5 | 1.6 | 7% | 0.69 |
| BPA | 1002 | 7.0 | 1.0 | 3% | 0.66 |
| COPRA | 3826 | 6.8 | 1.2 | 15% | 0.65 |
| Links | 2933 | 6.4 | 1.8 | 20% | 0.09 |
| SCP | 1969 | 4.9 | 3.2 | 37% | 0.22 |
| GCE | 682 | 4.1 | 4.0 | 29% | 0.43 |

degree — average node intra-cluster or internal degree

expansion — average node inter-cluster or external degree

Flake — % nodes with larger external than internal degree

# bibmetrics

bibliometric metrics of clusterings by representative methods

orders $\gg 1$ and increasing coverage from top/bottom

| method | size | orders | diameter | coverage | uncertainty |
|---|---|---|---|---|---|
| Graclus | 15.0 | 1.1 | 3.4 | 29% | 0.42 |
| OSLOM | 16.0 | 2.6 | 4.8 | 46% | 0.36 |
| Infomap | 17.3 | 2.7 | 4.3 | 62% | 0.13 |
| Louvain | 66.7 | 3.3 | 9.1 | 85% | 0.19 |
| Walktrap | 29.0 | 3.4 | 7.8 | 80% | 0.00 |
| BPA | 32.0 | 3.6 | 7.3 | 86% | 0.21 |
| COPRA | 8.8 | 4.0 | 6.9 | 85% | 0.22 |
| Links | 10.1 | 4.3 | 11.1 | 78% | 0.05 |
| SCP | 16.6 | 4.2 | 23.1 | 61% | 0.02 |
| GCE | 47.8 | 3.3 | 12.0 | 50% | 0.24 |

