Same Data, Different Results
Comparing Topic Extraction Solutions

Theresa Velden
University of Michigan, School of Information
(now: ZTG, TU Berlin)

Workshop on Identification, location and temporal evolution of topics
Budapest, 29/30 August 2016, Hungarian Academy of the Sciences

Topic Extraction Challenge *(Beta Version)*

- How to group publications algorithmically into topics?
- Method: Collaboration of several scientometric groups
- Data set: metadata of ~ 111.000 publications from 59 journals in astronomy and astrophysics 2003-2010 (Web of Science) —> ***AstroData***

Kevin Boyack (SciTech Strategies) · Nees van Eck (CWTS Leiden)· Wolfgang Glänzel & Bart Thijs (ECOOM) · Jochen Gläser (TU Berlin) · Frank Havemann & Michael Heinz (HU Berlin) · Rob Koopman & Shenghui Wang (OCLC Research), Andrea Scharnhorst (DANS-KNAW), Theresa Velden (UMSI)

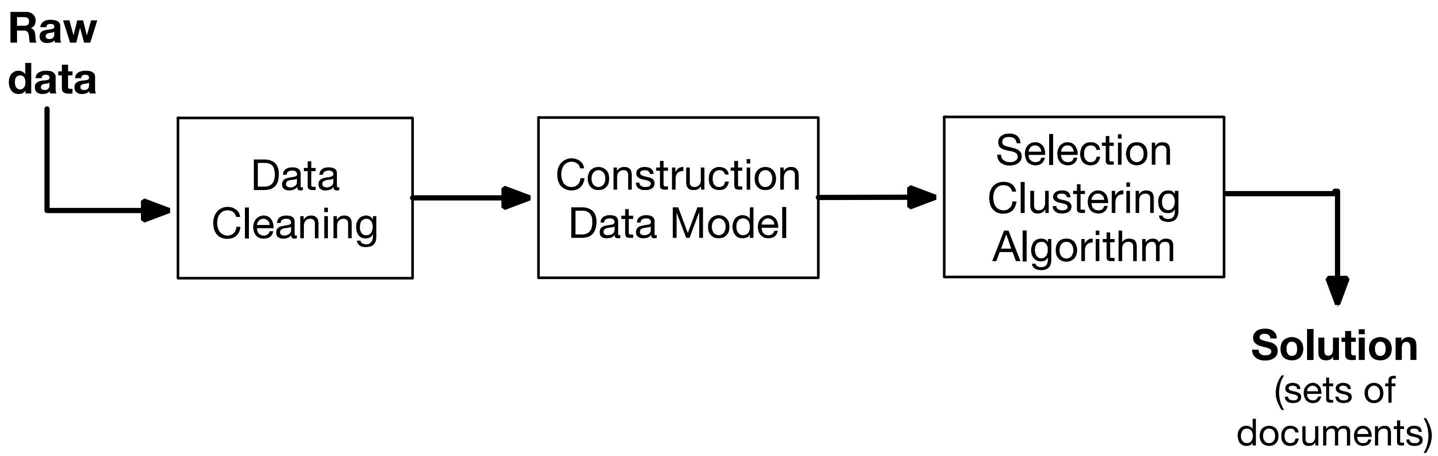
Same Data, Different Results
Problem Statement

- Need for ‘benchmarking’ of topic extraction approaches
  - Often developed and fine-tuned in-house with lack of replication
  - Usually data set not available for replication
  - Origin and scale of differences in results unclear
- Lack of ground truth
  - Depending on perspective more than one valid thematic structure can be constructed
  - Topical structures are reconstructed for specific purposes, so if at all, best method for a given purpose

→ Aim: Instead of finding best solution, aim at uncovering how results differ and how those differences relate to approaches

Secondary aim: Developing methods for comparison

Topic Extraction Workflow
& Sources of Variance



Label (short)	Coverage articles (%)	Data Model	Clustering	Parameters	Results # of topics
CWTS -C5 (**c**)	$101,828$ (91.23%)	direct citation giant component[2]	Smart Local Moving Algorithm (SLMA)	resolution min. cluster size	22
UMSI0 (u)	$101,831$ (91.23%)	direct citation giant component	Infomap (undirected)	random seed	22
OCLC-Louvain (ol)	$111,616$ 100%	semantic matrix[3]	Louvain (python library networkX)	word occurrence thresh. stopword list K most similar articles similarity value thresh.	32
OCLC-31 (ok)	$111,616$ (100%)	semantic matrix	k-means (python library sklearn.cluster. MiniBatchKMeans)	word occurrence thresh. stopword list number of clusters	31
ECOOM -BC13 (eb)	108,512 (97.22%)	bibliographic coupling	Louvain (pajek)	references < 11 years if indexed in TR product resolution link strength threshold	13
ECOOM -HY11 (eh)	109,376 (97.99%)	bibliographic & lexical coupling (NLP)	Louvain (pajek)	resolution link strength thresh. weight bc vs. text	11
STS-RG (sr)	107,304 (96.14%)	direct citation incl. non-source items cited at least twice	projection onto global science map (1996-2012) clustered by SLMA	resolution min. cluster size	555
HU-DC (hd)	101,762 (91,17%)	direct citation giant component	Memetic (random evolution + deterministic search)	seeds resolution population size other evolution param.	111 over-lapping

