Mapping Scientific Communities - Opening up the Black Box

Theresa Velden, PhD
University of Michigan, School of Information

Conference: "Visualisation for Science Policy"
September 18, 2015, Warsaw, Poland

Opening up the "Black Box"

Black box "broadly: anything that has mysterious or unknown internal functions or mechanisms" [Merriam Webster Dictionary]

In science and technology studies: looking 'under the hood' at the social mechanics involved in producing scientific knowledge typically neglected and de-emphasized in the official account of how scientific results are obtained.

Opening up the Black Box

Part 1: How Scientific Communities Produce Knowledge – Insights Gained From Maps of Science

Part 2: How Maps of Science Are Produced – Discussion of Challenges Encountered

Mixed Method Approach

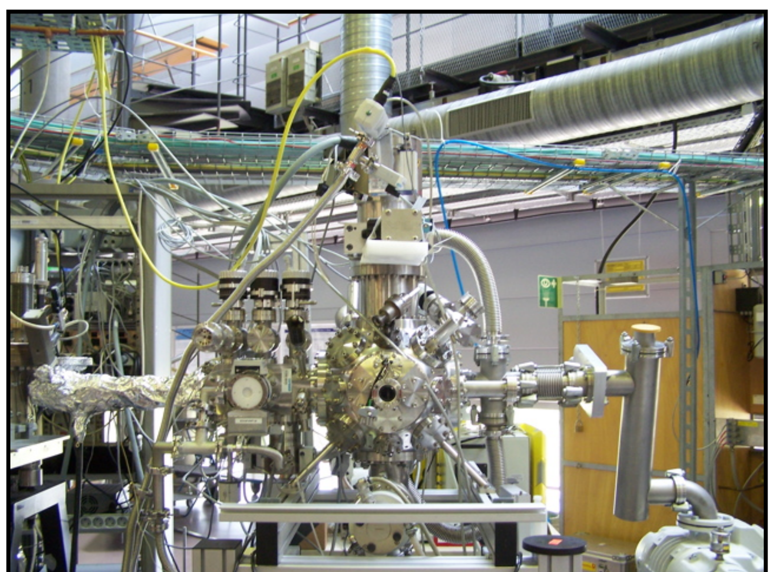
Network analysis

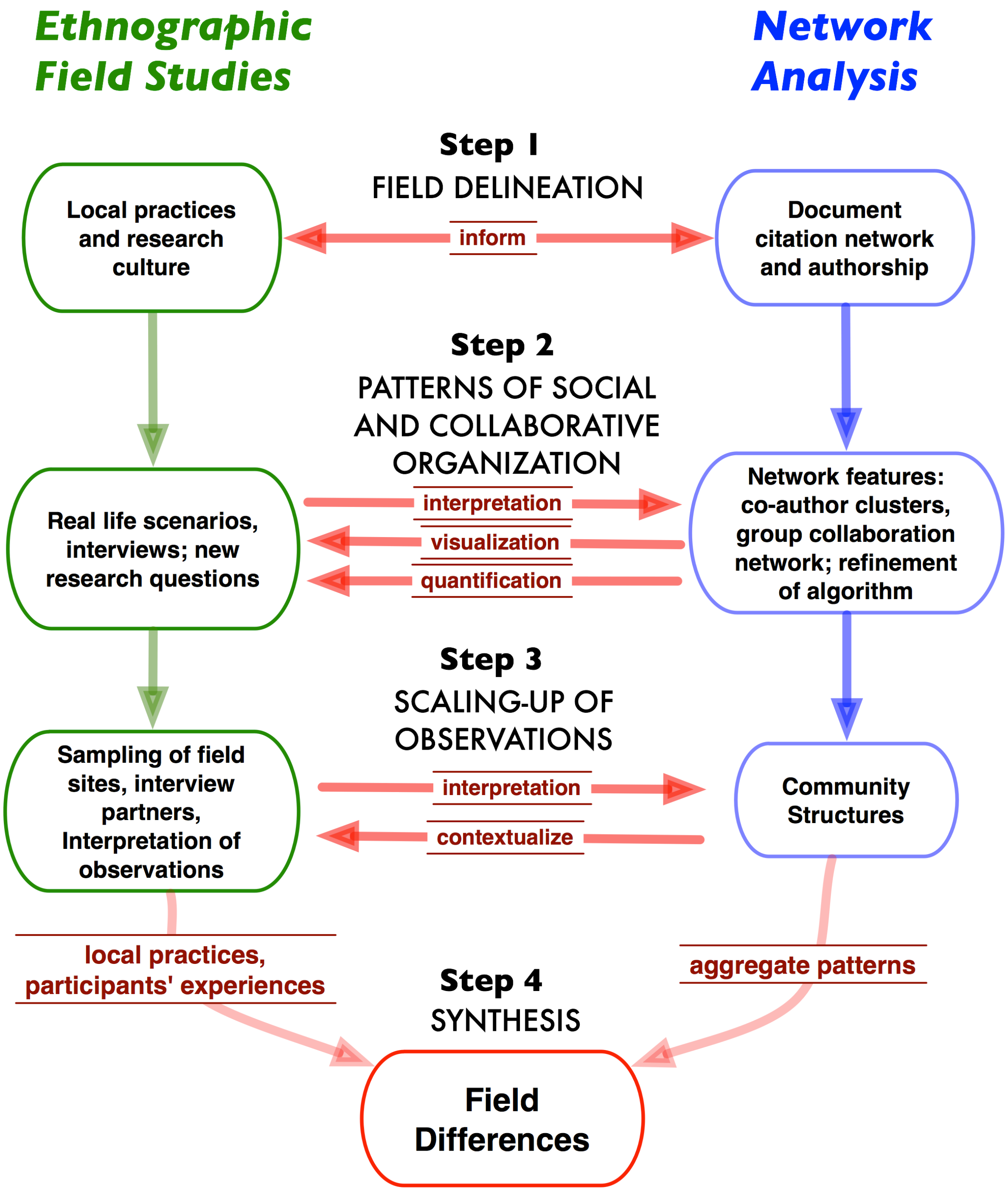
large publication networks (several 10,000 publications/authors)

&

Ethnographic field studies

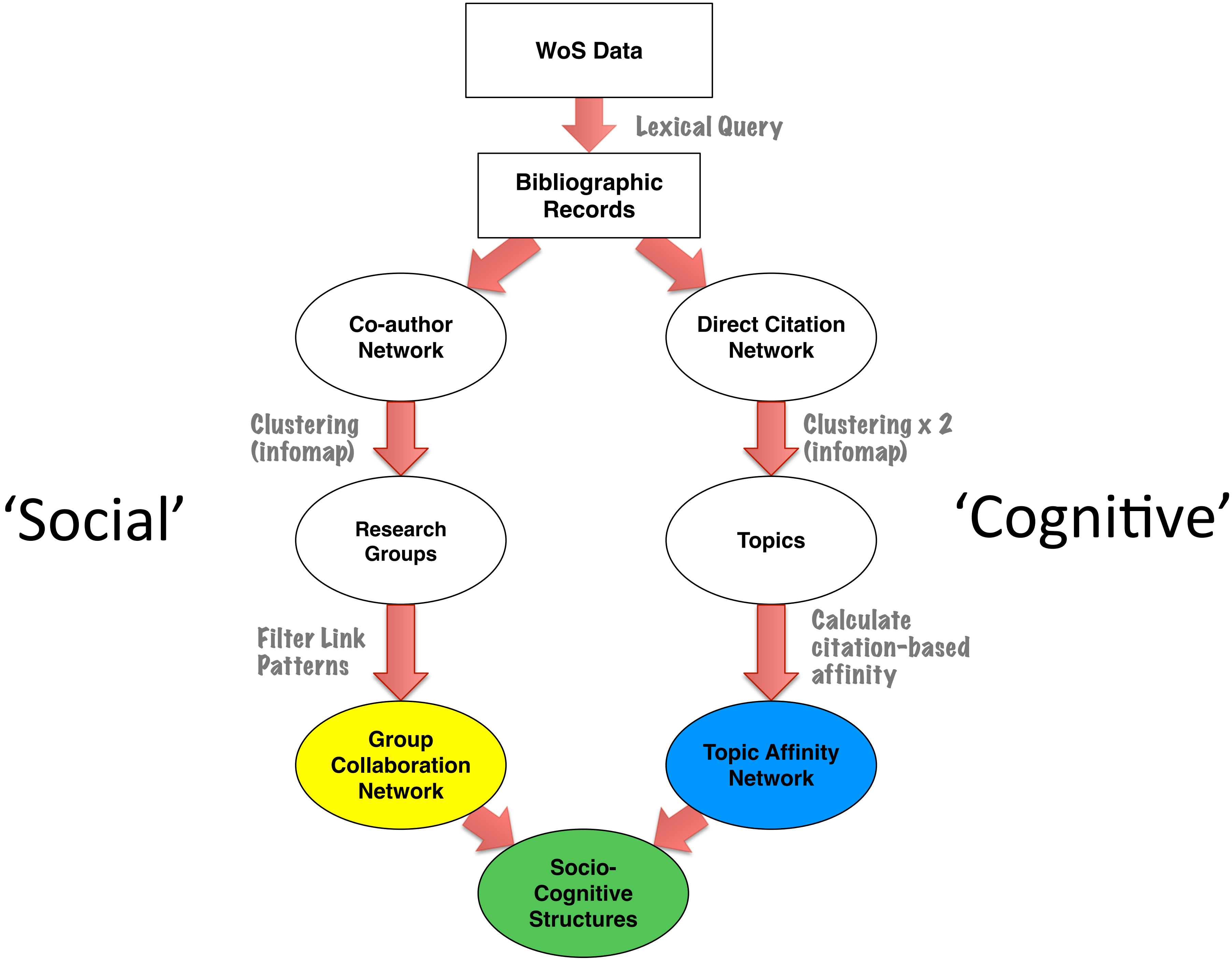
of scientific communities




Ethnographic Field Studies
Network Analysis
Step 1
FIELD DELINEATION
Local practices and research culture
inform
Document citation network and authorship
Step 2
PATTERNS OF SOCIAL AND COLLABORATIVE ORGANIZATION
Real life scenarios, interviews; new research questions
interpretation
visualization
quantification
Network features: co-author clusters, group collaboration network; refinement of algorithm
Step 3
SCALING-UP OF OBSERVATIONS
Sampling of field sites, interview partners, Interpretation of observations
interpretation
contextualize
Community Structures
local practices, participants' experiences
aggregate patterns
Step 4
SYNTHESIS
Field Differences