orders — orders of magnitude spanned by cluster sizes

diameter — average within cluster effective diameter

uncertainty — variation of information of clusterings

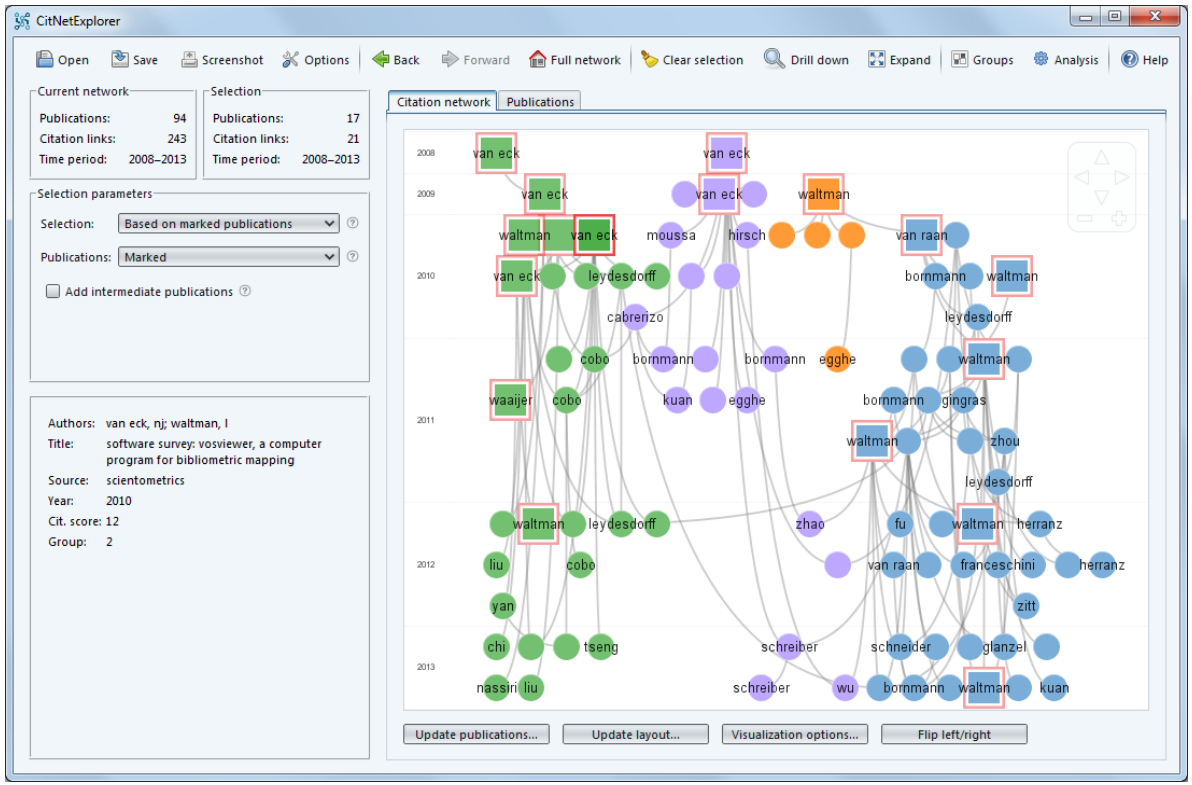
coverage — % links covered by clusters

# clustering tool

## assessment tool

CitNetExplorer for analyzing citation networks

freely available at www.citnetexplorer.nl



# clustering resolution

clusterings for L&IS by representative methods

hands-on expert assessment for scientometrics using CitNetExplorer

## low resolution

Walktrap and BPA

BPA returns one cluster covering scientometrics

## high resolution

Graclus(S|L) and METIS(S|L)

Graclus returns four clusters covering h-index

## topics resolution

OSLOM, Louvain, Metimap and Infomap

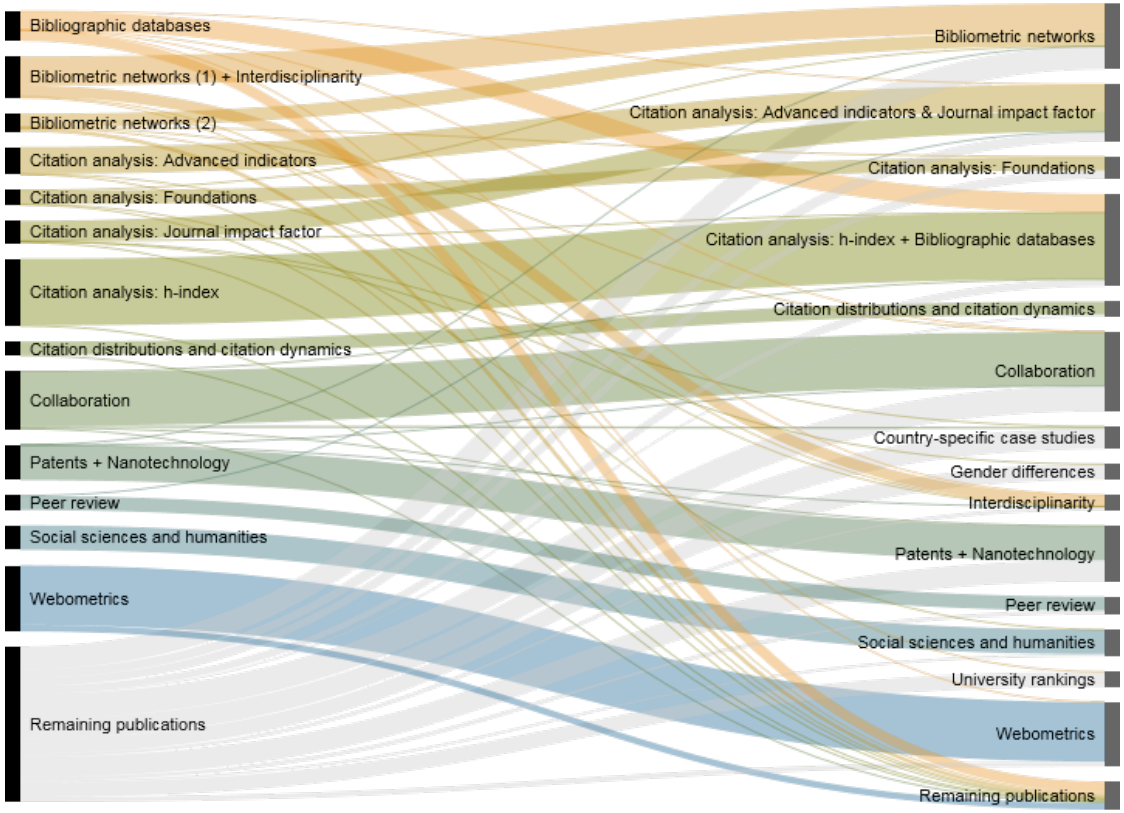
OSLOM, Louvain return ambigous/heterogeneous clusters