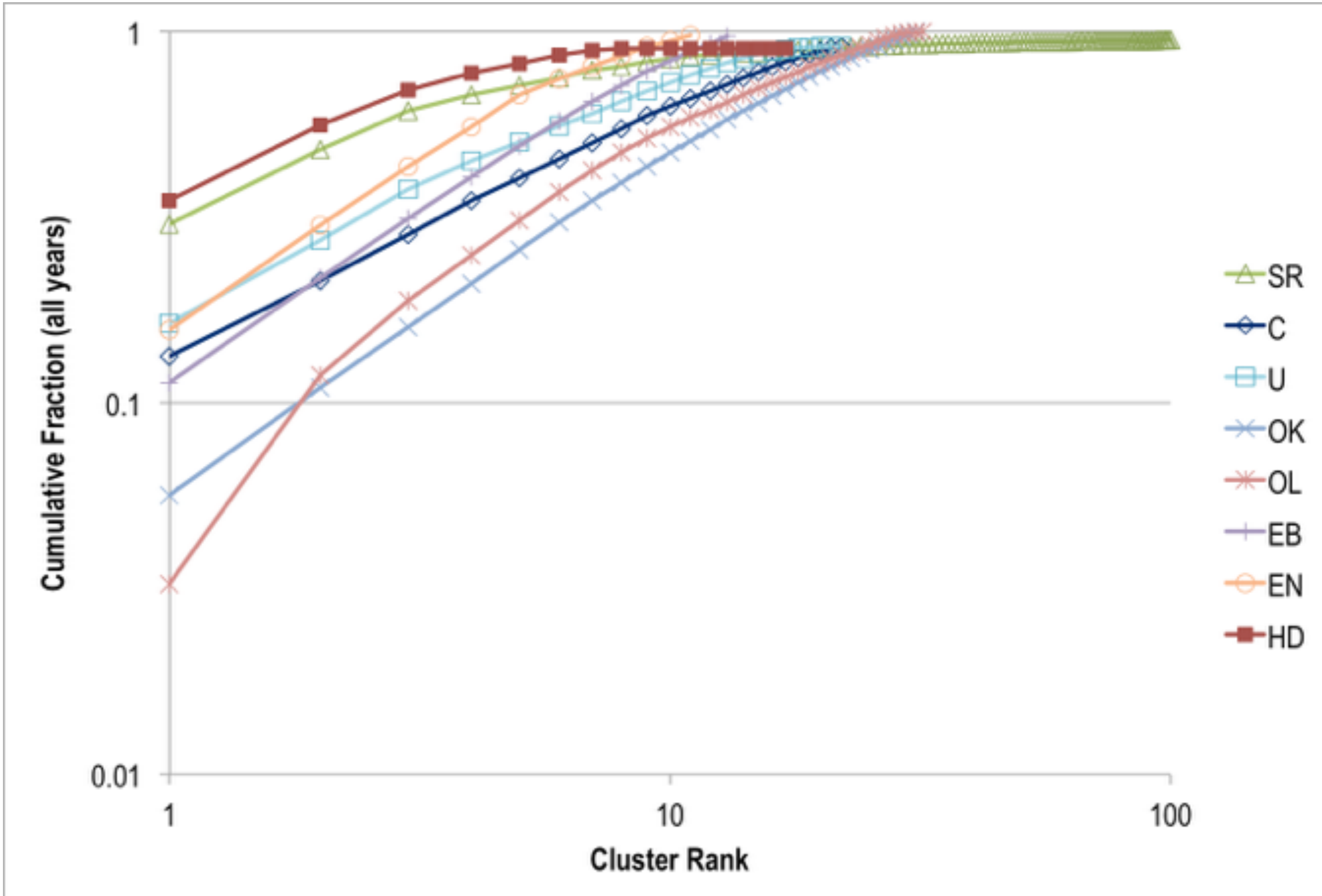
Realized Solutions

DATA MODEL

CLUSTERING ALGORITHM

	Direct Citation	Bibliographic Coupling	Hybrid (bc & terms/NLP)	Semantic matrix	Projection onto Global Direct Citation Map
Infomap	u	--	--	--	--
SLMA	c	--	--	--	sr
Memetic	hd	--	--	--	--
Louvain	--	eb	eh	ol	--
K-means	--	--	--	ok	--

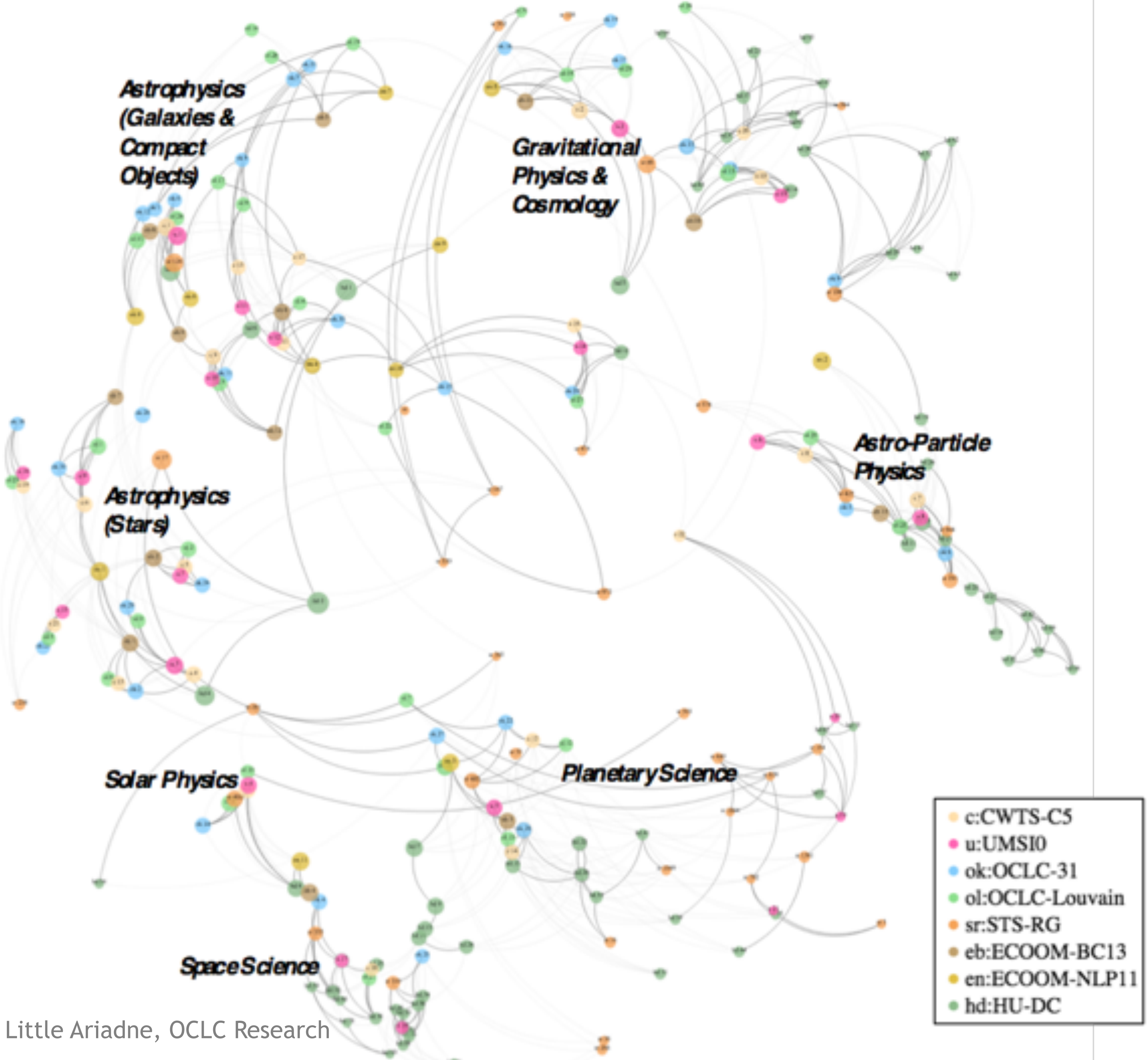
sr: Kevin Boyack (SciTech Strategies)
c: Nees van Eck, Ludo Waltman (CWTS Leiden)
eb, en: Wolfgang Glänzel & Bart Thijs (ECOOM)
hd: Jochen Gläser (TU Berlin), Frank Havemann & Michael Heinz (HU Berlin)
ol, ok: Rob Koopman & Shenghui Wang (OCLC Research)
u: Theresa Velden, Shiyang Yang, Carl Lagoze (UMSI)



