Same Data, Different Results-- On a Comparative Topic Extraction Exercise

Theresa Velden
University of Michigan School of Information (UMSI)

SIGMET Workshop at ASIST 2015
November 7, 2015

In collaboration with:

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)

The performative nature of topic extraction

- To what extent do topic extraction approaches capture the 'ground truth' of thematic structure in a field or how does the choice of approach shape the results and introduce artifactual features?
- In Scientometrics topic extraction approaches are rarely directly compared on same data set; lack of understanding of nature & origin, and implications of differences

Background

- Evolved from annual meetings of advisory project funded by German Ministry for Education and Research on ***'Measuring Diversity in Science'*** (Jochen Gläser, Frank Havemann & Michael Heinz)
- To measure epistemic diversity of a field, the field needs to be delineated and topics identified
  - Even slight changes in topic structure influence measure
- Compare solutions derived from same data set ('Astro Data')
- Series of workshops (Berlin 9/2014, Amsterdam 4/2015, Berlin 8/2015)
- Special session at ISSI 2015, July in Istanbul

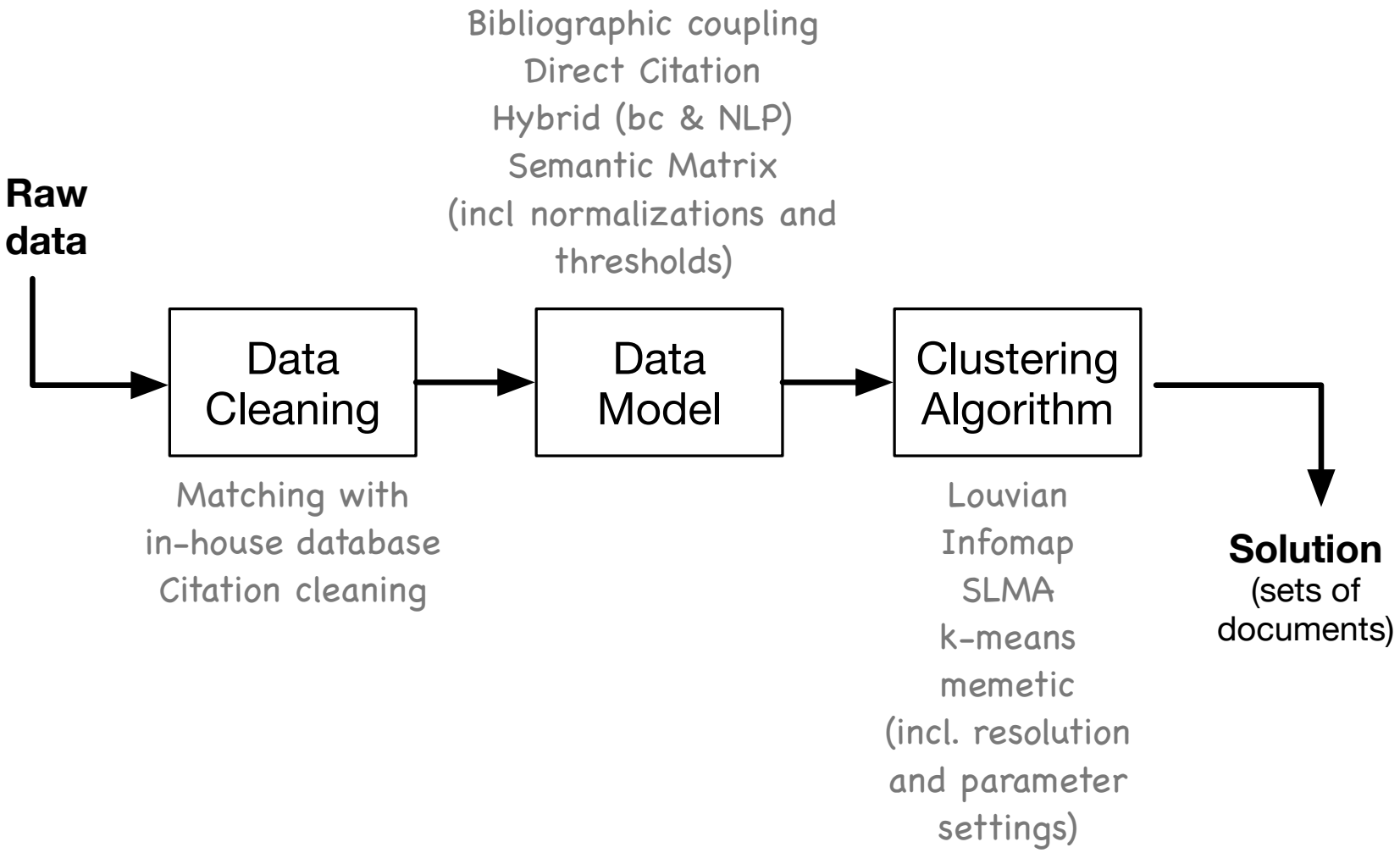
Premises & Objective

- More than one valid thematic structure can be constructed depending on the perspective applied to the knowledge.
- Topical structures are reconstructed for specific purposes, so if at all, there might be a best method for a given purpose.
- Instead of finding the one best solution, we aim at uncovering how results differ and how those differences relate to approaches

Data Set

- Source: Web of Science (Thomson Reuters)
- 8 years: 2003 -2010
- 59 astrophysics and astronomy journals
- **111,161** articles, letters & proceedings papers