# clustering assessment

## expert assessment

largest scientometrics clusters by Metimap and Infomap methods

identified research topics of clusters covering ≈ 75% publications



# clustering WoS

## clustering metrics for WoS by fastest methods

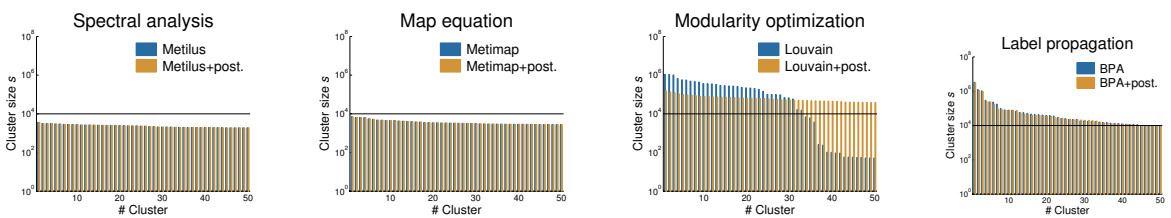
| method | size | orders | degree | coverage | Flake | complexity |
|---|---|---|---|---|---|---|
| Metilus | 50.0 | 2.3 | 5.9 | 27% | 69% | 30 min |
| Metimap | 33.2 | 3.6 | 10.3 | 47% | 45% | 94 min |
| Louvain | 334.4 | 5.7 | 18.5 | 84% | 5% | 52 min |
| BPA | 105.4 | 6.2 | 18.5 | 84% | 7% | 66 min |

# post-processing

## tiny clusters $< 15$ nodes merged by maximizing likelihood

| method | size | orders | degree | coverage | Flake | complexity |
|---|---|---|---|---|---|---|
| Metilus+post. | 51.5 | 2.2 | 5.9 | 27% | 69% | 34 min |
| Metimap+post. | 58.9 | 3.6 | 10.3 | 47% | 45% | 99 min |
| Louvain+post. | 320.9 | 4.9 | 15.2 | 69% | 17% | 79 min |
| BPA+post. | 167.1 | 6.2 | 18.0 | 82% | 9% | 114 min |

## giant clusters $> 10^4$ nodes repartitioned by same method



# study summary

## conclusions

methods return substantially different clusterings

no method performs satisfactory by all criteria

straightforward post-processing performs poorly

map equation methods provide good trade-off

## limitations

limitations of expert assessment of clusterings

limited number of methods with default parameters

no directed, overlapping, multi-resolution, principled methods

no equivalence clusters or co-citation and bibliographic coupling

# arXiv:1512.09023


Lovro Šubelj
University of Ljubljana
lovro.subelj@fri.uni-lj.si
lovro.lpt.fri.uni-lj.si

Nees Jan van Eck
Leiden University
ecknjpvan@cwts.leidenuniv.nl
www.neesjanvaneck.nl

Ludo Waltman
Leiden University
waltmanlr@cwts.leidenuniv.nl
www.ludowaltman.nl