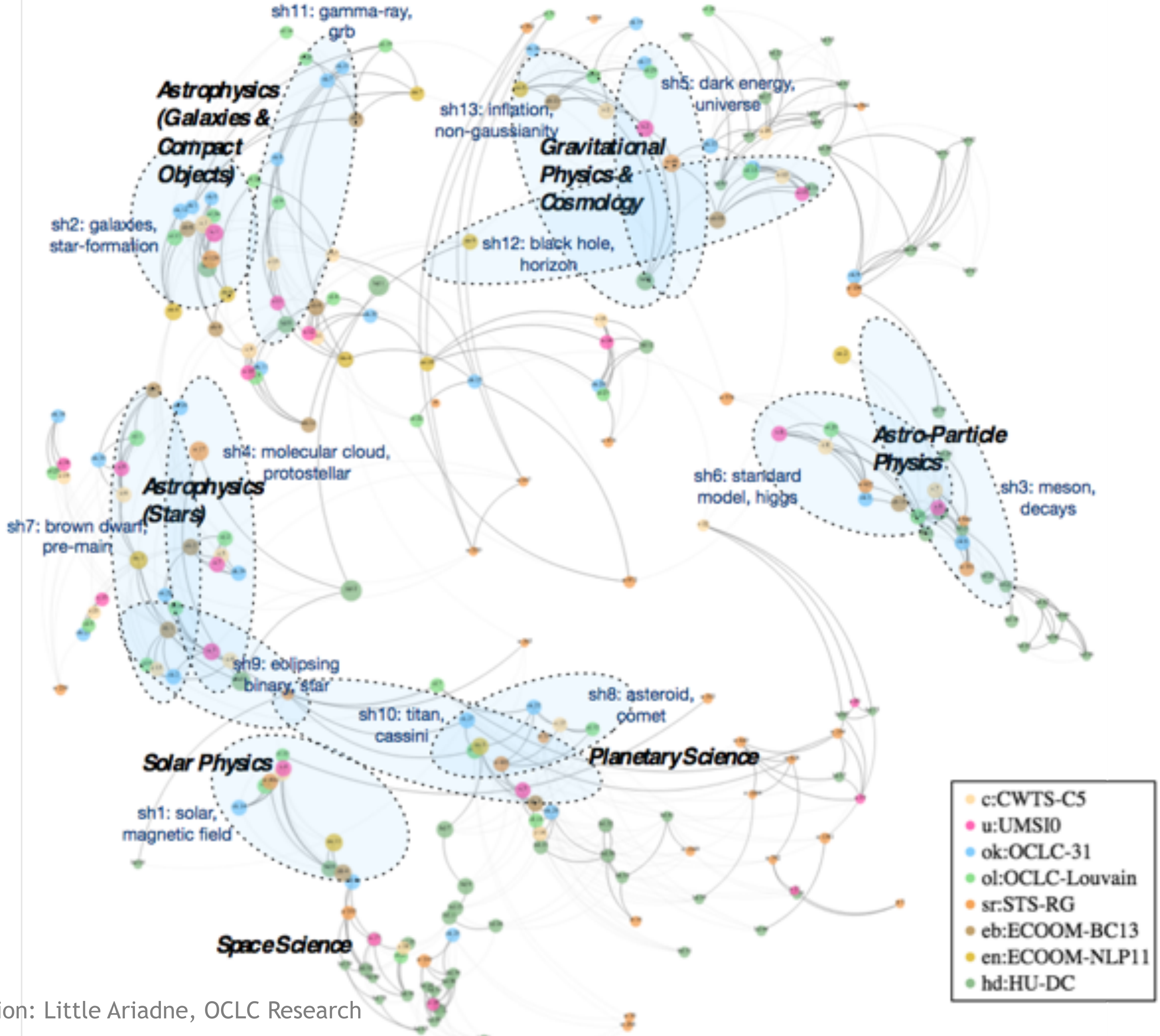
1
0.1
0.01
Cumulative Fraction (all years)
1
10
100
Cluster Rank
SR
C
U
OK
OL
EB
EN
HD

Labeling approaches

- Cluster-level labels (Word & Thesaurus based)
  - Mutual Information based score (Boyack, special issue)
  - Labels: thesaurus terms (Unified Astronomy Thesaurus (UAT, http://astrothesaurus.org) or words extracted from titles and abstracts
- Assigning clusters to domains (Journal Signature)
  - Journals in a cluster ranked by a score that combines popularity and idiosyncrasy (Velden et al, special issue)
  - Reveals sub-disciplinary groupings
  - Labels for groupings created using subject knowledge

