network analysis in bibliometrics

Lovro Šubelj

University of Ljubljana,

Faculty of Computer and Information Science

CWTS ‘17

Slovenia “chicken”

University of **Ljubljana**

- **since 1919** *271st in CWTS Leiden Ranking 2017*
- **26 members** *23 faculties & 3 academies*
- **40,110 students** & 5,730 staff *in 2016*

Faculty of **Computer and Information Science**

- **since 1996** *cs study since 1973*
- **≈1,300 students** & ≈180 staff
- **BSc, MSc, PhD** *cs, prog, math, mm*
- **research** *cs, db, is, dm, ml, ai, nets*

networks courses

University of Ljubljana
Faculty of Computer and Information Science

elective courses on NETWORKS in 2017/18

Networks or graphs are ubiquitous in everyday life. Examples include online social networks, the Web, references between WikiLeaks cables, Supervizor, terrorist affiliations, LPP bus map, plumbing systems and your brain. Many such real networks reveal characteristic patterns of connectedness that are far from regular or random. Networks have thus been a prominent tool for investigating real-world systems since the 18th century. However, while small networks can be drawn by hand and analyzed by a naked eye, real networks require specialized computer algorithms, techniques and models. This led to the emergence of a new scientific field about 20 years ago...

INA
Introduction to Network Analysis

Network analysis concepts and techniques

Course code **63545B** | eUcilnica **#183**
MSc students | Lecturer Lovro Šubelj
Summer semester | Starts **Feb 19, 2018**
Introductory 15 week course to get started

ASI
Social and Information Network Analysis

Modern analysis of large-scale real networks

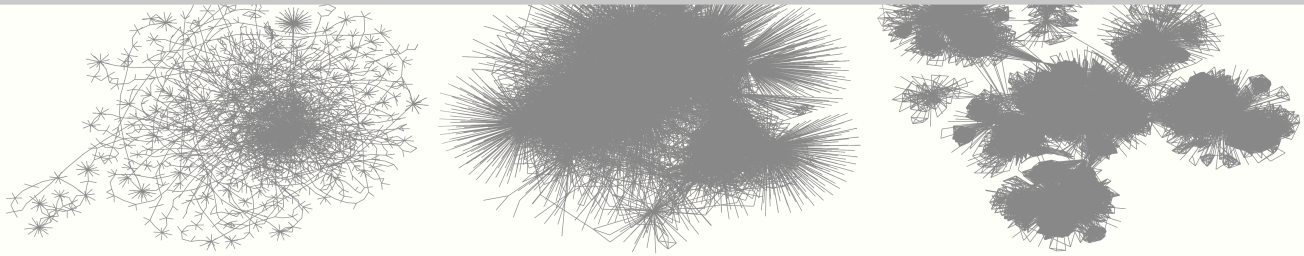
Course code **63225B** | eUcilnica **#66**
MSc students | Lecturer Jure Leskovec
Winter semester | Starts **Sep 25, 2017**
Fast pace 10 week course by a leading expert

ANTS
Advanced Topics in Network Science

Thorough review of modern network science

Course code **63835A** | eUcilnica **#170**
PhD students | Lecturers Šubelj & Leskovec
Summer semester | Starts **Feb 26, 2018**
Research-oriented 12 week course & invited talks

Course enrollment is **not** possible in order **from right to left** | lovro.subelj@fri.uni-lj.si



talk **outline**

1. **reliability of bibliographic databases**
   Šubelj, L., Fiala, D., & Bajec, M. (2014). Scientific Reports, 4, 6496.
   Šubelj, L., Bajec, M., Boshkoska, B. M., et al. (2015). PLoS ONE, 10(5), e0127390.

2. modeling paper citation networks
   Šubelj, L., & Bajec, M. (2013). In Proceedings of the LSNA '13, pp. 527–530.
   Šubelj, L., Žitnik, S., & Bajec, M. (2014). In Proceedings of the NetSci '14, p. 1.

3. clustering paper citation networks
   Šubelj, L., Van Eck, N. J., & Waltman, L. (2016).PLoS ONE, 11(4), e0154404.

bibliographic databases **reliability**

- databases basis for research & **evaluation**

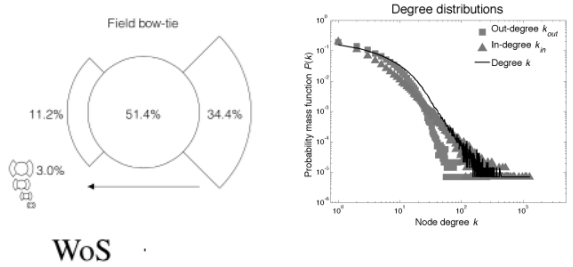
- databases can **differ substantially**
  different databases often give quite different conclusions
- content & **structure** can differ substantially
  coverage, timespan, features, accuracy, acquisition etc.