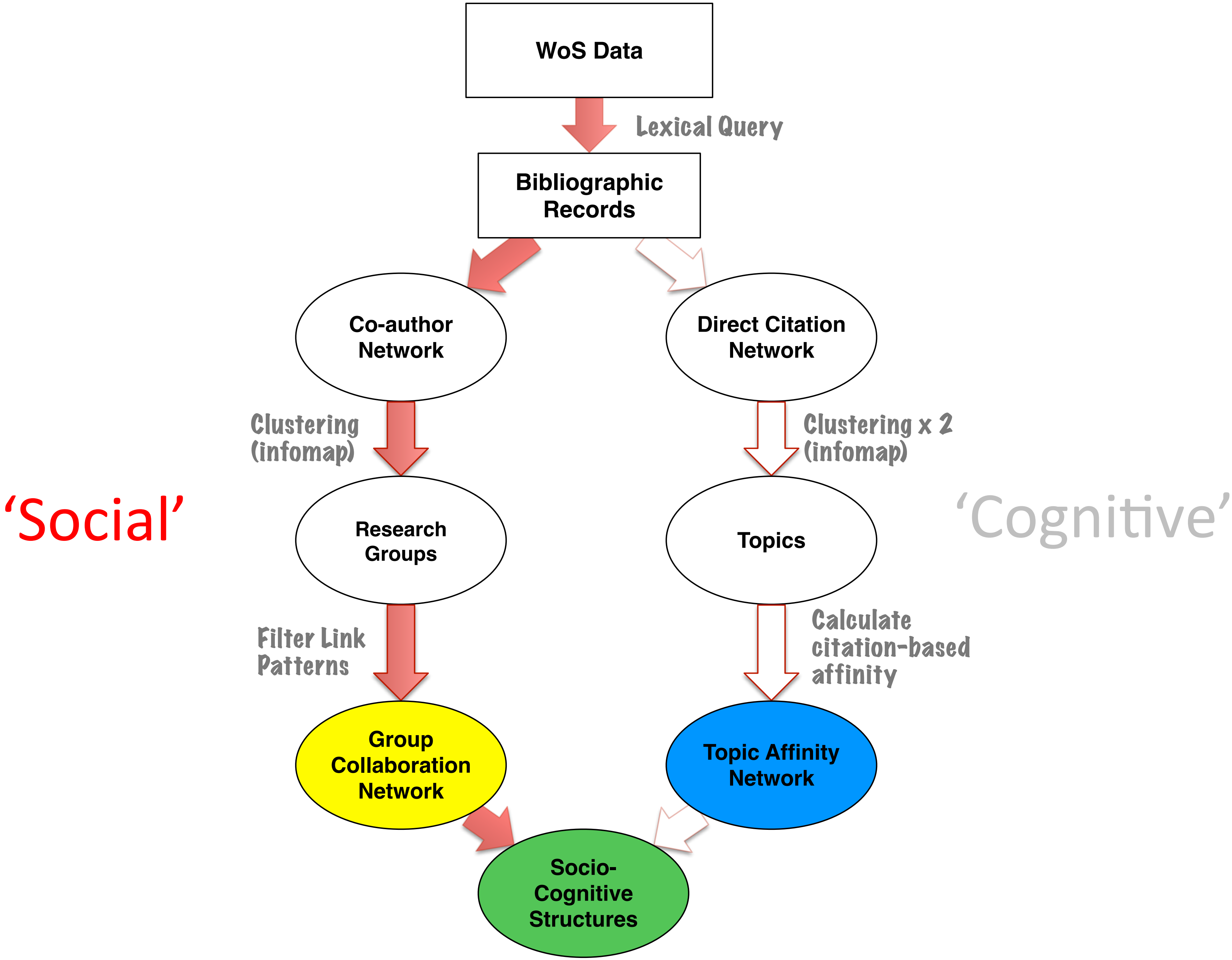
Co-author & citation networks

PART 1: MAPS OF SCIENCE

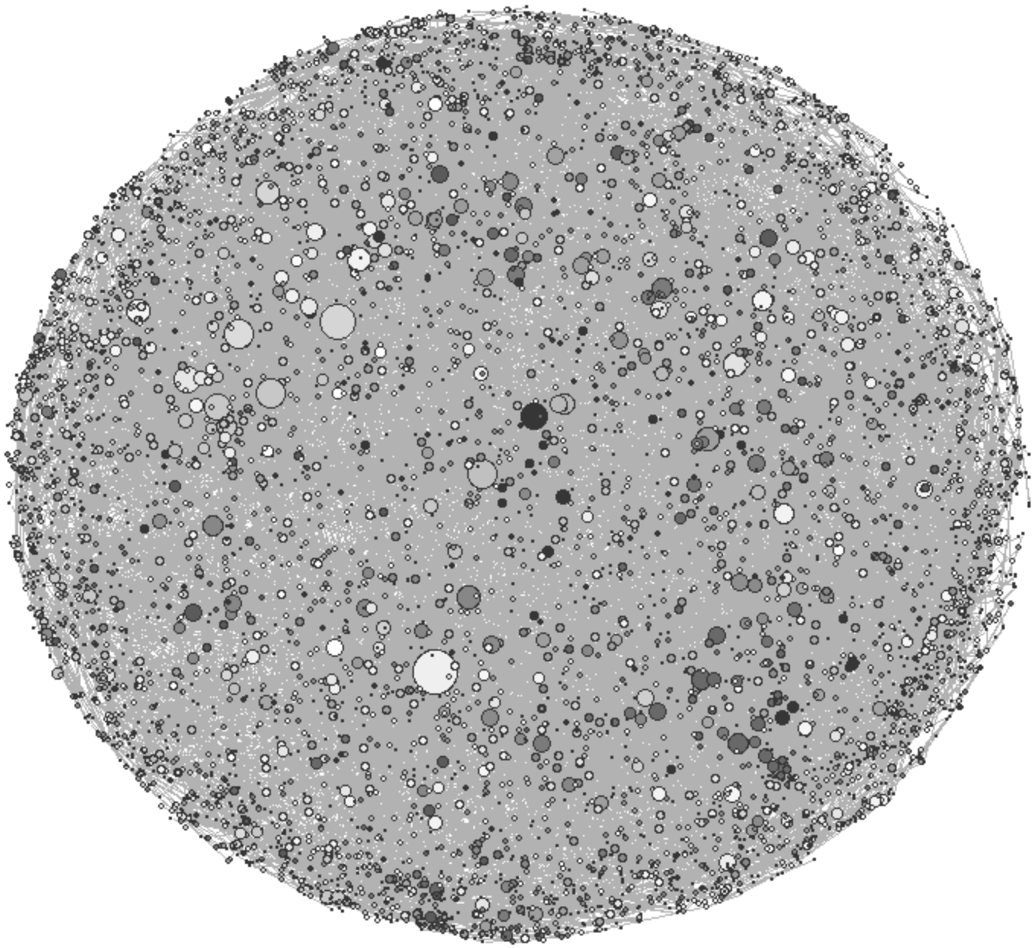

WoS Data
Lexical Query
Bibliographic Records
Co-author Network
Direct Citation Network
Clustering (infomap)
Clustering x 2 (infomap)
'Social'
Research Groups
Topics
'Cognitive'
Filter Link Patterns
Calculate citation-based affinity
Group Collaboration Network
Topic Affinity Network
Socio-Cognitive Structures

Data To Represent a Research Specialty

- Science Citation Index Expanded (SCI) edition, Web of Science (October 2013)
- Lexical query on title field
  - 20-year period (1991 - 2010)
  - Developed during ethnographic field studies between 2007-2009 to capture two research specialties in the physical and chemical sciences
  - Optimized recall and precision (Velden & Lagoze, JASIST 2013)
- Data preprocessing:
  - Include only records of type 'article'
  - Author name disambiguation (Velden et al, JCDL 2011)
  - Remove transient, one-time authors (~ 60%)
  - Final data sets:
    - For field A: **55,648 records** and **40,808 unique authors**
    - For field B**: 13,910 records** and **9,116 unique authors**


WoS Data
Lexical Query
Bibliographic Records
Co-author Network
Direct Citation Network
Clustering (infomap)
Clustering x 2 (infomap)
'Social'
Research Groups
Topics
'Cognitive'
Filter Link Patterns
Calculate citation-based affinity
Group Collaboration Network
Topic Affinity Network
Socio-Cognitive Structures