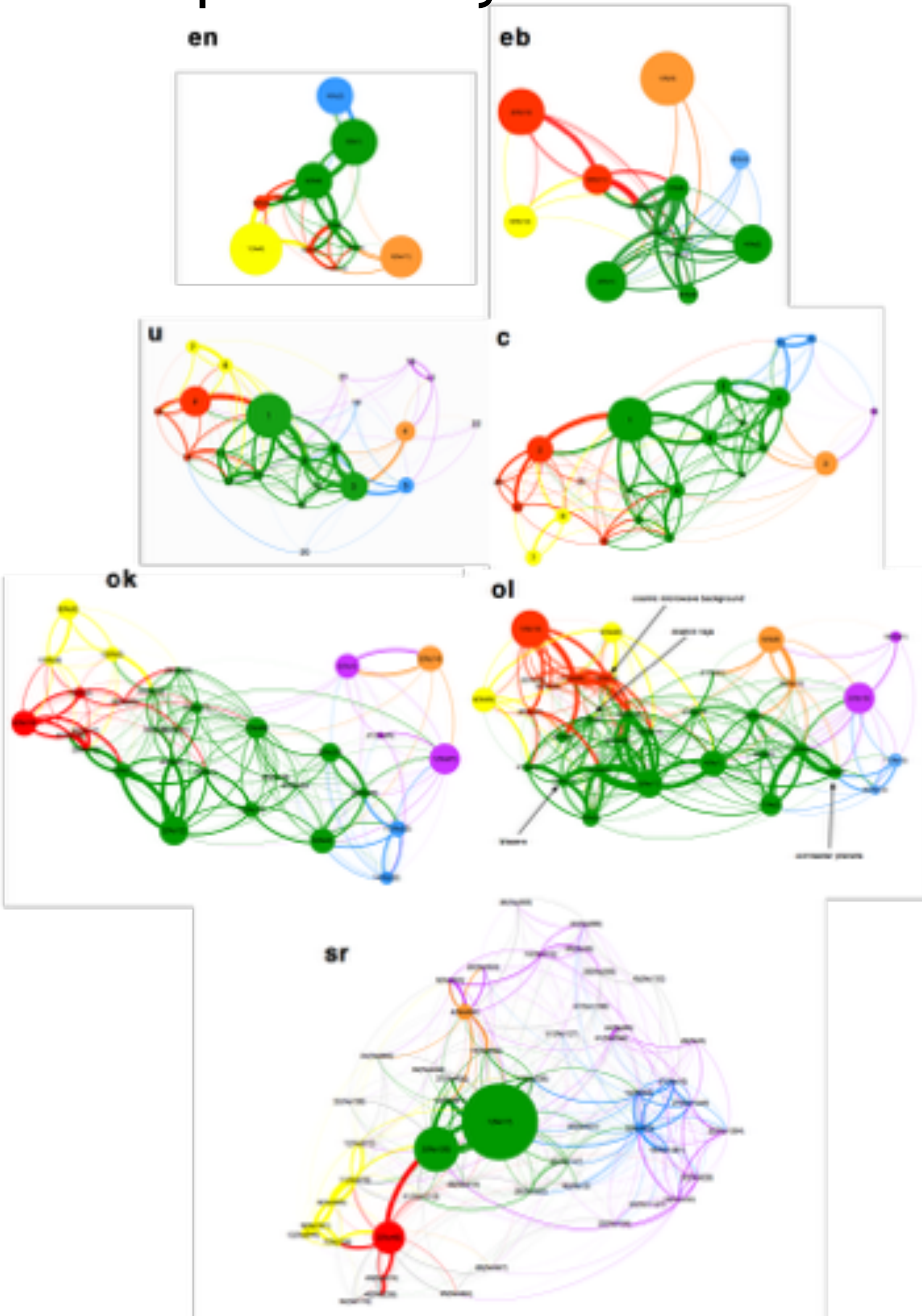


Visualization: Little Ariadne, OCLC Research

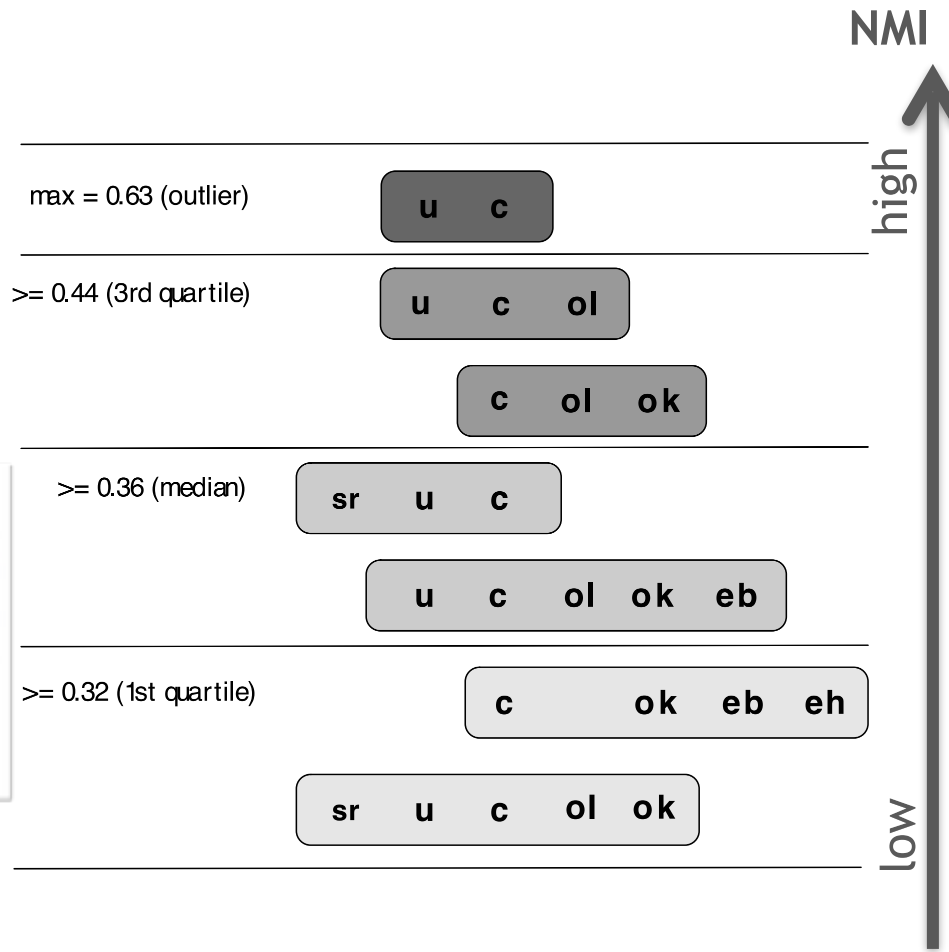


Visualization: Little Ariadne, OCLC Research

Topic Affinity Networks



Grouping of solutions based on Similarity Metric (NMI)