- only informal notions on their **reliability**
  particular case of reliability of structure of citation networks

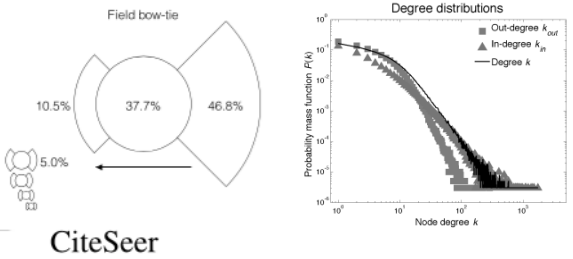
structure of **citation networks**

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

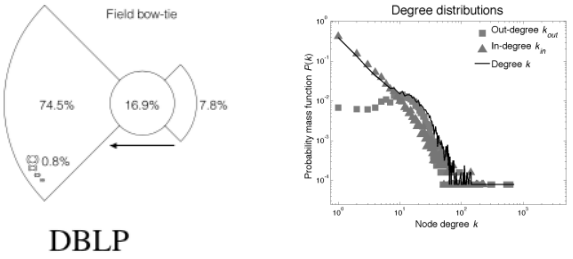
- comparison over **one statistic**
- comparison over **many statistics?**
  same problem in machine learning community



WoS



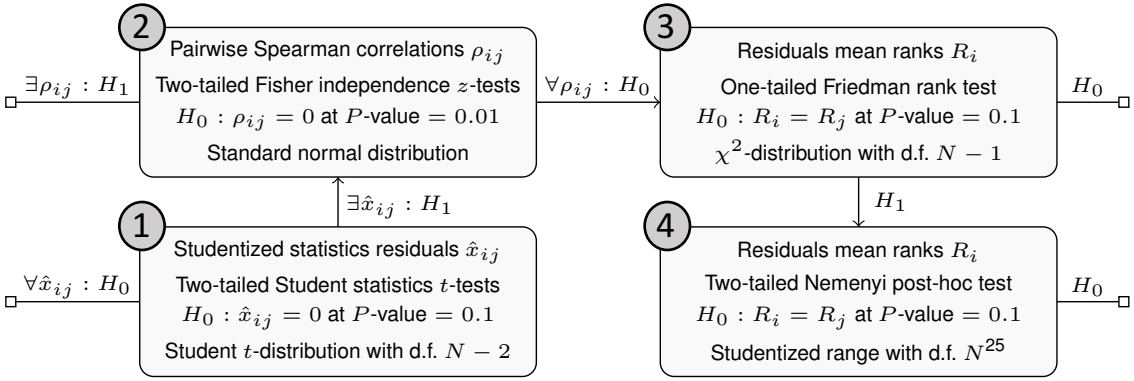
CiteSeer



DBLP

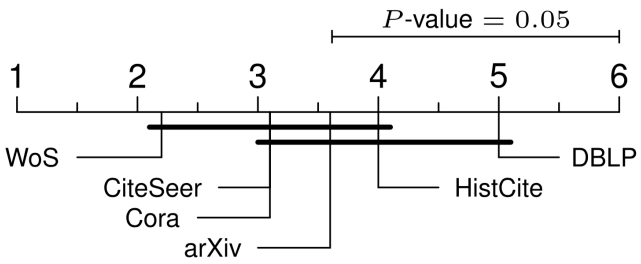
methodology of database comparison

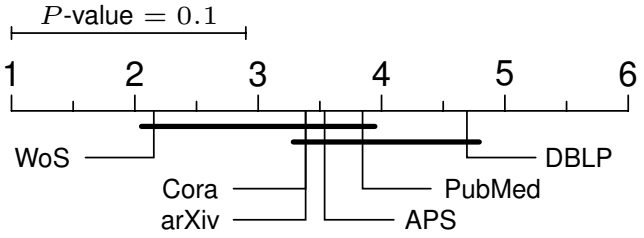
- network **statistics** — residuals — database rank
- mean **ranks** of databases over many statistics
- **residuals** since “true database” is not known
  database reliability seen as consistency with other databases