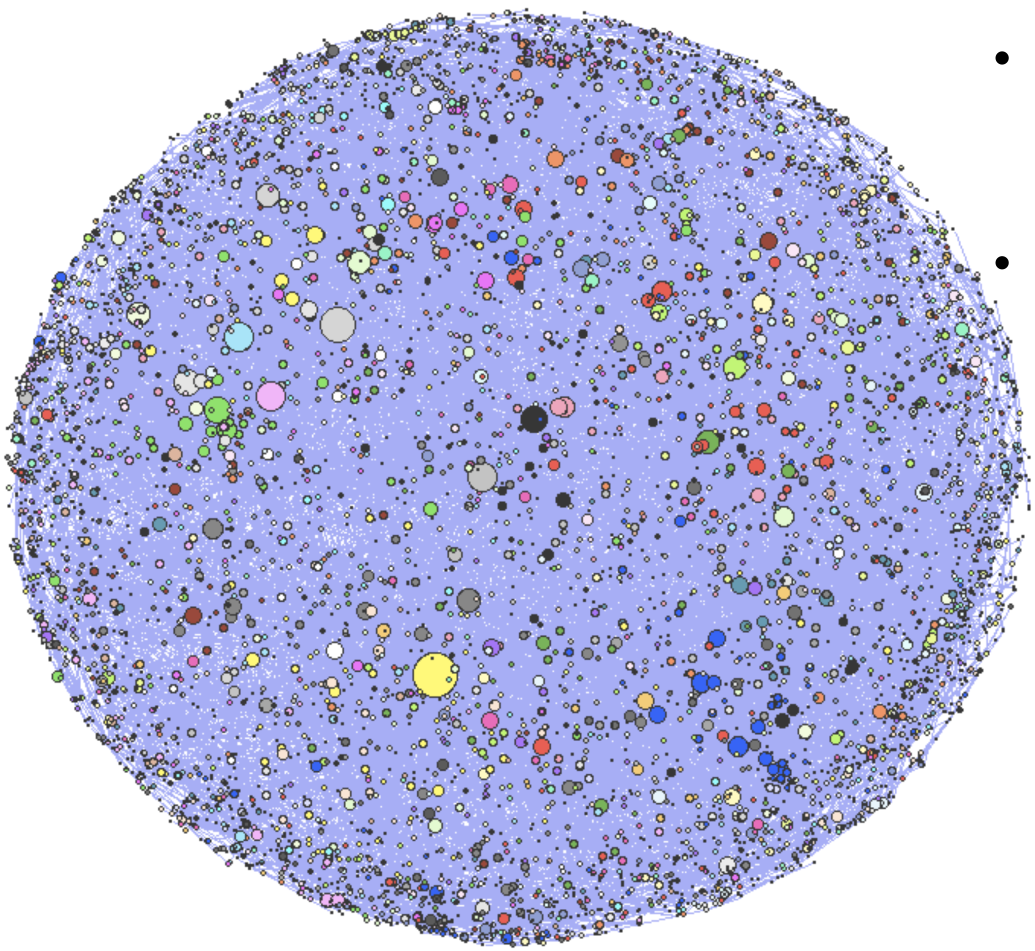
Co-Author Network



- weighted (weight =1 per co-authored paper)
- undirected
- Field B: ~ 7,000 authors in giant component

Visualization: pajek, Fruchterman Rheingold algorithm

Clustered Coauthor Network

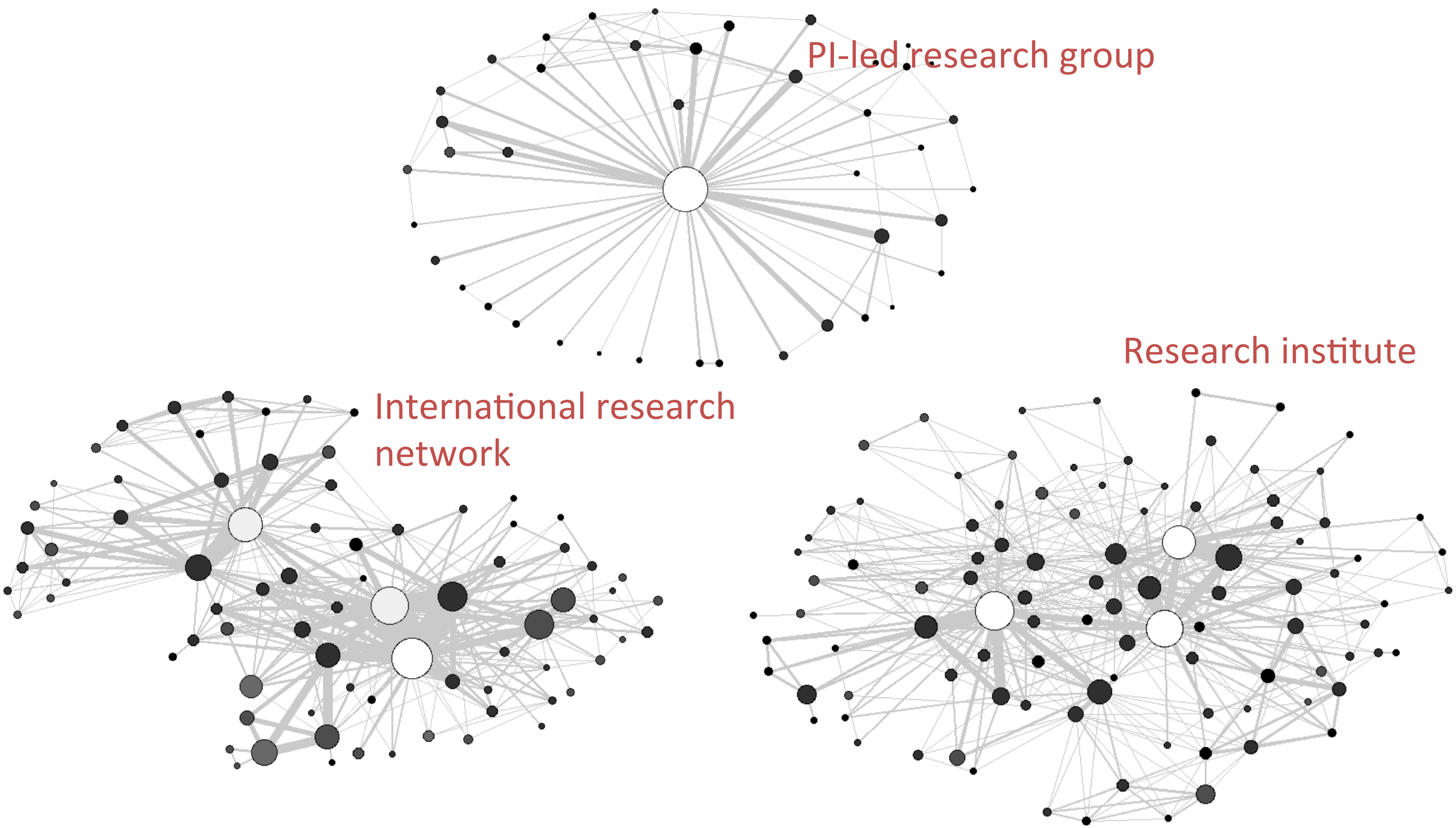


- Apply clustering algorithm to extract groups of closely collaborating authors
- Key properties of infomap algorithm:
  - Disjoint clusters
  - Unbiased cluster size
  - Fast

Clustering: Rosvall, M., & Bergstrom, C. (2007). An information-theoretic framework for resolving community structure in complex networks. PNAS, 104(18), 7327.

Co-Author Clusters

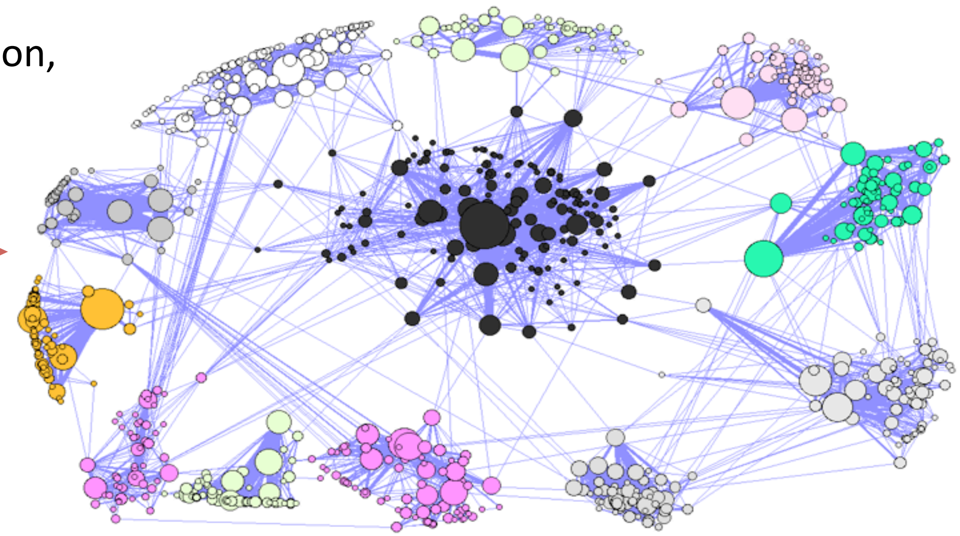
‘functional research groups’ [Seglen & Aksnes 2000, microbiology]