Topic Extraction Workflow



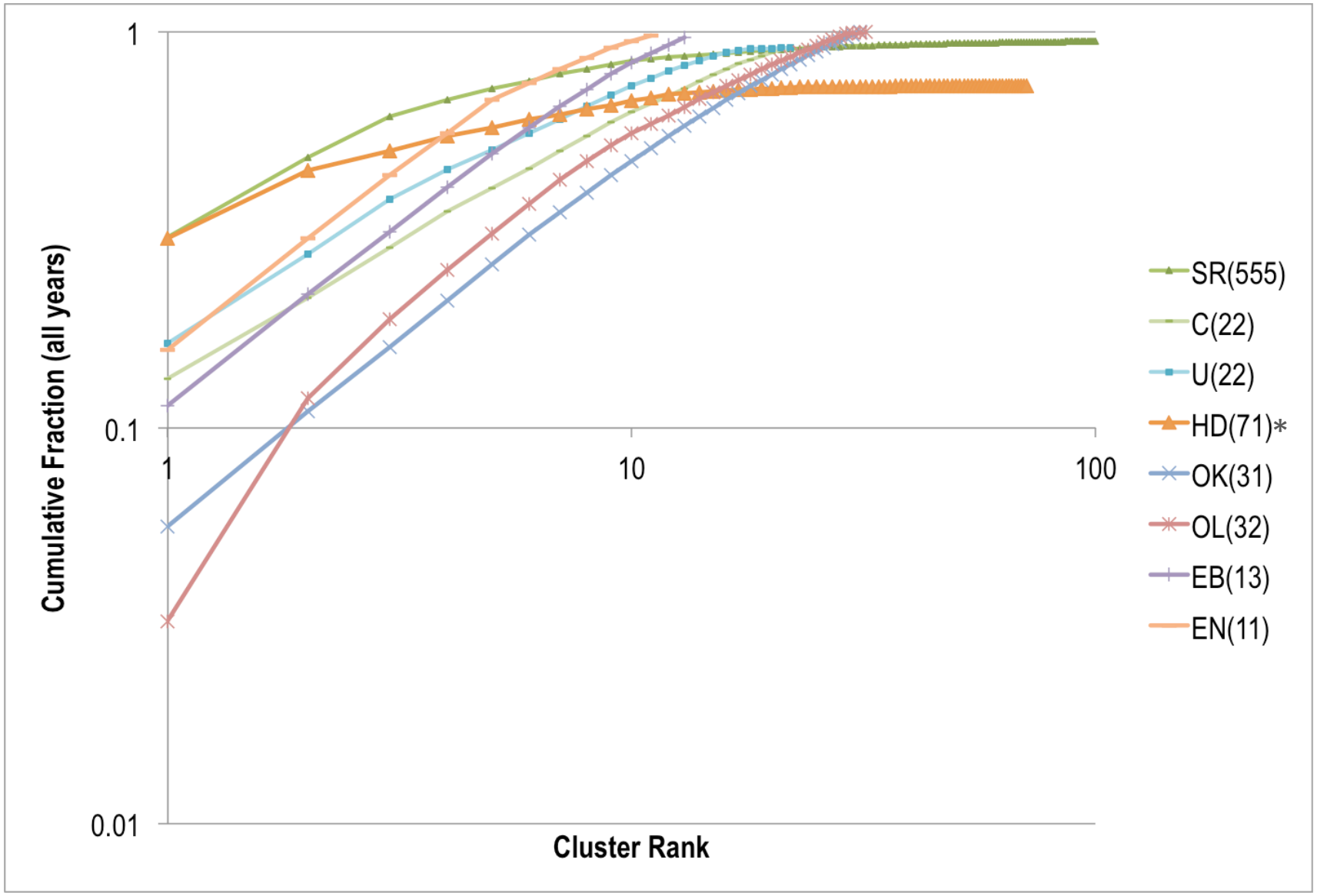
Overview Approaches

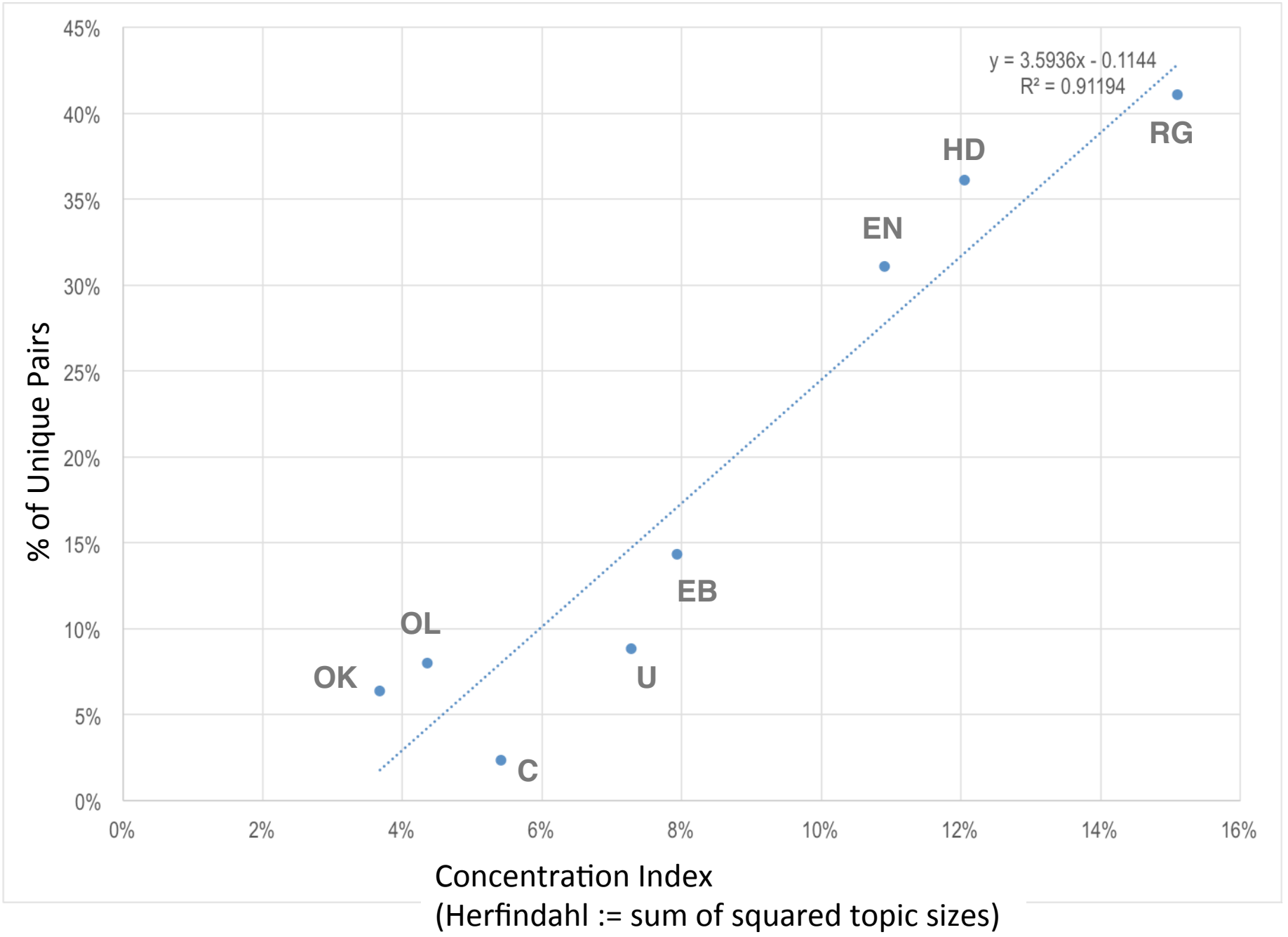
	Direct Citation	Bibliogr. Coupling	Hybrid (bc & terms/ NLP)	Semantic matrix	Projection onto Global Direct Citation Map
Infomap	UMSI	--	--	--	--
SLMA	CWTS	--	--	--	STS
Memetic	HU	--	--	--	--
Louvian	--	ECOOM	ECOOM	OCLC	--
K-means	--	--	--	OCLC	--