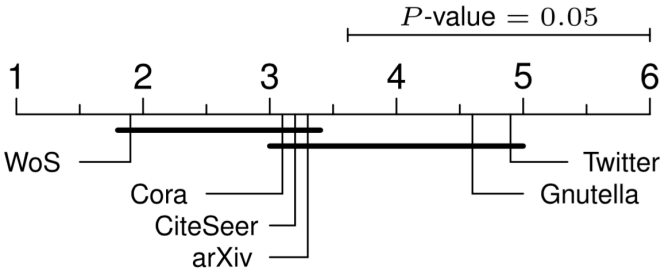
comparison of **citation networks**

- comparison of different **citation networks**
  results robust to selection of networks, statistics, patterns etc.



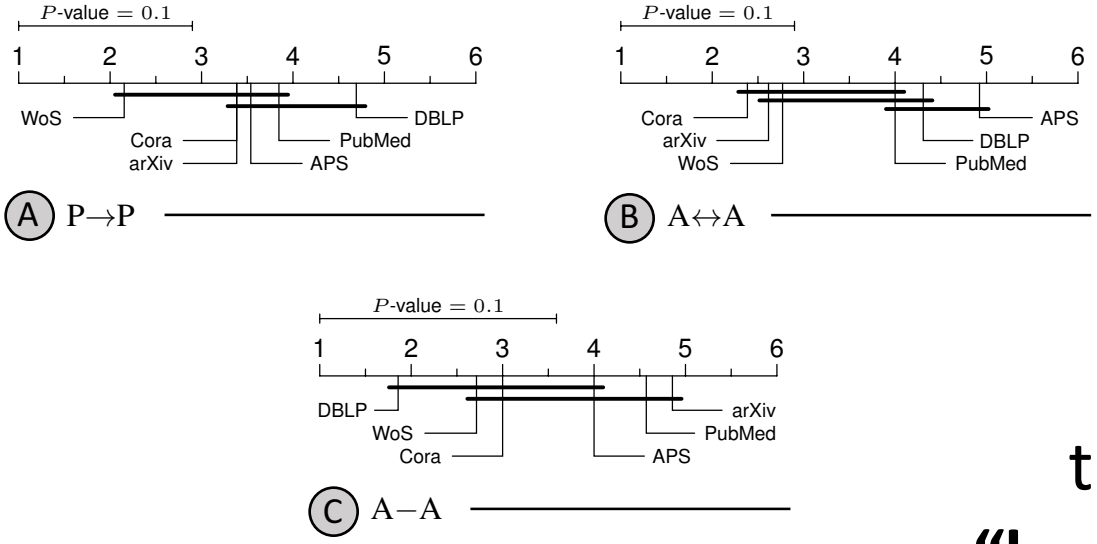


- comparison of different **information networks**



comparison of **bibliographic networks**

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



there is **no** **“best”** database!

talk **outline**

1. reliability of bibliographic databases
   Šubelj, L., Fiala, D., & Bajec, M. (2014). Scientific Reports, 4, 6496.
   Šubelj, L., Bajec, M., Boshkoska, B. M., et al. (2015). PLoS ONE, 10(5), e0127390.

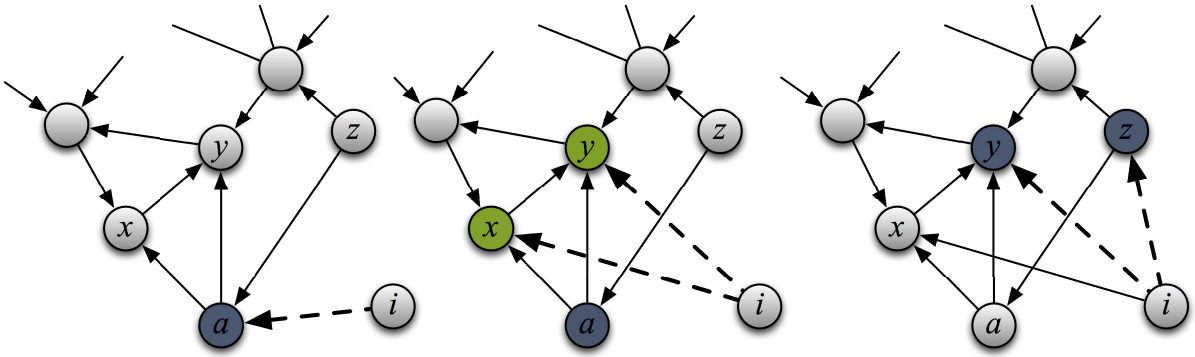
2. **modeling paper citation networks**
   Šubelj, L., & Bajec, M. (2013). In Proceedings of the LSNA '13, pp. 527–530.
   Šubelj, L., Žitnik, S., & Bajec, M. (2014). In Proceedings of the NetSci '14, p. 1.