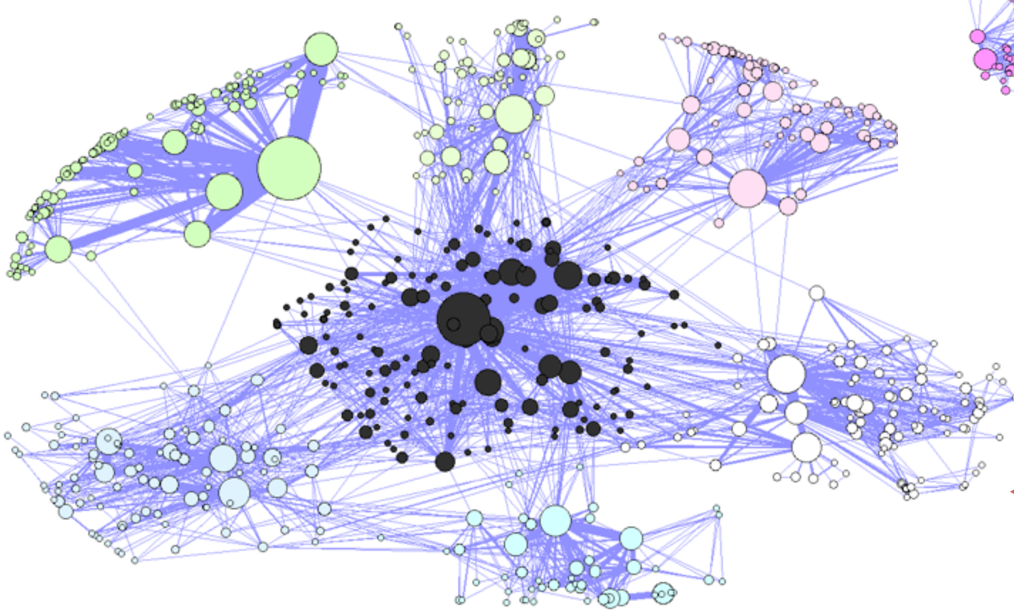


Mesoscopic Structure

Linking patterns between groups

Transfer links: career migration, sample exchange, measurement services





Inter Group Collaboration: Intensive collaboration between subgroups

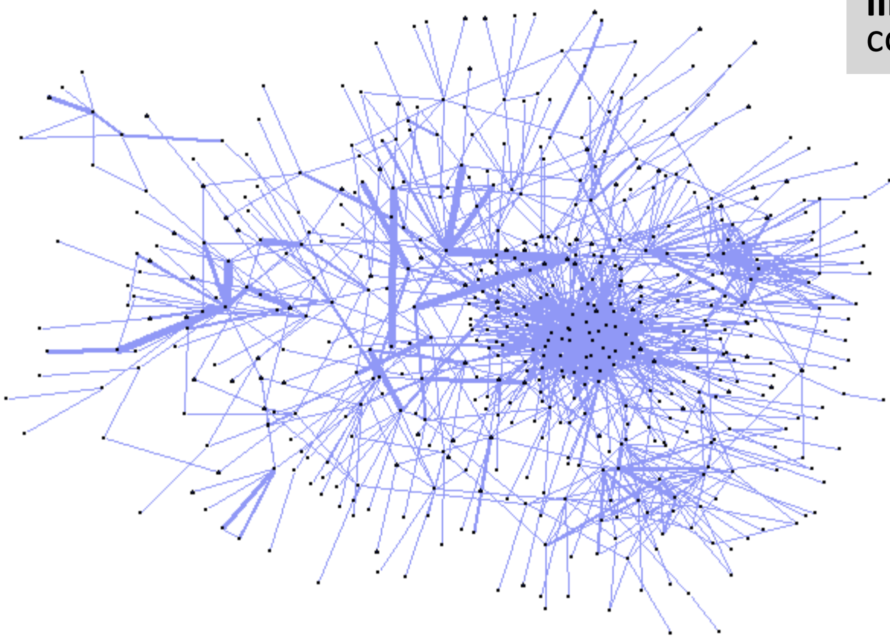
Velden, Haque, & Lagoze, *Scientometrics 85(1) 2010*

Field Differences: Group Collaboration Network

nodes: research groups

links: collaboration

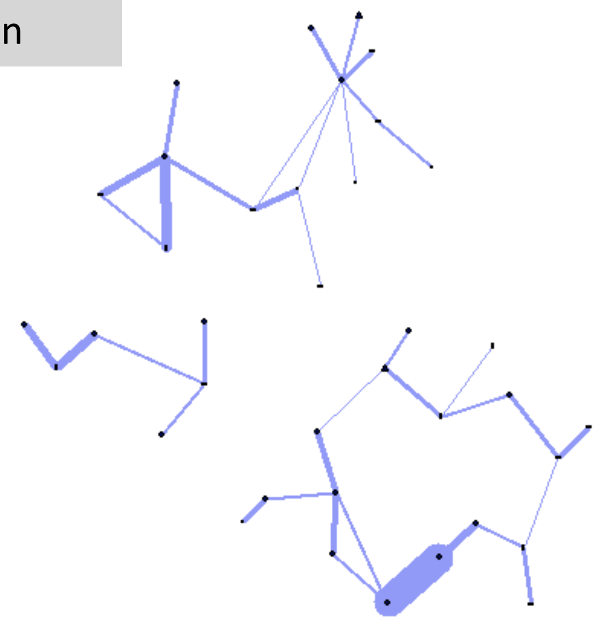
Field A



~ **23%** of groups from the giant component of the co-author network collaborate

Large giant component

Field B

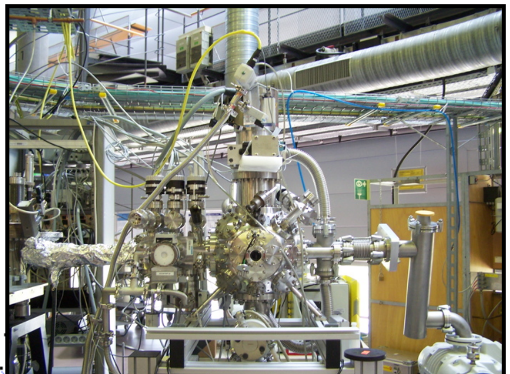


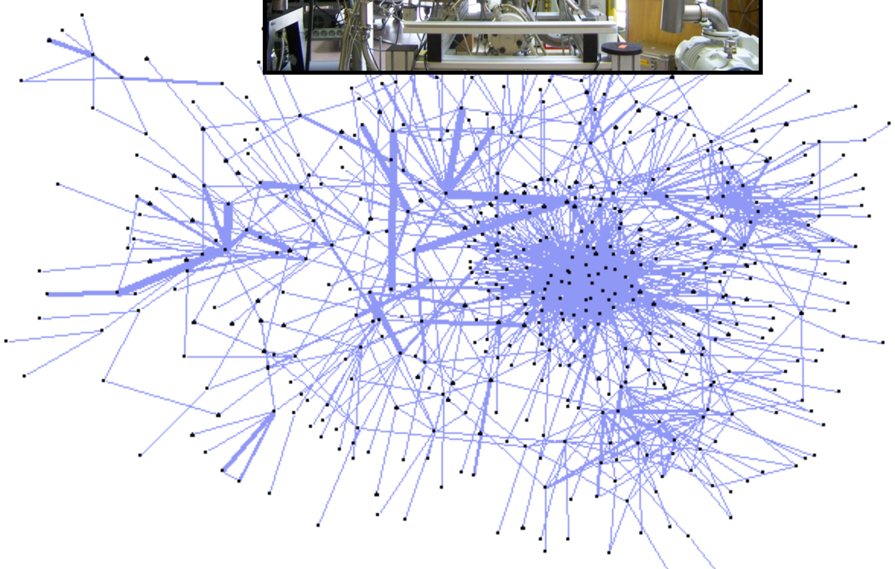
~ **9%** of groups from the giant component of the co-author network collaborate

Small unconnected components

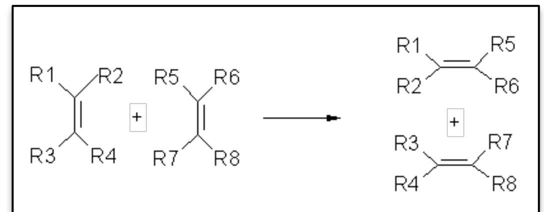
Field Differences: Group Collaboration Network