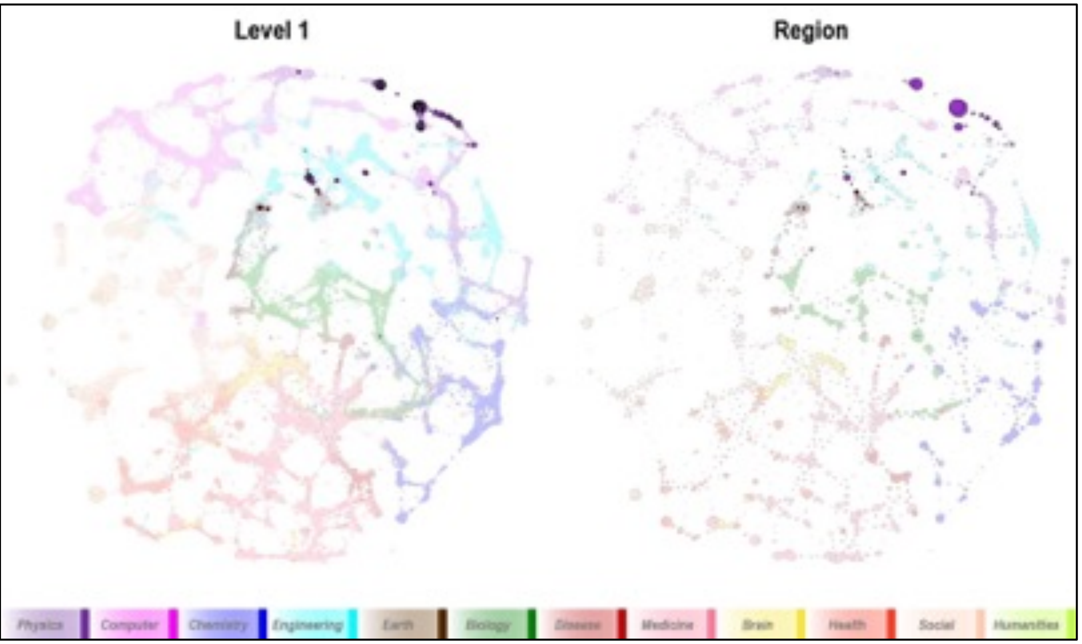


Specific Pairwise Comparisons *Dimensions*

1. Internal versus external perspective
2. Semantic versus citation based
3. Local versus global clustering

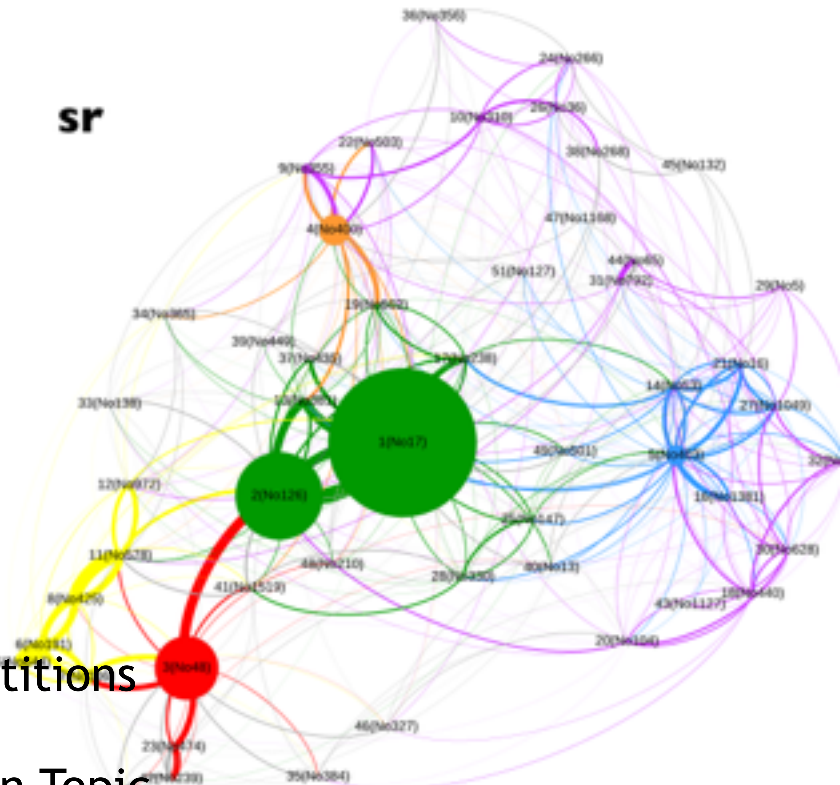
Construction of ‘external’ perspective:
Projection of AstroData onto Global Science Map

Global Map (Scopus 1996-2012)