3. clustering paper citation networks
   Šubelj, L., Van Eck, N. J., & Waltman, L. (2016).PLoS ONE, 11(4), e0154404.

models of **citation networks**

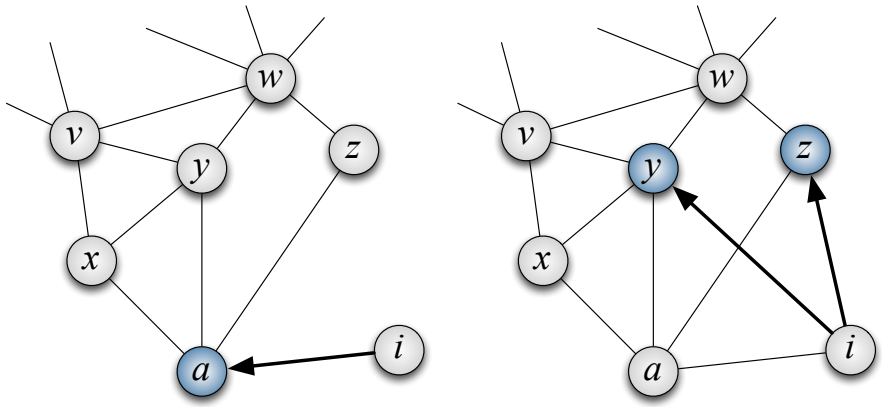
- **generative models** of citation networks
  to reason about structure, evolution, dynamics, future etc.

- many possible **applications** in bibliometrics



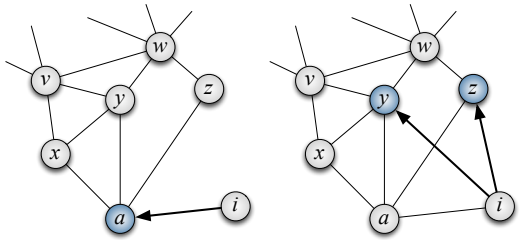
forest fire network model

- each new **node *i* forms links** as follows
  1. *i* selects initial **ambassador *a*** and **links to *a***
  2. *i* selects its **neighbors *y*, *z*** and **links to *y*, *z***
  3. *y*, *z* are taken as **new ambassadors** of *i*



forest fire **citation model**

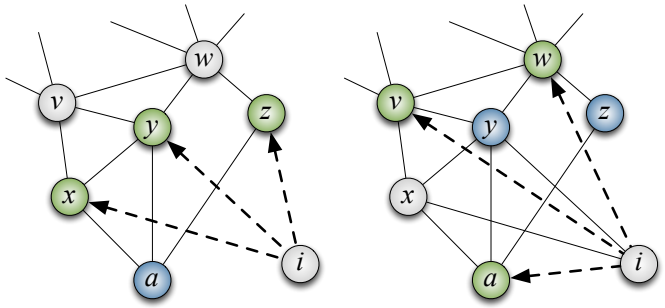
- each new **paper *i* cites** as follows
  1. *i* selects initial **paper *a*** and **cites *a***
  2. *i* selects its **references *y*, *z*** and **cites *y*, *z***
  3. *y*, *z* are taken as **new reading** for *i*



- then authors **read all cited** papers and vice-versa
- only **≈20% references read** (Simkin & Roychowdhury, 2003)