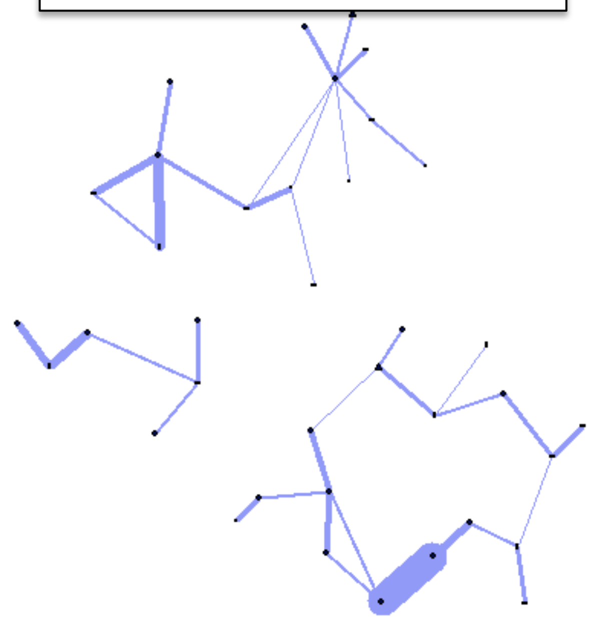
Field A: experimental physics

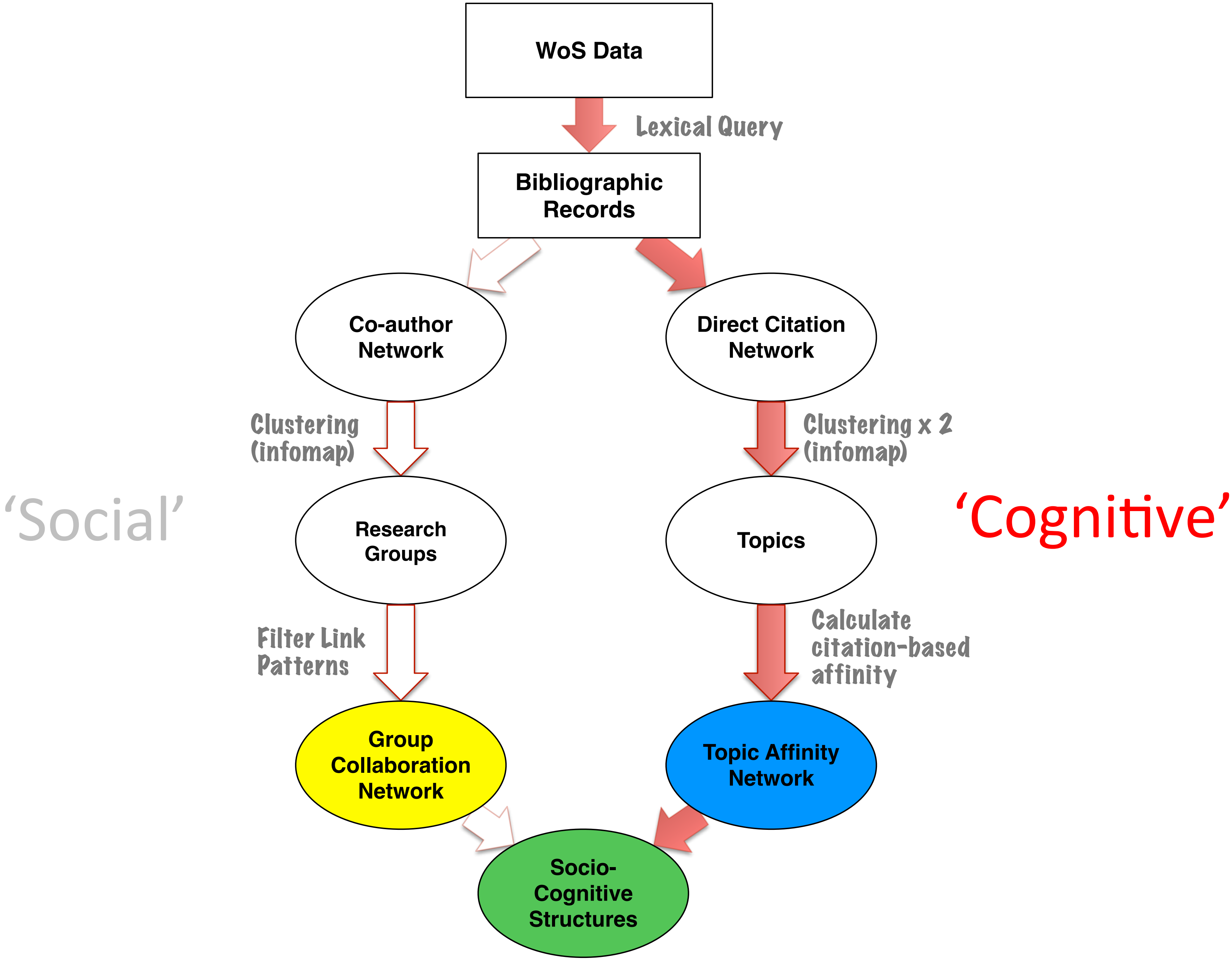




Field B: synthetic chemistry






WoS Data
Lexical Query
Bibliographic Records
Co-author Network
Direct Citation Network
Clustering (infomap)
Clustering x 2 (infomap)
'Social'
Research Groups
Topics
'Cognitive'
Filter Link Patterns
Calculate citation-based affinity
Group Collaboration Network
Topic Affinity Network
Socio-Cognitive Structures

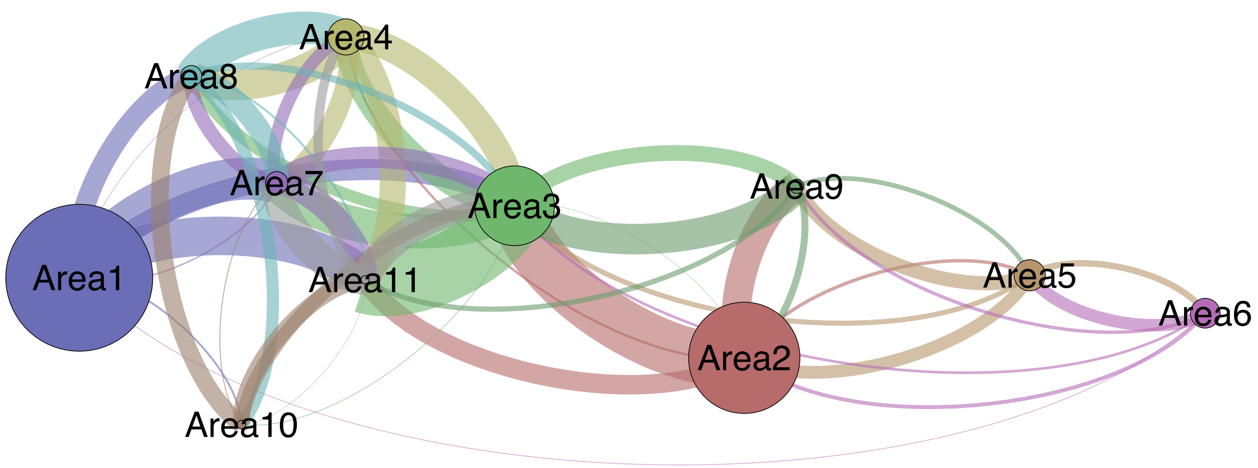
Topic Affinity Map

Topic: Clusters of clusters of documents (twice clustered citation network; infomap clustering algorithm)

Affinity: disproportionally strong citation links

Affinity ($area_i \rightarrow area_j$) := (actual count – expected count) / √ (expected count)^2

where expected count is proportional to relative size of $area_j$



Velden T. & Lagoze, C. (JASIST, 2013)

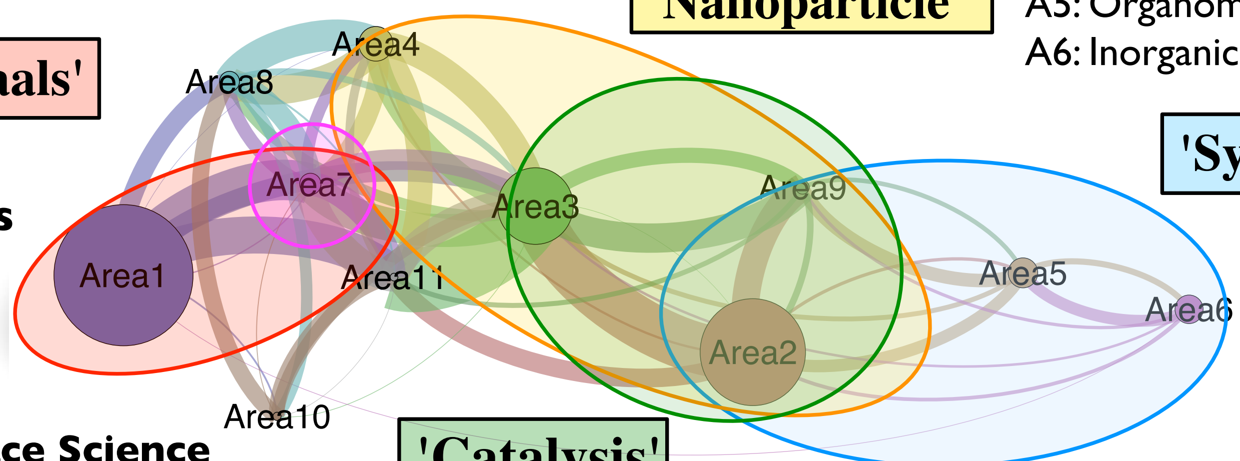
Visualization: gephi, Force Atlas 2 Layout algorithm

Topic Affinity Map

‘Disciplinary Orientations’

Surface Science
A3: surface science
A4: surface science

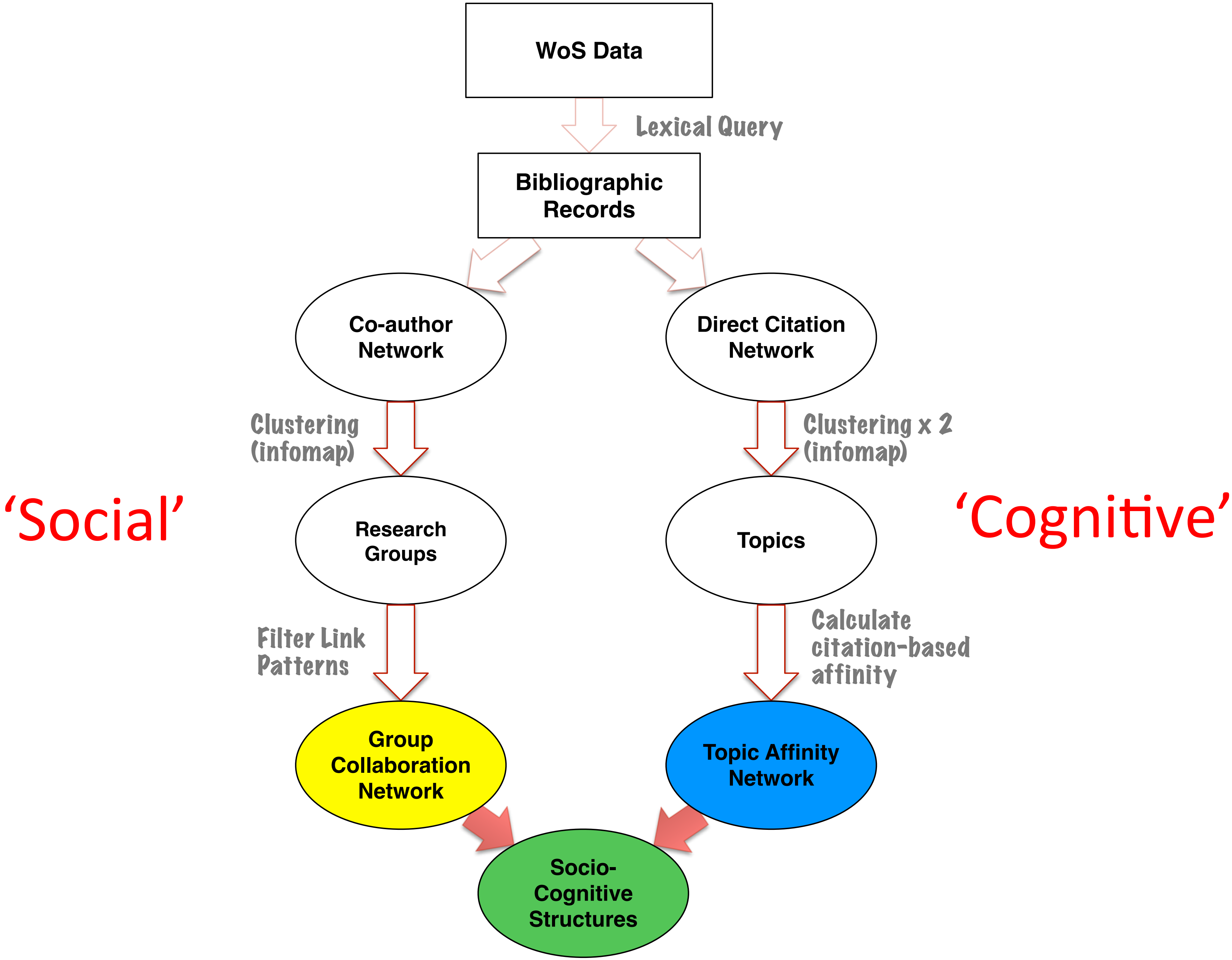
Cluster Chemistry
A9: Catalysis
A5: Organometallics
A6: Inorganic Chemistry