Level 1: ~100,000 partitions   Regions: 1,649 partitions

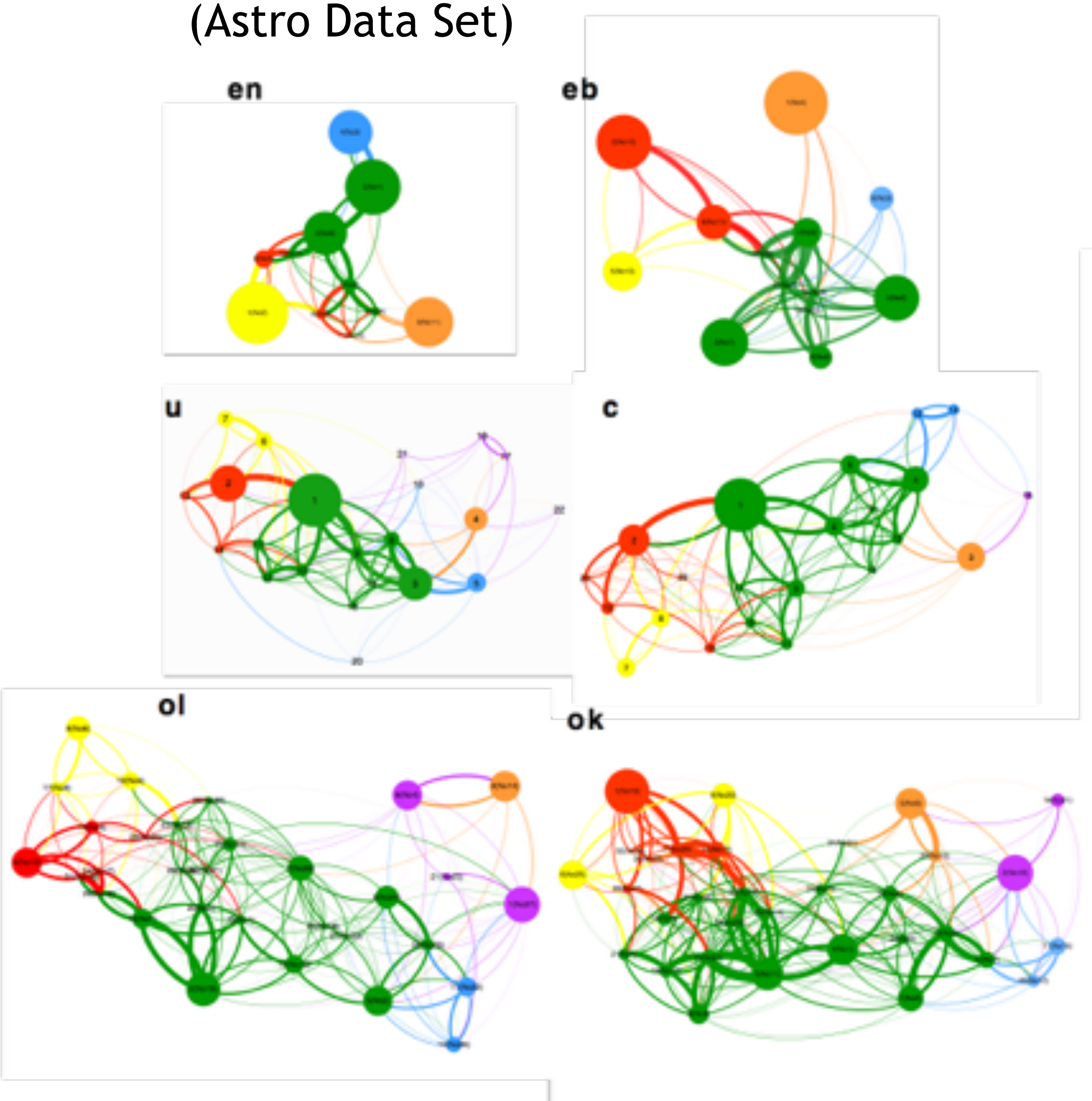
sr (Astro Data Set, WoS 2003-2010)

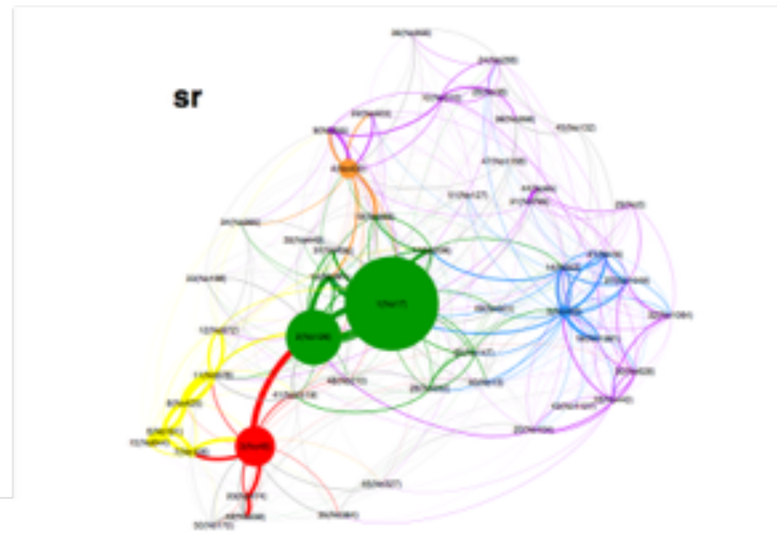


[Boyack, Investigating the Effect of Global Data on Topic Detection (forthcoming) *Scientometrics* Special Issue]

Internal perspective versus External perspective

(Astro Data Set) (Global Science Map)