realistic **citation model**

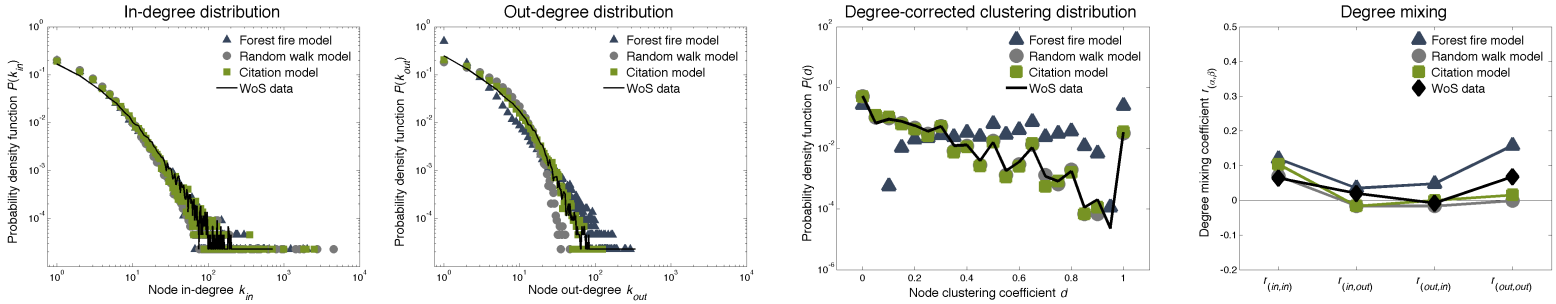
- each new **paper *i* cites** as follows
  1. *i* selects initial **paper *a*** and **can cite *a***
  2. *i* selects its **references *y*, *z*** and **can cite *y*, *z***
  3. some **references** are taken as **new reading** for *i*

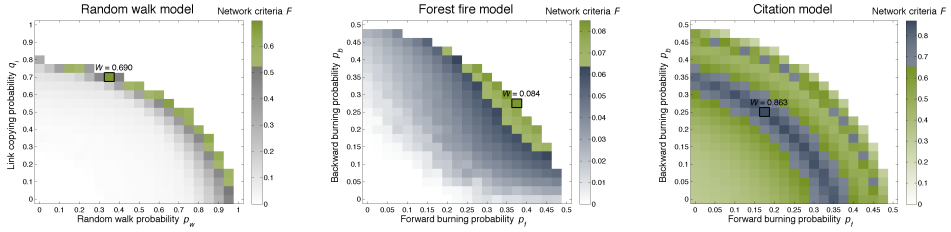


- read & cited papers **modeled independently**

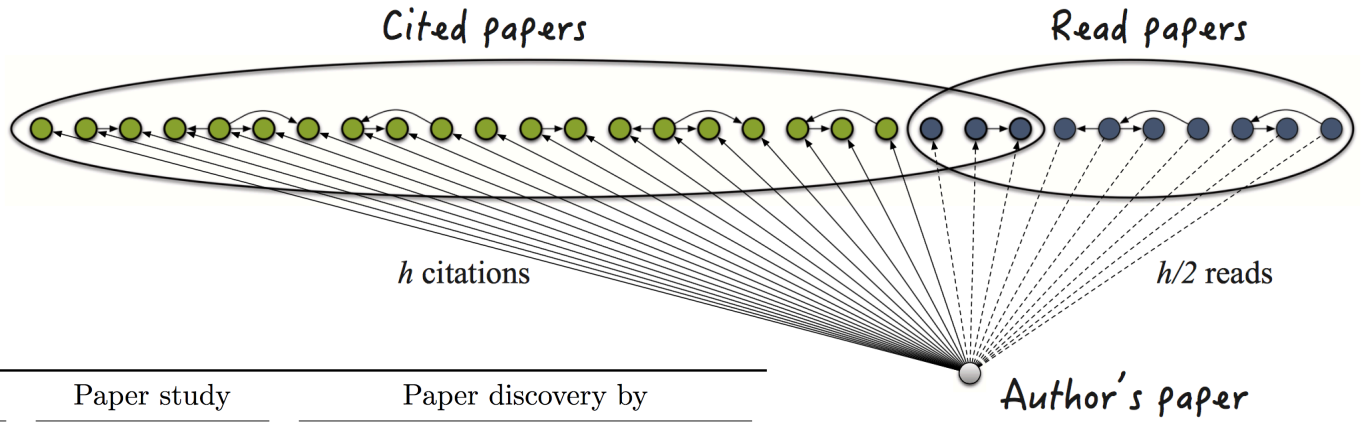
directed **citation model**

- **directed dynamics** much more complicated
- model **reproduces WoS** citation networks
- **clear optima** (peak) in model parameters





implications of citation model



Data	Paper citation		Paper study		Paper discovery by		
	# Cite	% Copy	# Read	% Cite	% Citation	% Service	% Other
ILS	3.98	86.1%	2.14	27.9%	29.2%	41.0%	29.8%
TM	2.93	79.7%	1.47	45.2%	74.7%	0.5%	24.9%
AI	4.52	87.3%	1.47	40.9%	25.8%	47.6%	26.6%
SE	2.78	81.5%	1.58	36.4%	68.8%	2.0%	29.2%
CY	2.18	69.6%	1.59	43.2%	24.5%	37.8%	37.6%