Cluster Physics
A1: cluster physics
A7: dynamics

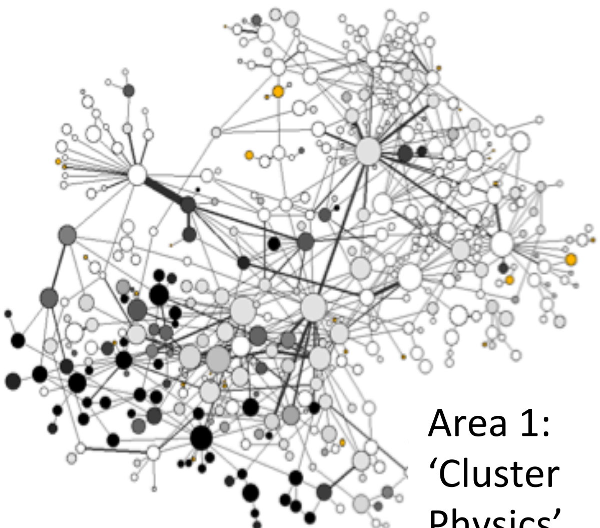
Surface Science /Cluster Physics
A8: incl. fullerene
A11: Condensed Matter/Materials Science
A10: Interfaces, interactions

Materials Chemistry
A2: materials chemistry

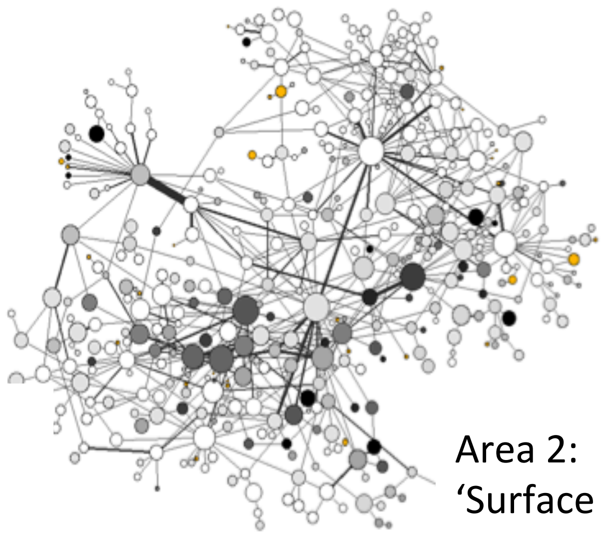

WoS Data
Lexical Query
Bibliographic Records
Co-author Network
Direct Citation Network
Clustering (infomap)
Clustering x 2 (infomap)
'Social'
Research Groups
Topics
'Cognitive'
Filter Link Patterns
Calculate citation-based affinity
Group Collaboration Network
Topic Affinity Network
Socio-Cognitive Structures

Visualization of the socio-cognitive fabric of a research field

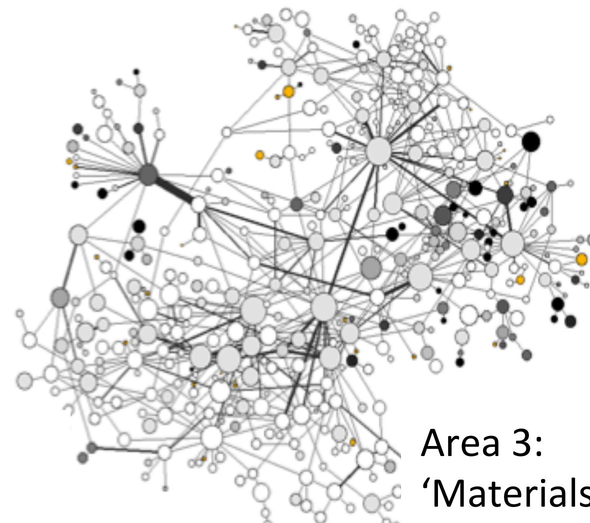
Disciplinary Differential in Cohesiveness?



Area 1: 'Cluster Physics'



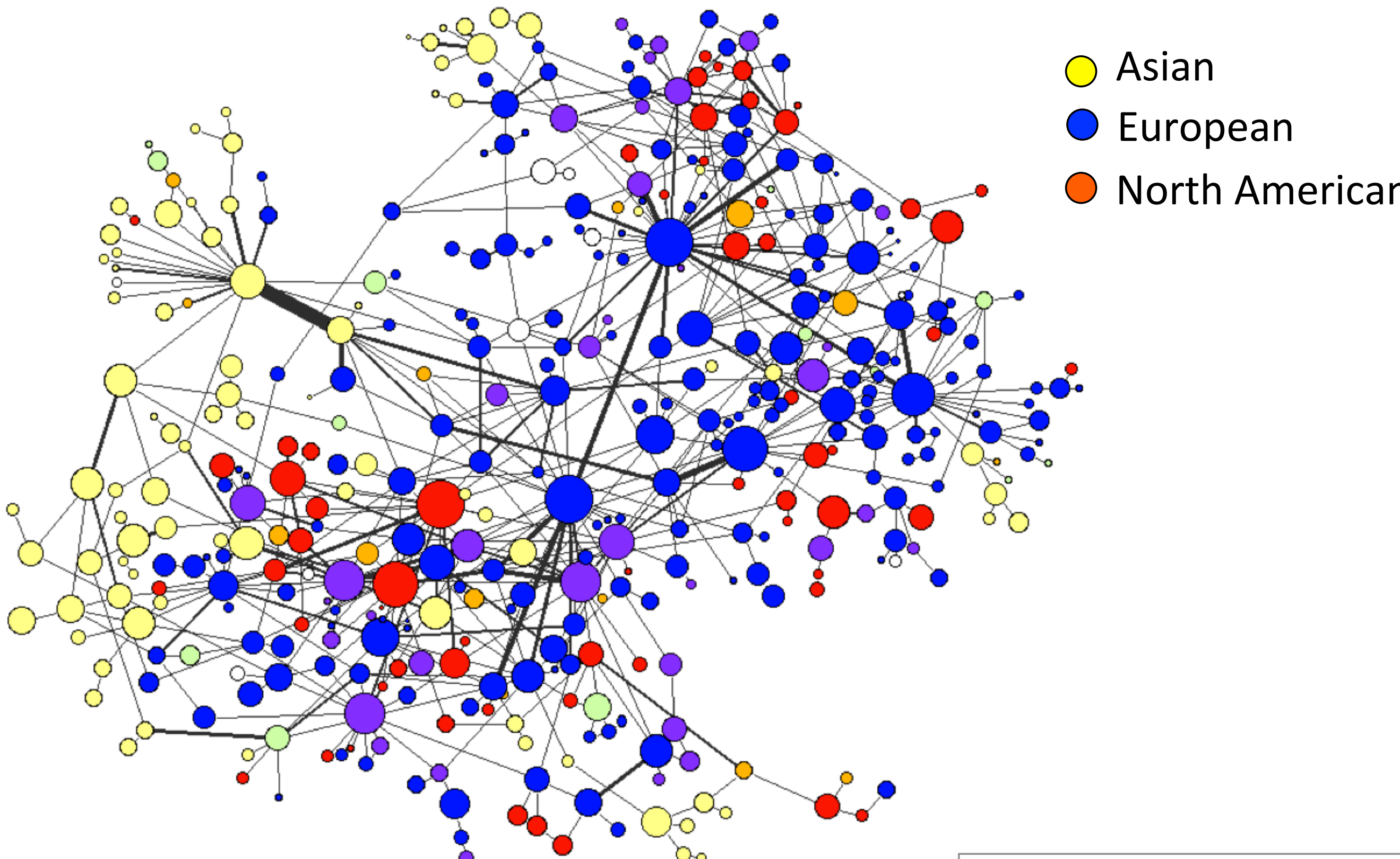
Area 2: 'Surface Science'



Area 3: 'Materials Chemistry'

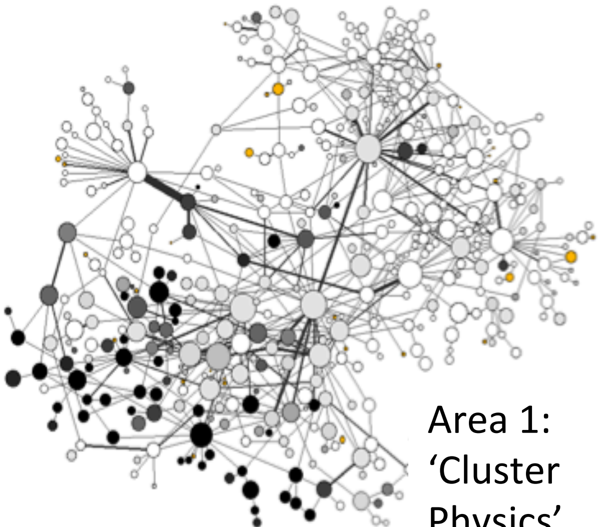
Group collaboration networks: color of nodes indicates intensity of (publication) activity of a group in the respective topic area.