HU: Humboldt University; **CWTS**: Centre for Science and Technology Studies, Leiden**; ECOOM**: Expertisecentrum Onderzoek en Ontwikkelingsmonitoring; **UMSI**: University of Michigan School of Information**, OCLC**: Online Computer Library Center, Inc.; **STS**: SciTech Strategies

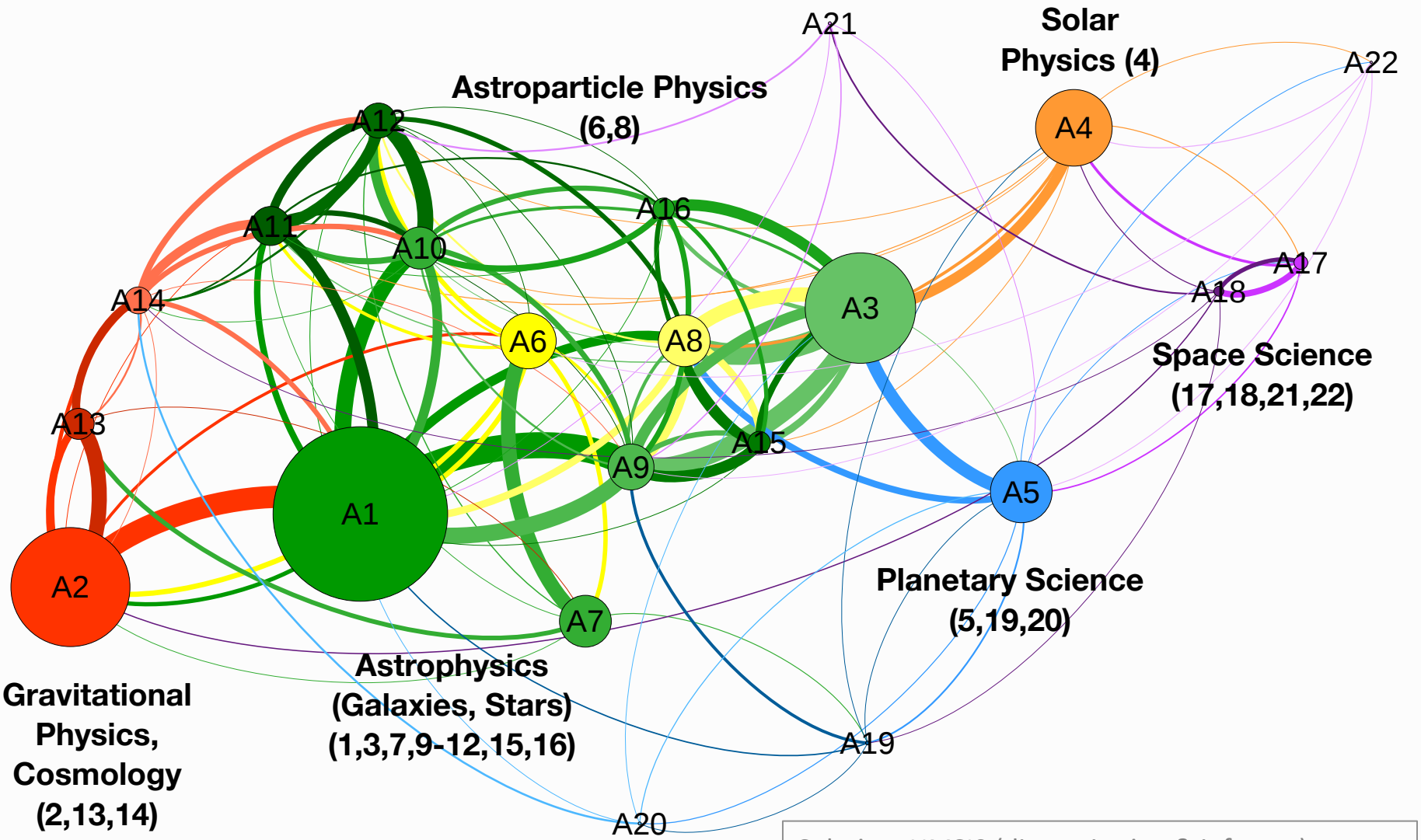
Results:

Cluster (topic) size distributions



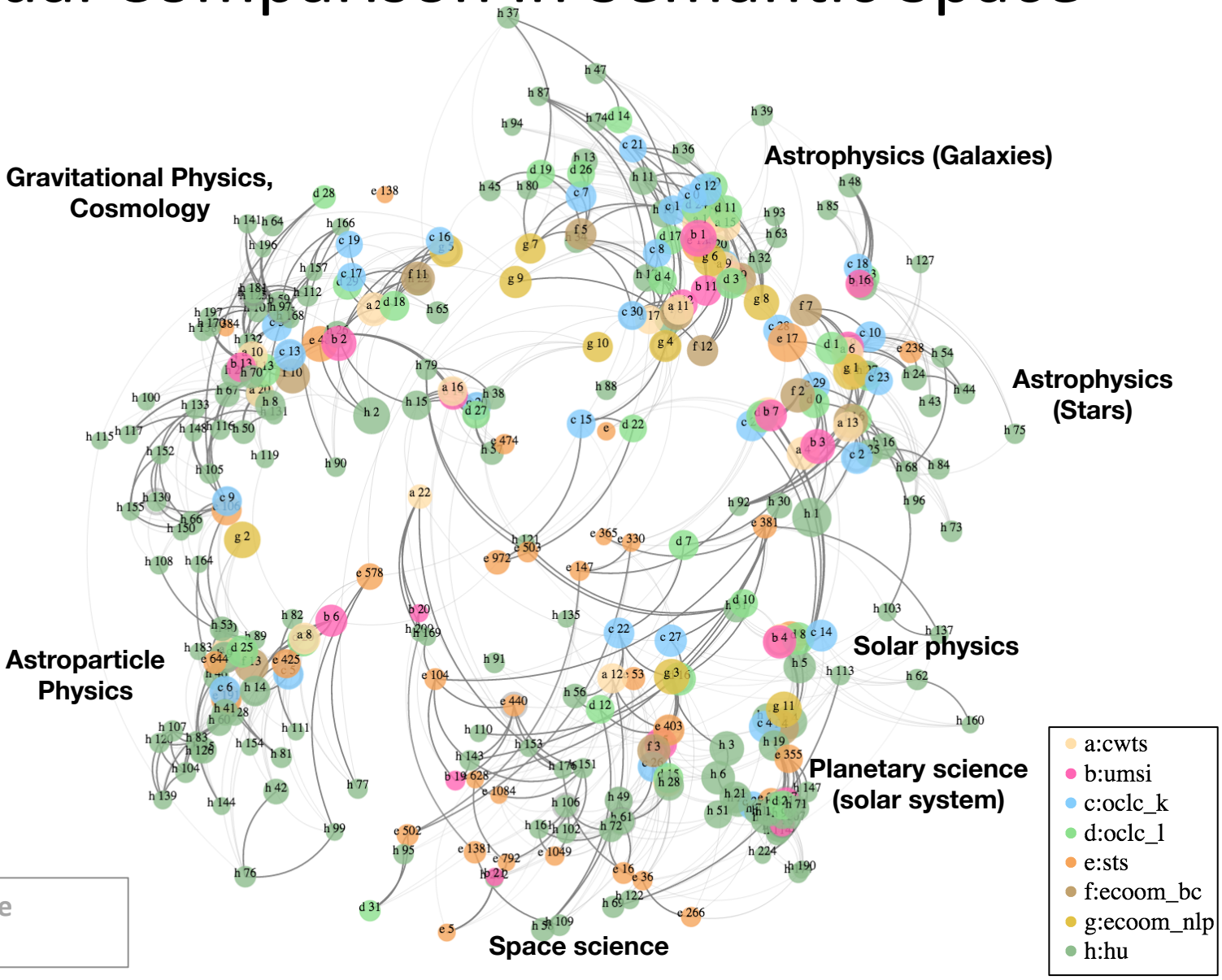

y = 3.5936x - 0.1144
R² = 0.91194
RG
HD
EN
EB
OL
OK
U
C
% of Unique Pairs
45%
40%
35%
30%
25%
20%
15%
10%
5%
0%
0%
2%
4%
6%
8%
10%
12%
14%
16%
Concentration Index
(Herfindahl := sum of squared topic sizes)

Topic Affinity Map



Solution: UMSI0 (direct citation & infomap)
Network vis: gephi, Force Atlas 2 Layout algorithm
Labeling: based on journal signature

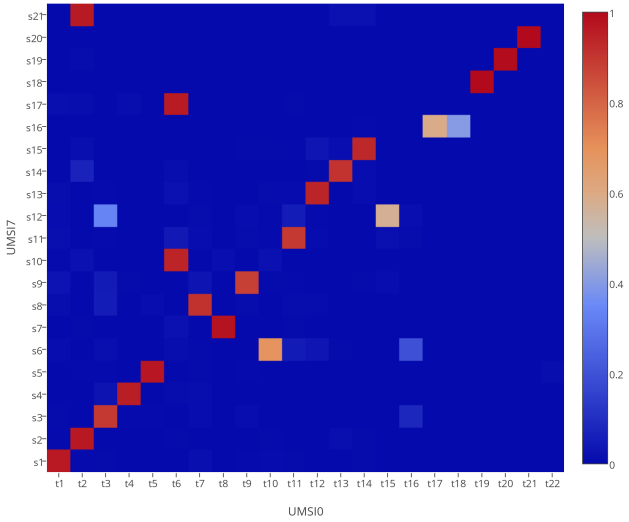
Visual Comparison in Semantic Space



Visualization: Little Ariadne (OCLC)

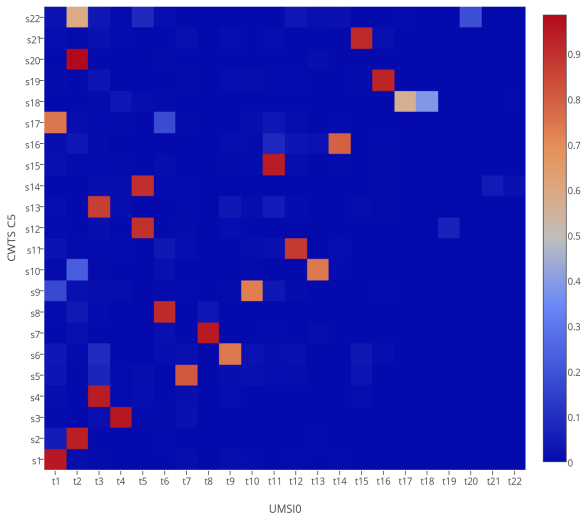
Sources of variability

UMSI7 vs UMSI0



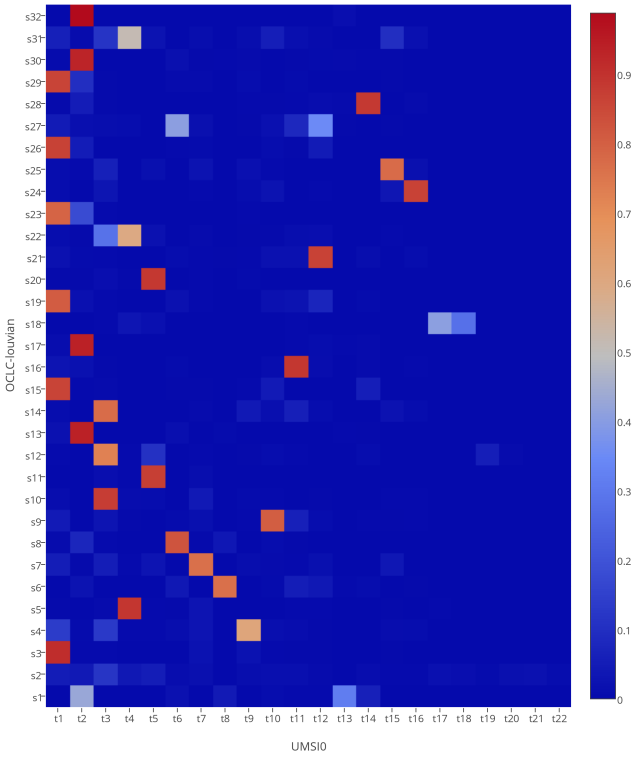
Same model & same algorithm (stochastic variation)