Period	Paper citation		Paper study		Paper discovery by		
	# Cite	% Copy	# Read	% Cite	% Citation	% Service	% Other
1945–2013	3.98	86.1%	2.14	27.9%	29.2%	41.0%	29.8%
1970–1980	2.23	52.1%	3.39	33.5%	41.4%	0.0%	58.5%
1980–1990	2.62	65.1%	2.96	33.0%	48.3%	1.1%	50.6%
1990–2000	3.42	81.6%	2.38	29.0%	40.3%	23.2%	36.5%
2000–2010	5.06	83.6%	2.90	32.2%	40.7%	27.5%	31.7%

one read paper ≈ ~~**five**~~ **two cited** papers!

talk **outline**

1. reliability of bibliographic databases
   Šubelj, L., Fiala, D., & Bajec, M. (2014). Scientific Reports, 4, 6496.
   Šubelj, L., Bajec, M., Boshkoska, B. M., et al. (2015). PLoS ONE, 10(5), e0127390.

2. modeling paper citation networks
   Šubelj, L., & Bajec, M. (2013). In Proceedings of the LSNA '13, pp. 527–530.
   Šubelj, L., Žitnik, S., & Bajec, M. (2014). In Proceedings of the NetSci '14, p. 1.

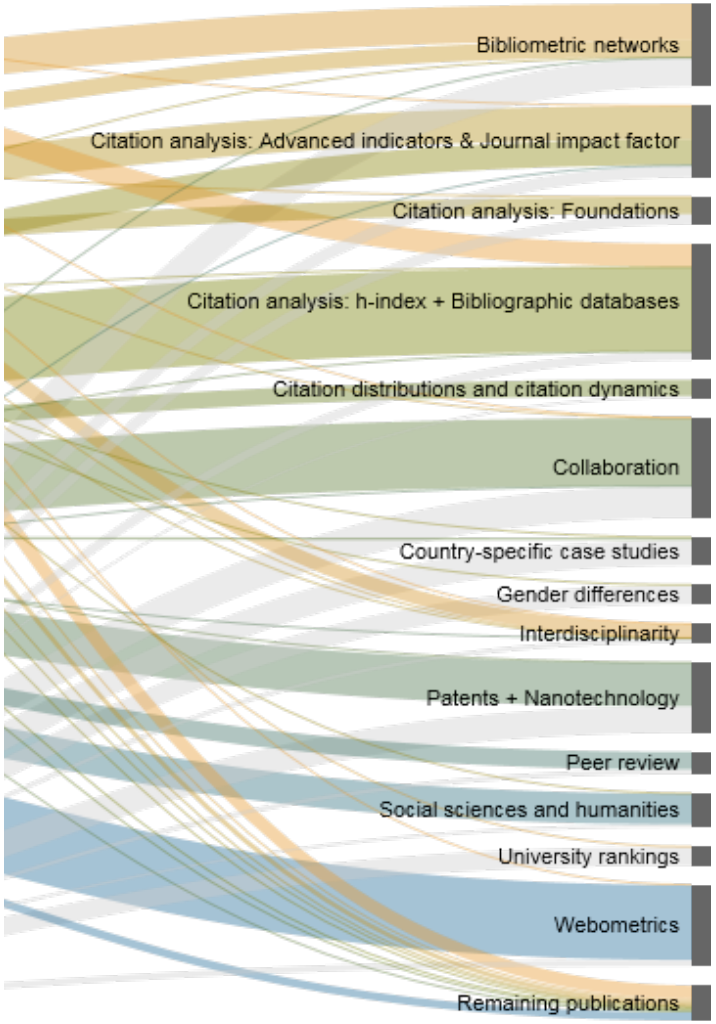
3. **clustering paper citation networks**
   Šubelj, L., Van Eck, N. J., & Waltman, L. (2016).PLoS ONE, 11(4), e0154404.

clustering citation networks

- **clustering papers** based on direct citation relations
  research areas or topics of papers

- **systematic comparison** of large number of methods
  network clustering and partitioning

there is **no**
“best” method!



thank you!

network convexity

LCN2 seminar next Friday at 4pm in Snellius