International Group Collaboration Network

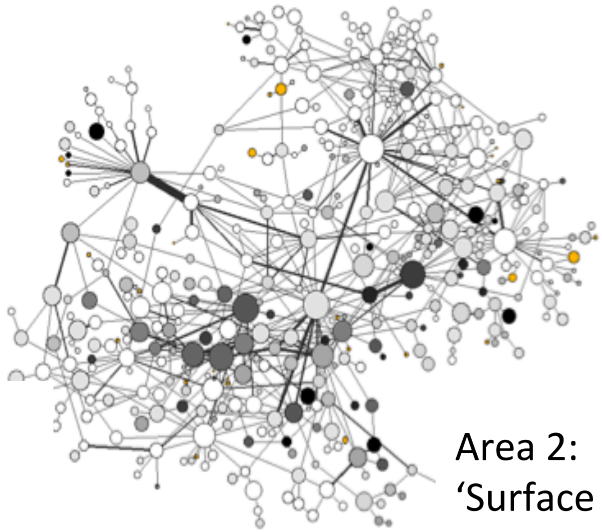


Visualization: pajek, Kamada-Kawai algorithm

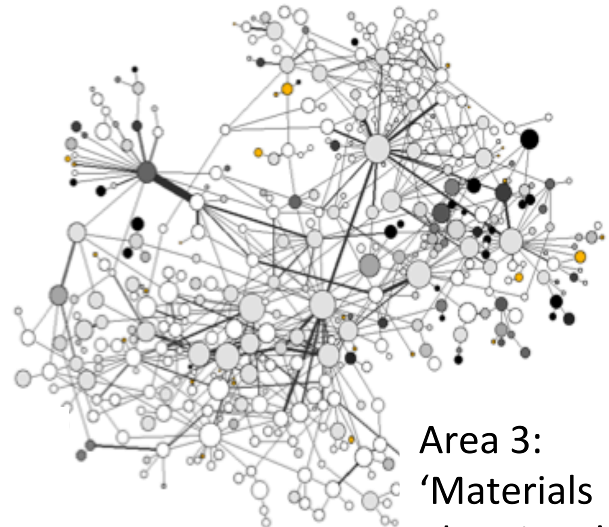
Topical versus Geographic Ordering of Collaboration Links



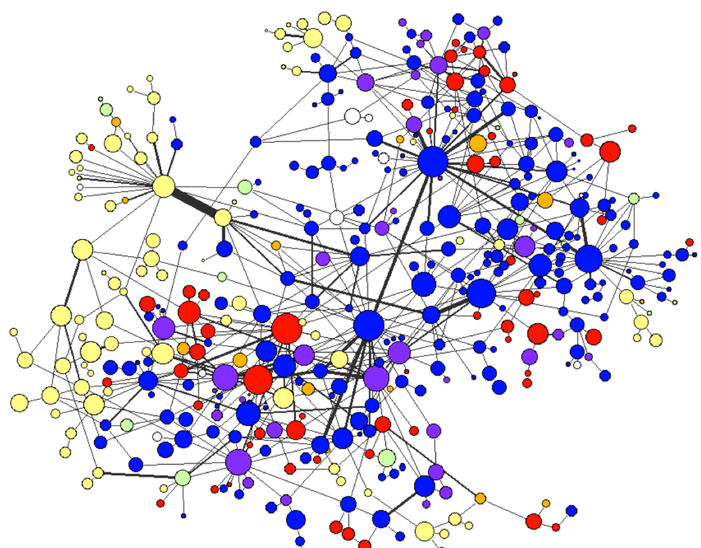
Area 1: 'Cluster Physics'



Area 2: 'Surface Science'



Area 3: 'Materials Chemistry'



Geographic affiliation

Data & Methods

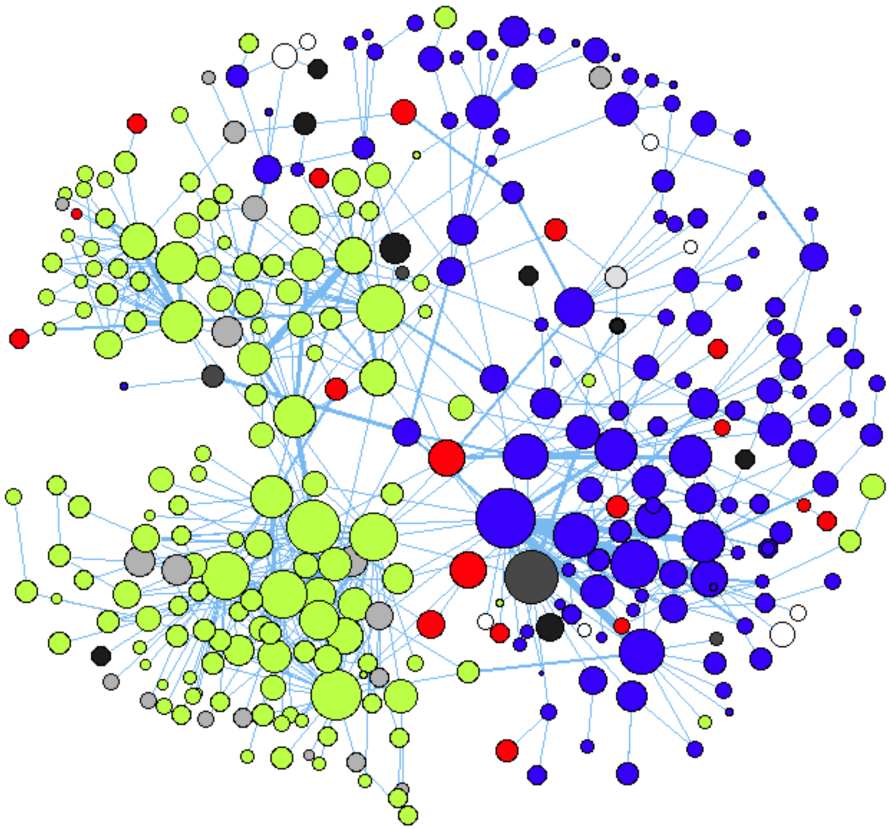
PART 2: MAJOR CHALLENGES IN THE MAKING OF MAPS OF SCIENCE

Data!

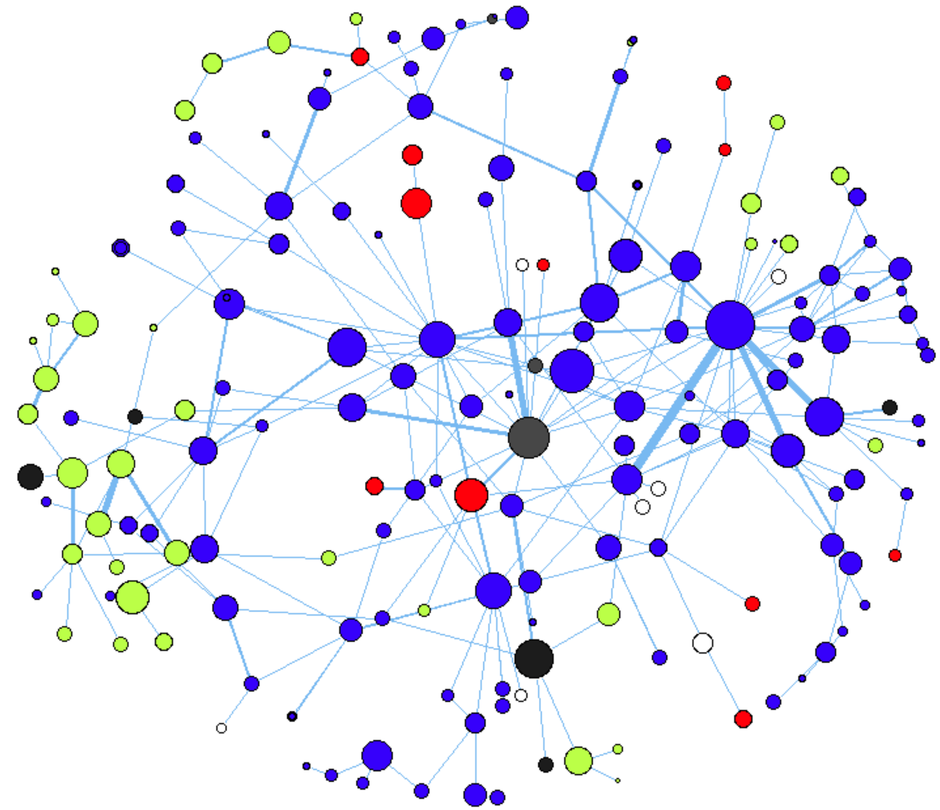
- Selection of Data:
  - Field delineation: how adequate is a data set to represent a field?
    - Research specialties have fuzzy boundaries (dynamic, overlapping, poly-hierarchical)
    - subject classification usually insufficient
    - Most thorough approaches (growing from seed) require comprehensive database access
- Access:
  - Can others reproduce or expand on my results?
- Quality:
  - Are references uniquely identified?
  - Author name disambiguation

Author Name Disambiguation

Before



After disambiguation



Proportion of Asian affiliated author clusters: reduced from 43% to 19%
average node degree decrease from 3.9 to 2.8

Velden, T., Haque, A. & Lagoze, C. (2011) Resolving Author Name Homonymy to Improve Resolution of Structures in Co-author Networks. JCDL 2011

Methods!

- Need for ‘benchmarking’ and validation
  - Often developed and fine-tuned in-house with lack of replication
  - Usually data set not available for replication
  - Limited understanding of origin and scale of differences in results obtained by different approaches

Example: Topic extraction

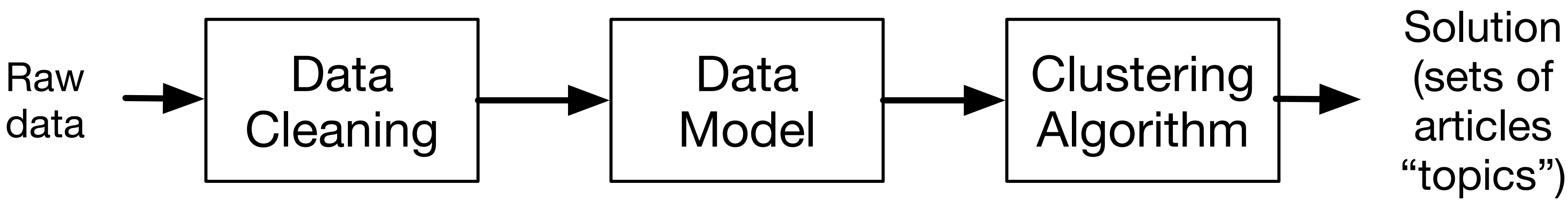
Same data, different results?