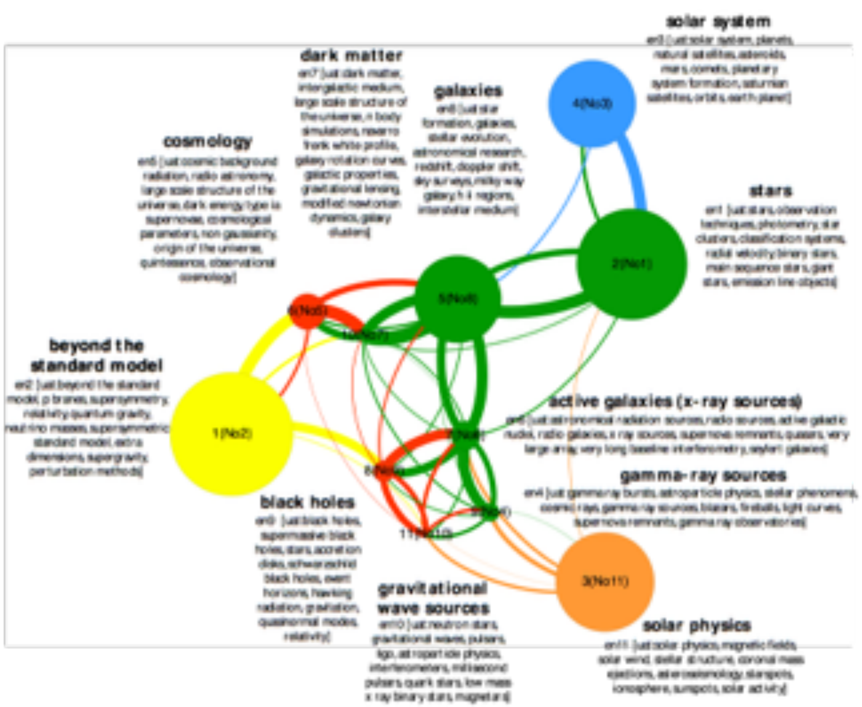


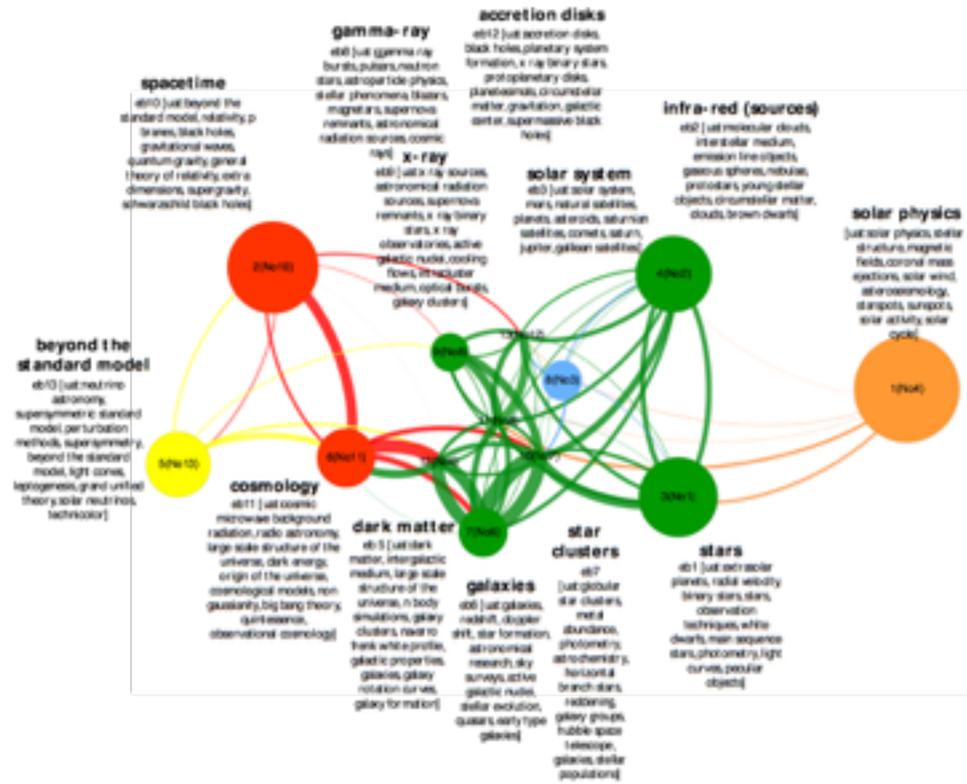


Observations

1. sr lacks resolution in center (regions!)
2. sr provides high-resolution at periphery
3. Some topics in internal solutions artifacts?

Hybrid (bibliographic coupling + NLP) versus bibliographic coupling

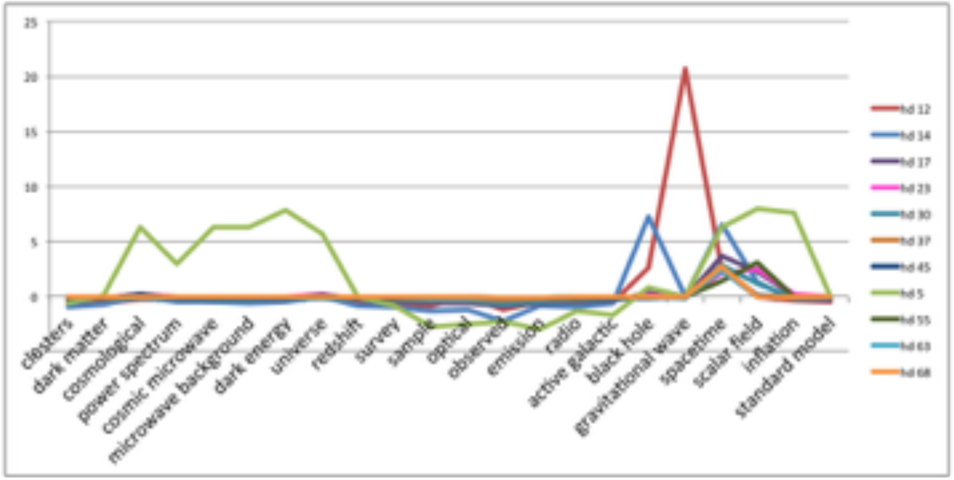


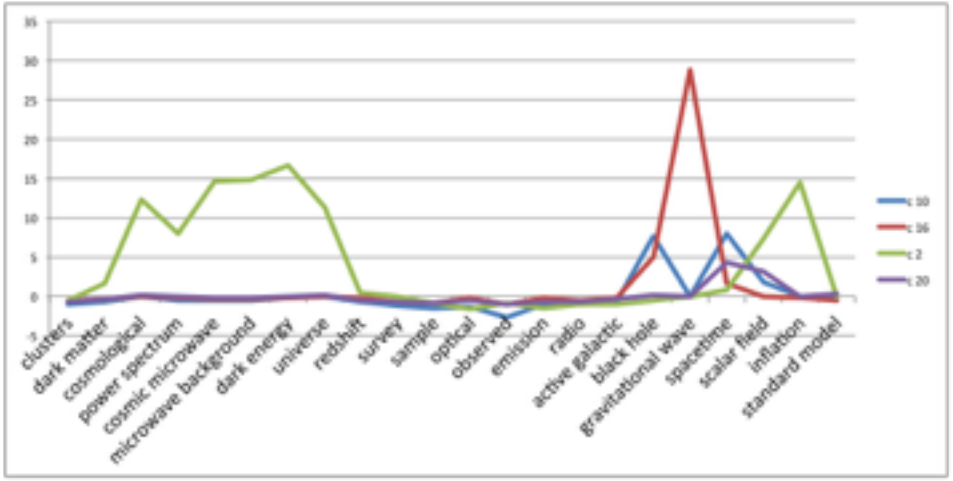


<u>Observation</u>:
Semantic similarities lead to different aggregations of documents and lead to distinctively different topic sizes and changes in topology of affinity network (e.g.´extrasolar planets‘, ´plasma’)