CWTS-C5 vs UMSI0



Same model & different algorithm

OCLC-louvian vs UMSI0



Different model & different algorithm

Overlap Between Clusters: Comparison with UMSI0 Cluster Solution (22 clusters)

Comparison: Set based metrics

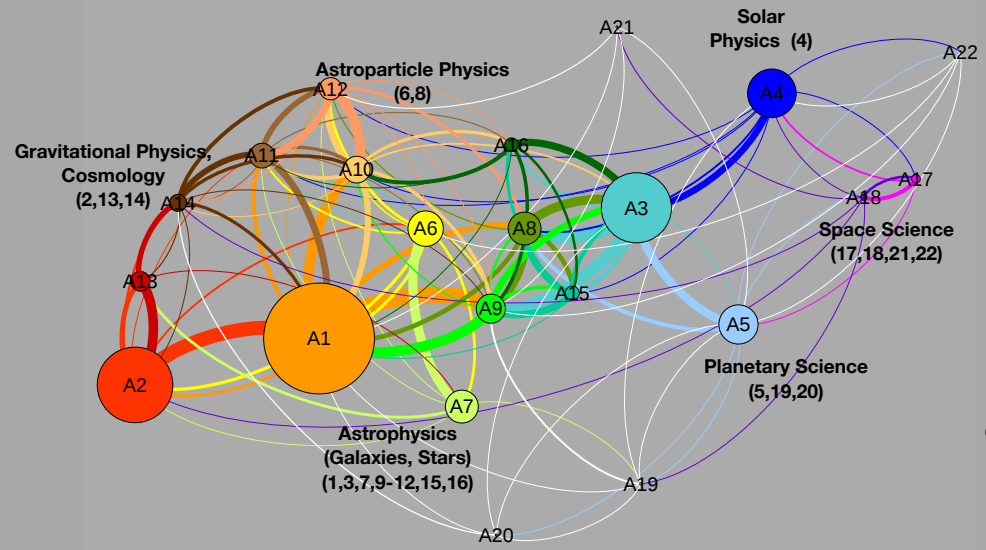
Normalised Mutual Information

	sr	c	u	ok	ol	eh	eb
sr	1.000	0.359	0.372	0.329	0.333	*0.243*	*0.306*
c	0.359	1.000	0.633	0.464	0.516	0.316	0.380
u	0.372	**0.633**	1.000	0.424	0.471	*0.295*	0.356
ok	0.329	0.464	0.424	1.000	**0.515**	0.334	0.363
ol	0.333	**0.516**	0.471	0.515	1.000	0.307	0.362
eh	0.243	0.316	0.295	0.334	0.307	1.000	0.330
eb	0.306	0.380	0.356	0.363	0.362	0.330	1.000

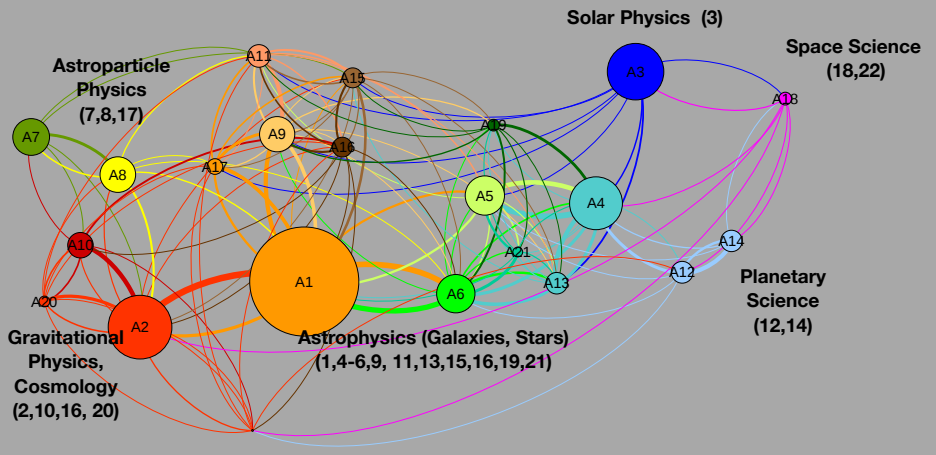
Overlap Index

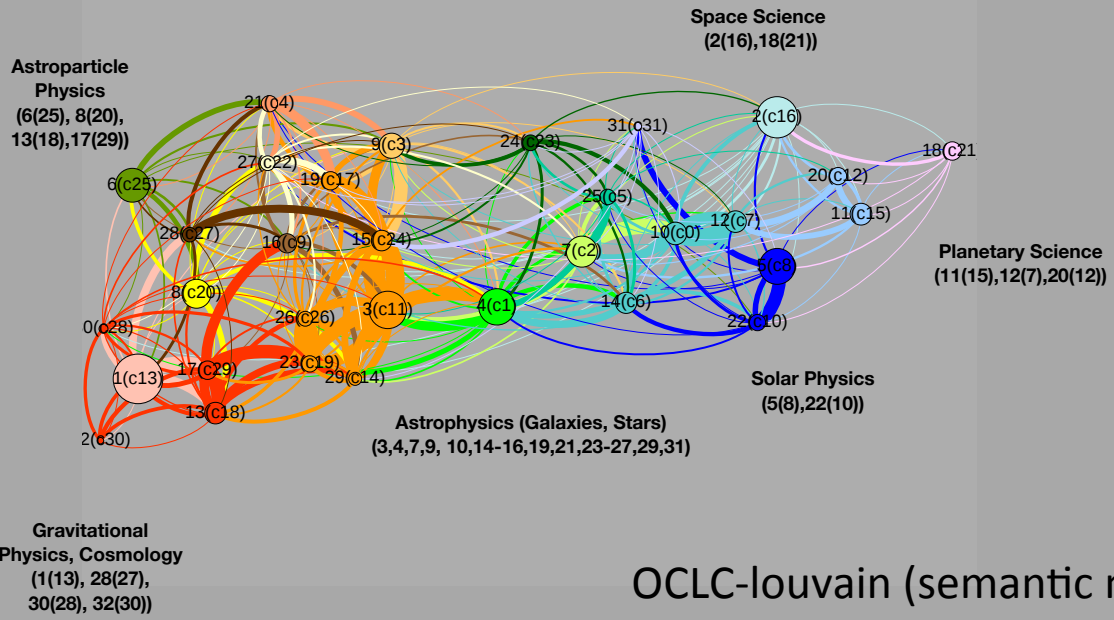
	sr	c	u	ok	ol	eh	eb
sr	1.000	0.698	0.706	0.686	0.692	0.591	0.662
c	0.698	1.000	**0.835**	0.622	**0.708**	*0.546*	0.593
u	0.706	0.835	1.000	0.645	**0.725**	*0.526*	0.574
ok	0.686	0.622	0.645	1.000	0.619	0.609	0.576
ol	0.692	0.708	0.725	0.619	1.000	0.553	0.567
eh	0.591	0.546	0.526	0.609	0.553	1.000	*0.541*
eb	0.662	0.593	0.574	0.576	0.567	0.541	1.000