- How to group publications algorithmically into topics?
- Ongoing collaboration of several scientometric groups
- Start from same raw data set: ~ 111.000 publications from 59 journals in astronomy and astrophysics (Web of Science), 2003-2010

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)

Workflow(s) for Topic Extraction

Sources for Variation



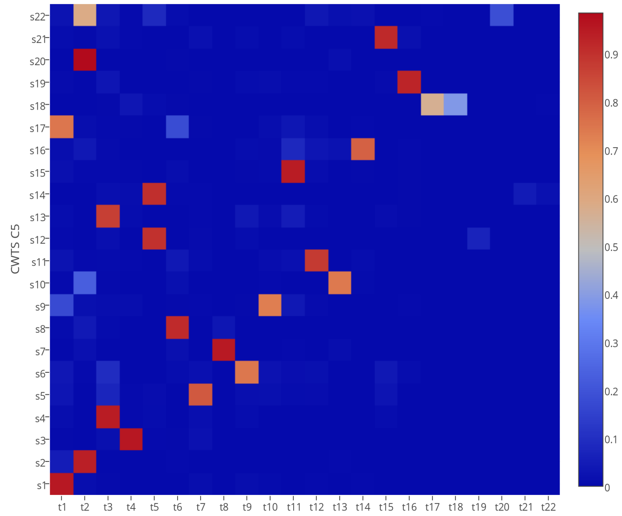
Data Cleaning:
Matching with in-house database
Reference resolution

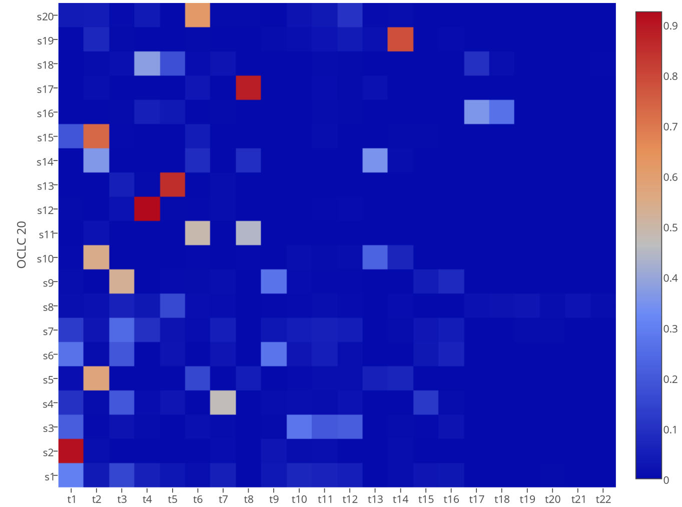
Data Model:
Bibliographic coupling
Direct Citation
Hybrid (bc & NLP)
Semantic Matrix
(incl normalizations and thresholds)

Clustering Algorithm:
Louvian
Infomap
SLMA
k-means
memetic
(incl. resolution and parameter settings)

Sources of variability between solutions

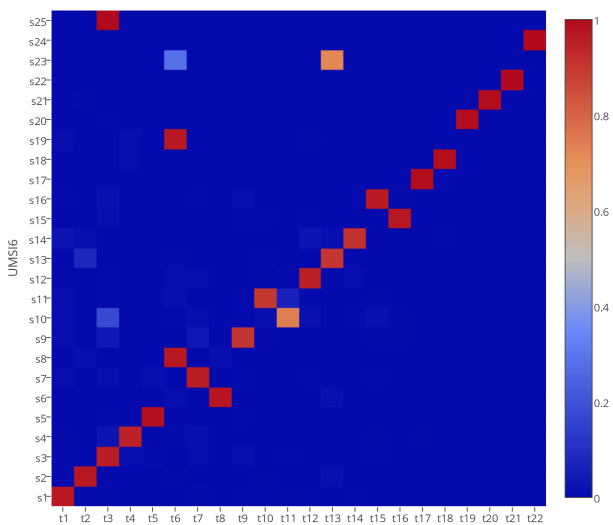
Same model & **different algorithm**

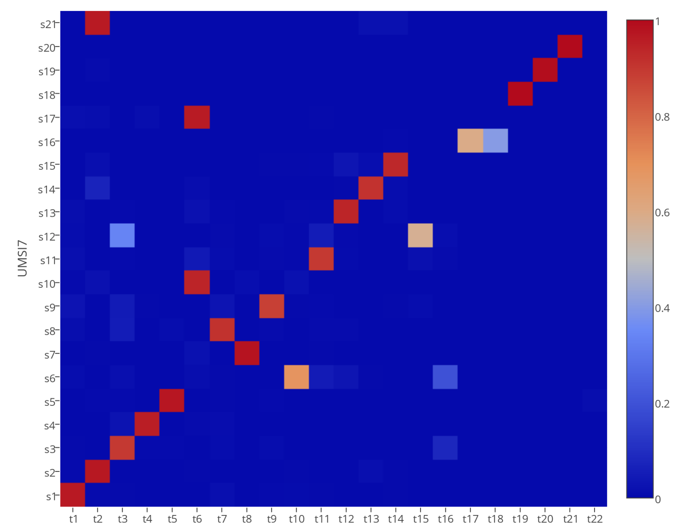




Different model & **different algorithm**

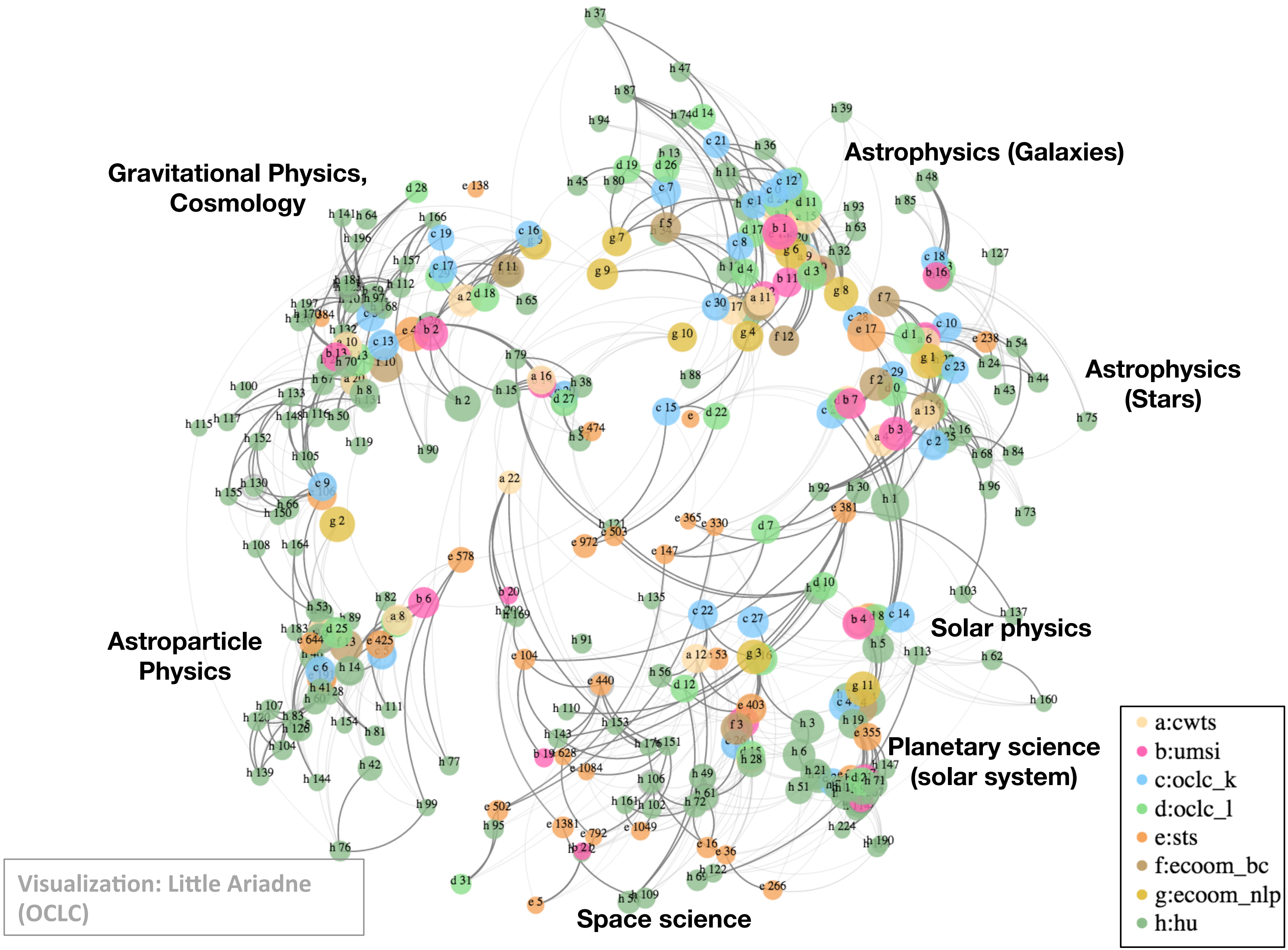
Same model & **same algorithm**





Same model & **same algorithm**

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


Gravitational Physics, Cosmology
Astrophysics (Galaxies)
Astrophysics (Stars)
Astroparticle Physics
Solar physics
Planetary science (solar system)
Space science
a:cwts
b:umsi
c:oclc_k
d:oclc_l
e:sts
f:ecoom_bc
g:ecoom_nlp
h:hu
Visualization: Little Ariadne (OCLC)

Stay tuned…

- Work in progress
- Special Issue for Scientometrics in Preparation
- In Planning: Topic extraction challenge
  - Invitation to other groups to provide their solutions for comparison

Conclusions

Visualizations for Science Policy

- Great potential for science maps, especially as an explorative and hypothesis generating tool
- Careful validation a key concern
  - Require comprehensive access to data to enable reproducibility and comparison
  - Need more rigorous comparison of methods
  - Benefit from mixed methods to ground interpretations

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