Local clustering **versus** Global cluster
(hd, memetic) (c, SLMA)

Topics identified in Gravitation & Cosmology (Lexical Fingerprint Analysis)

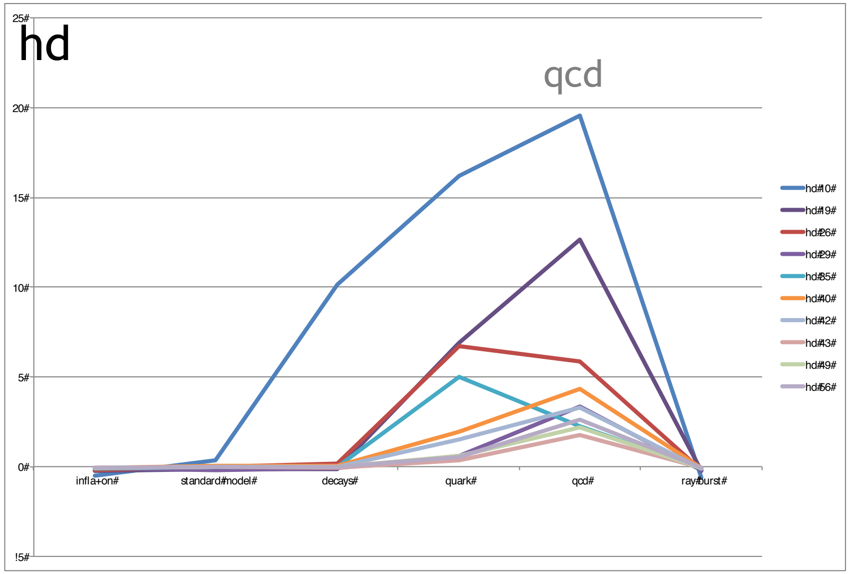


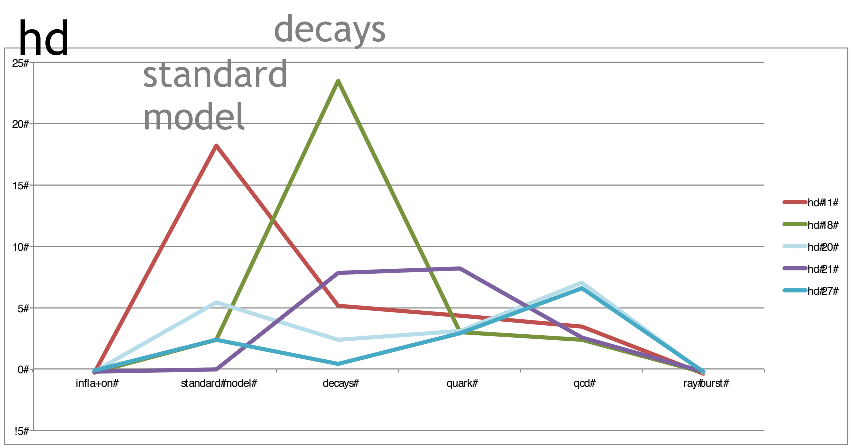


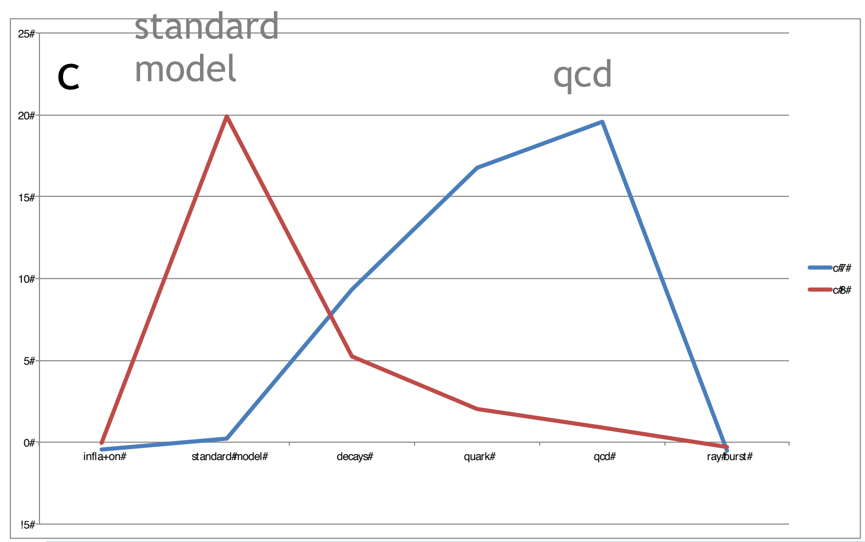
<u>Observation</u>:
- Local clustering delivers topics very similar to the global clustering approach
- Additional topics are close and smaller variants of detected topics

Local clustering **versus** Global clustering
(hd, memetic) (c, SLMA)

Topics identified in Astroparticle Physics (Lexical Fingerprint Analysis)







<u>Observation</u>:

- Local clustering delivers the two topics the global clustering approach identifies plus 'scatter'
- Additional **new topic** in the middle

Conclusions

- Significant differences depending on approach
- Differences can be tentatively explained by features of data model and clustering algorithm
- Open challenges:
  - Identifying best approach for a given purpose
  - Validate a topic extraction solution in context of purpose