UMSI-0 (direct citation, Infomap)



CWTS-C5 (direct citation, SLMA)

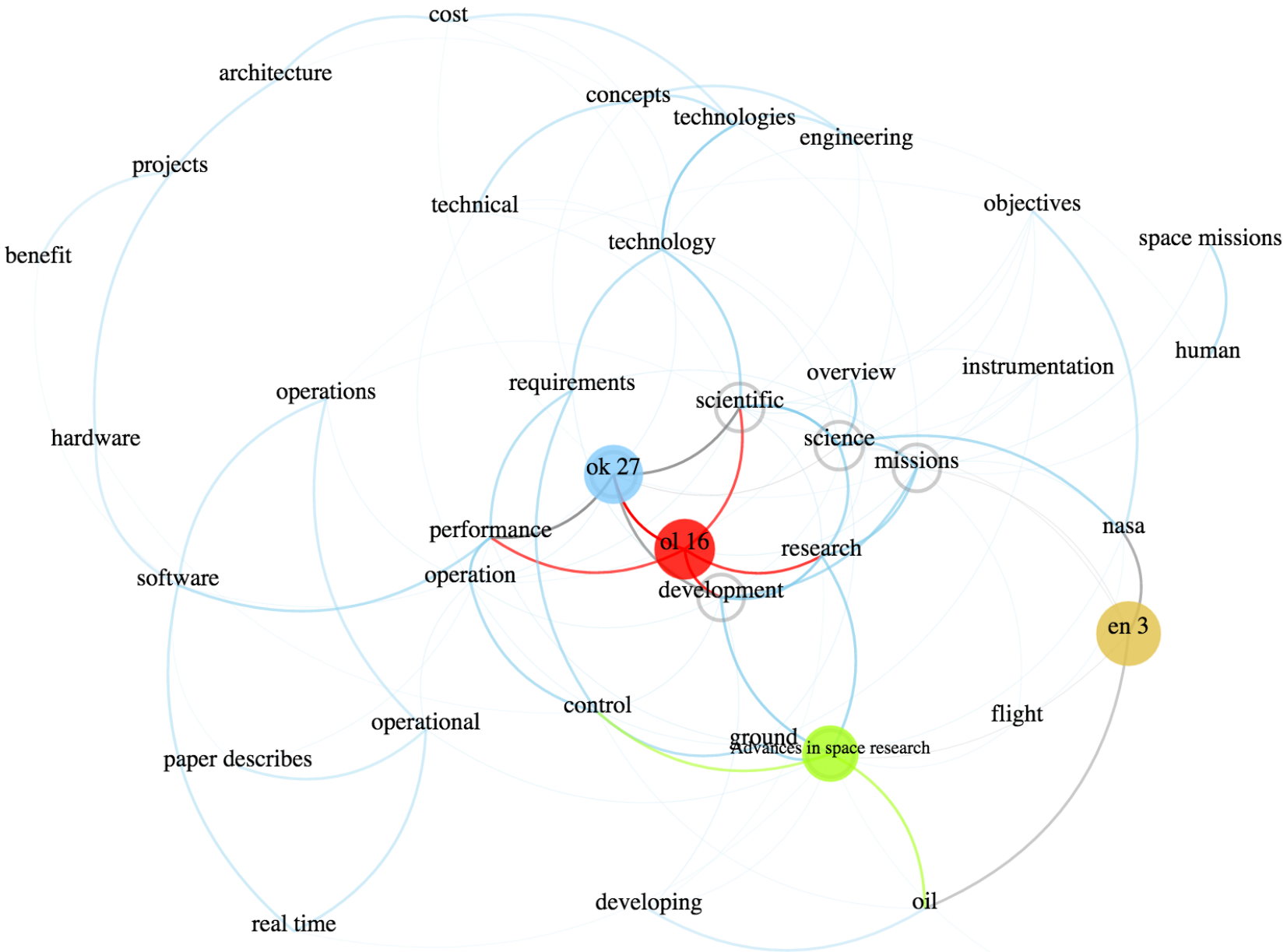




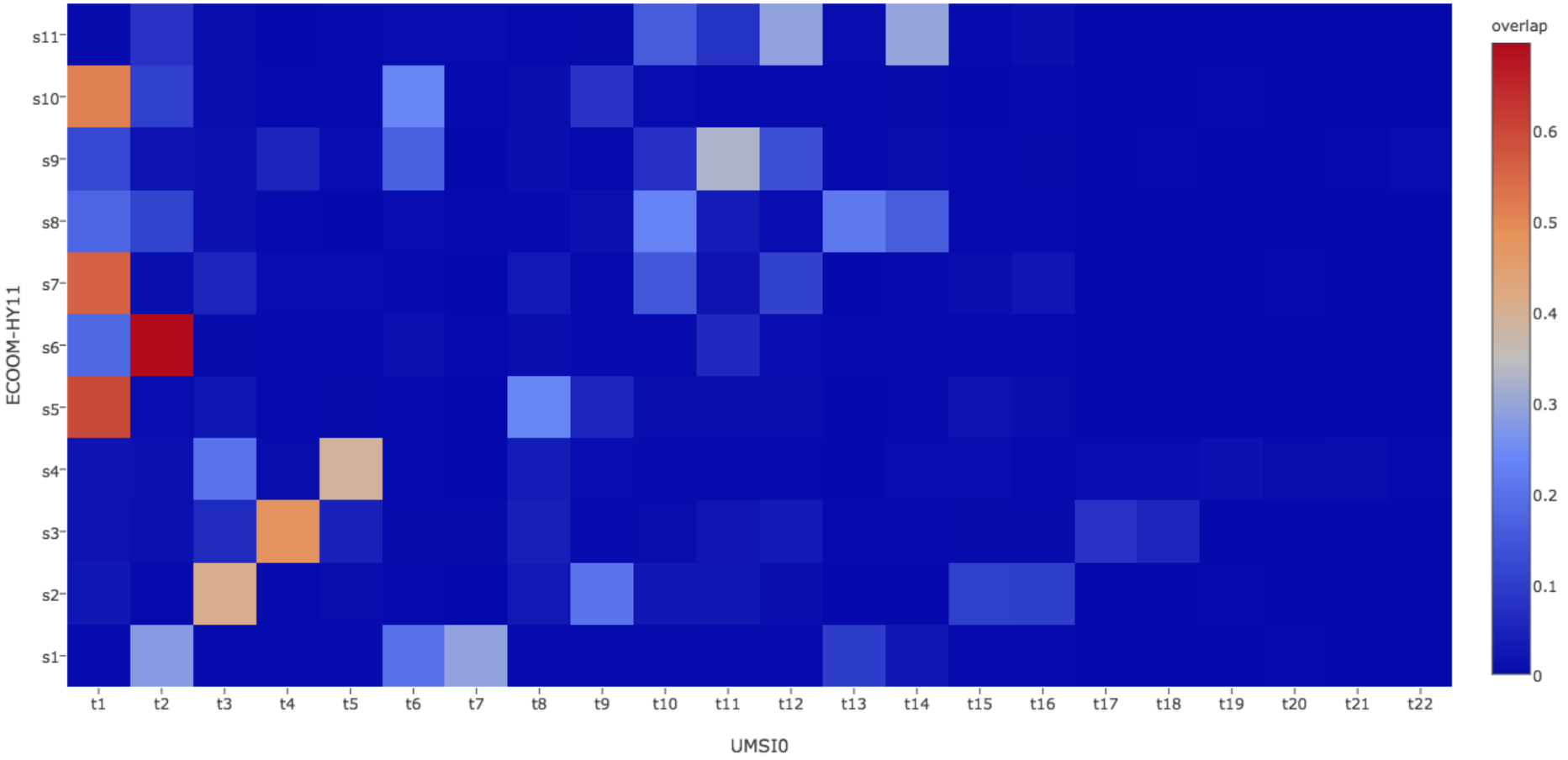
OCLC-louvain (semantic matrix, Louvain)

Outside of the giant component:

Blind spot of a citation based approach

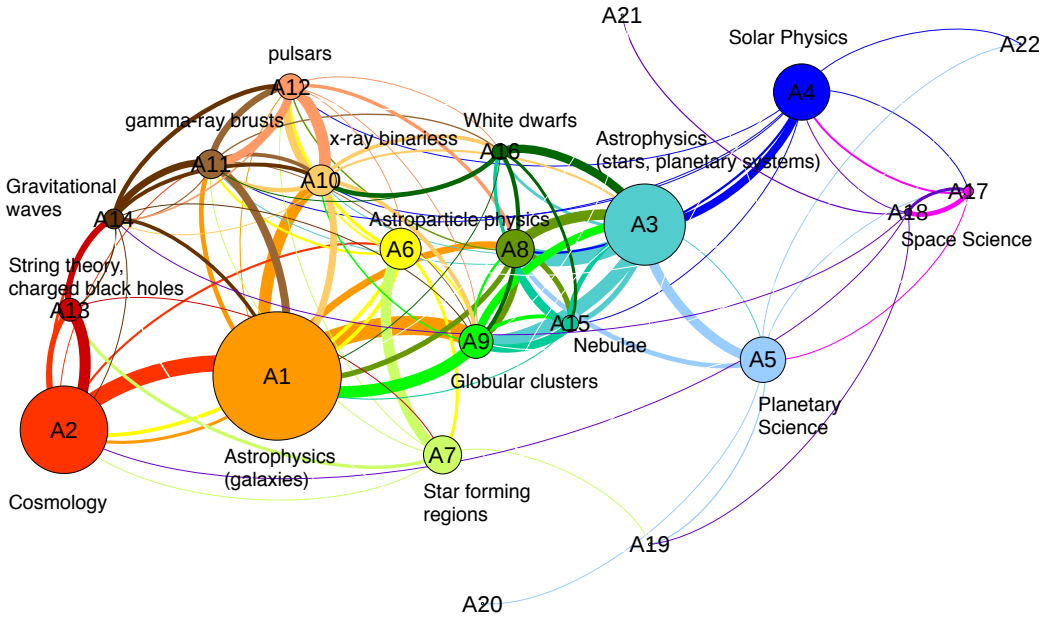


Two very different solutions

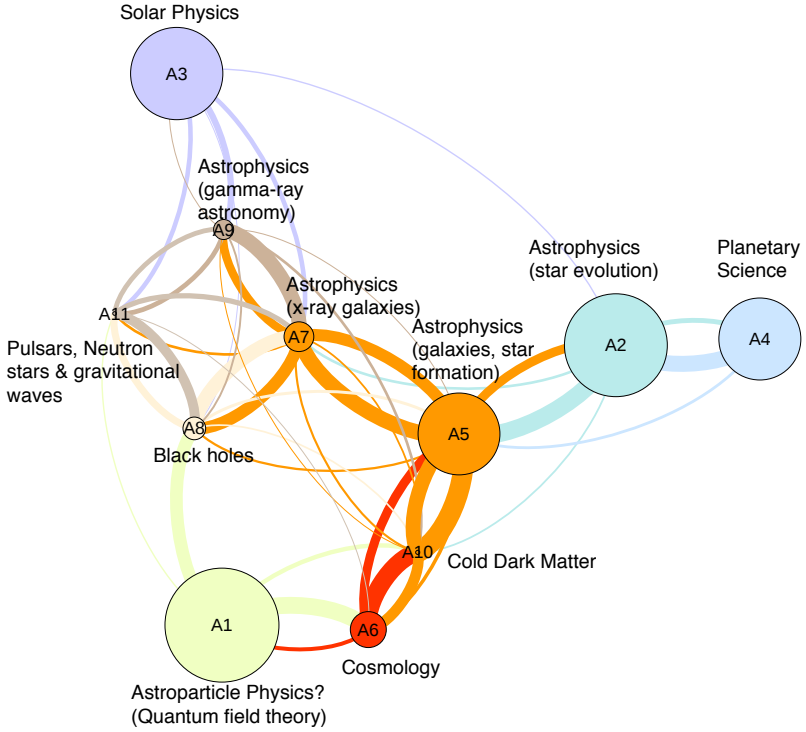


Two very different solutions

UMSI0 (Direct citation & Infomap)



ECOOM-HY (bibliographic coupling/NLP term extraction & Louvain)



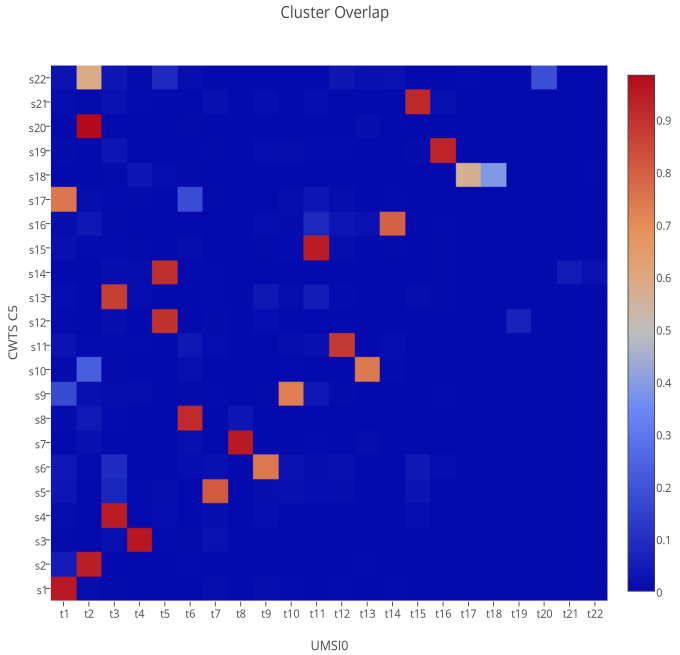
Network vis: gephi, Force Atlas 2 Layout algorithm
Labeling: based on Little Ariadne related words/subjects

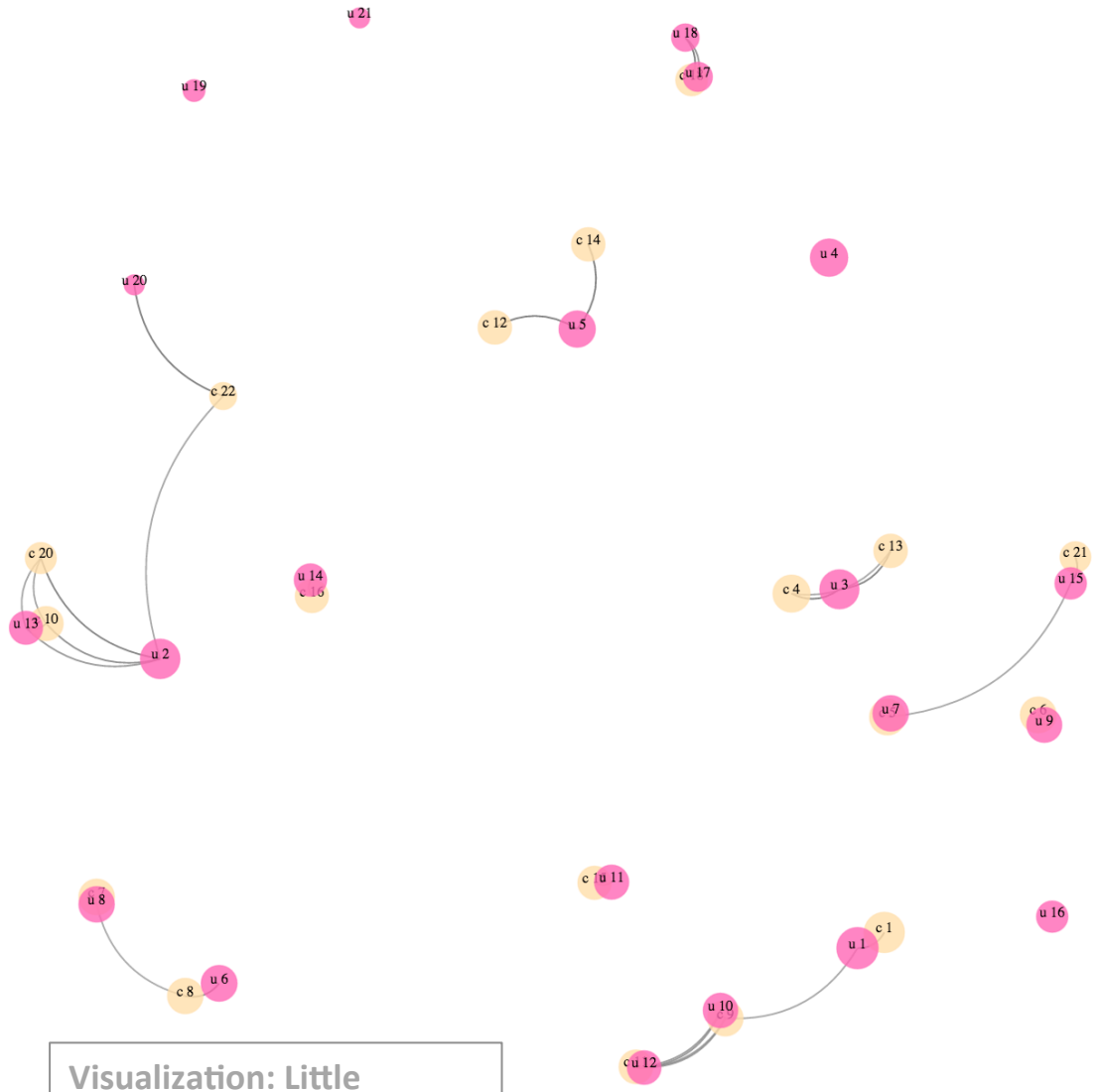
Conclusions & Outlook

- Developing methods for meaningful comparison a major challenge
  - combination of quantitative metrics & visualizations
- Variations due to coverage, modeling & clustering
- Comparative analysis ongoing:
  - Case studies (instances of agreement and divergence)
  - Blind spots (areas left out by some approaches)
  - Mapping onto Unified Astronomy Thesaurus
- Special Issue in Scientometrics in preparation
- Call to join 'Topic Extraction Challenge'

Theresa Velden, tvelden@umich.edu, University of Michigan School of Information

Comparison of UMSI0 and CWTS-C5





Visualization: Little Ariadne (OCLC)