



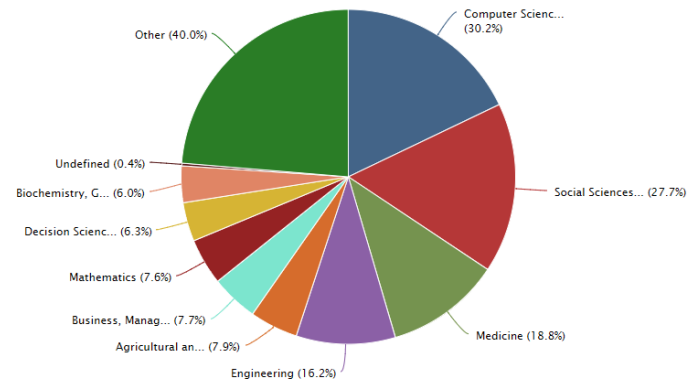
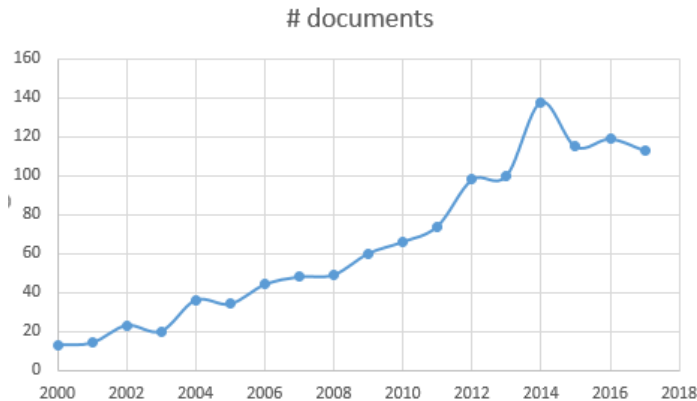
UNDERSTANDING THE DYNAMICS OF SCIENCE

**Interdisciplinary Workshop
November 23th, 2017**



UNDERSTANDING THE DYNAMICS OF SCIENCE

Scopus: (TITLE-ABS-KEY (scientometric*) OR TITLE-ABS-KEY (bibliographic*) OR TITLE-ABS-KEY (bibliometric*)) AND TITLE-ABS-KEY (dynamic*)



➔ A topic of growing interest, involving researchers from many fields.

Our goals today:

- Discuss the relationship between our different approaches
- Identify and discuss promising strategies to analyze and visualize the DYNAMICS of scientometric networks

AGENDA

9:30-10:00 / *Meet up & coffee*

10:00-11:00 / S. Grauwin & P. Jensen (IXXI, Lyon). *Brief review of current approaches & work in progress*

11:00-12:00 / Marta Sales-Pardo (SEES, Tarragona). *From complex networks to scientometrics analysis*

12:00-13:30 / *lunch*

13:30-14:30 / Marion Maisonobe (Toulouse). *Geography of science.*

14:30-15:30 / J-P (Paris, Médialab) Cointet & Ale Hannud Abdo. *Tools for science dynamics.*

15:30-16:30 / P-P Combes (GATE, Lyon). *Analyzing scientometrics data with econometric tools.*

16:30 / GENERAL DISCUSSION

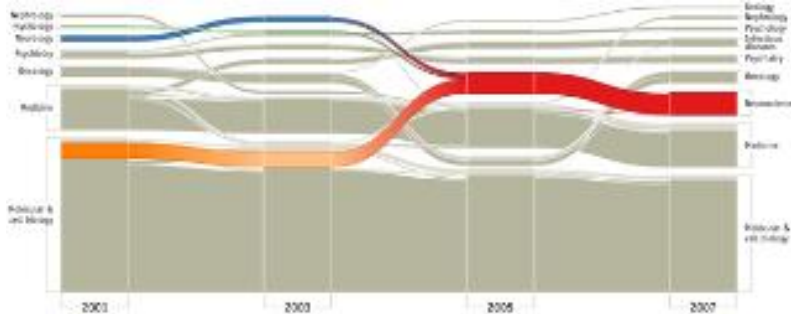
DYNAMICS OF SCIENTIFIC COMMUNITIES

A BRIEF REVIEW OF CURRENT APPROACHES + WORK IN PROGRESS

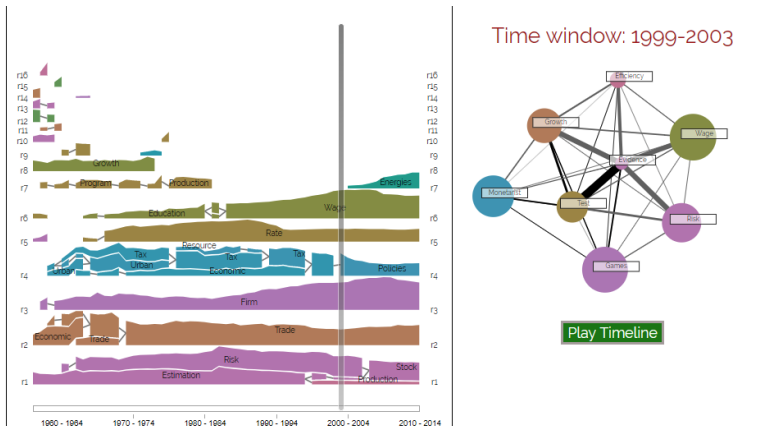
Sebastian Grauwin (IXXI, ENS Lyon)

MOTIVATION

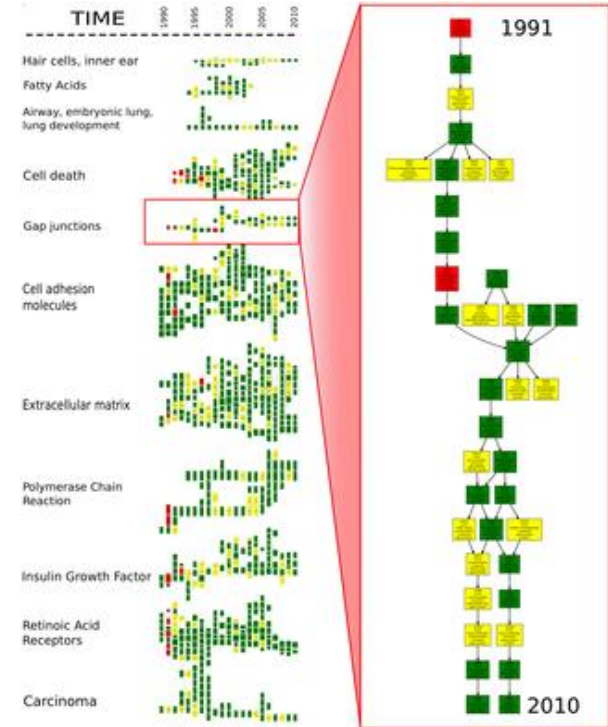
GOAL: BUILDING A « PERTINENT » HISTORY OF SCIENCE



Rosvall & Bergstrom (2010), *Mapping change in large networks*



Claveau & Gingras (2016), *Macrodynamics of Economics: A Bibliometric History*



Chavalarias & Cointet (2013), *Phylomemetic Patterns in Science Evolution*

3 ESSENTIAL STEPS

INTERDISCIPLINARITY NEEDED 😊



DEFINING scientific communities

→ « Social science »



DETECTING scientific communities

→ « Computer science »



VISUALIZING scientific communities

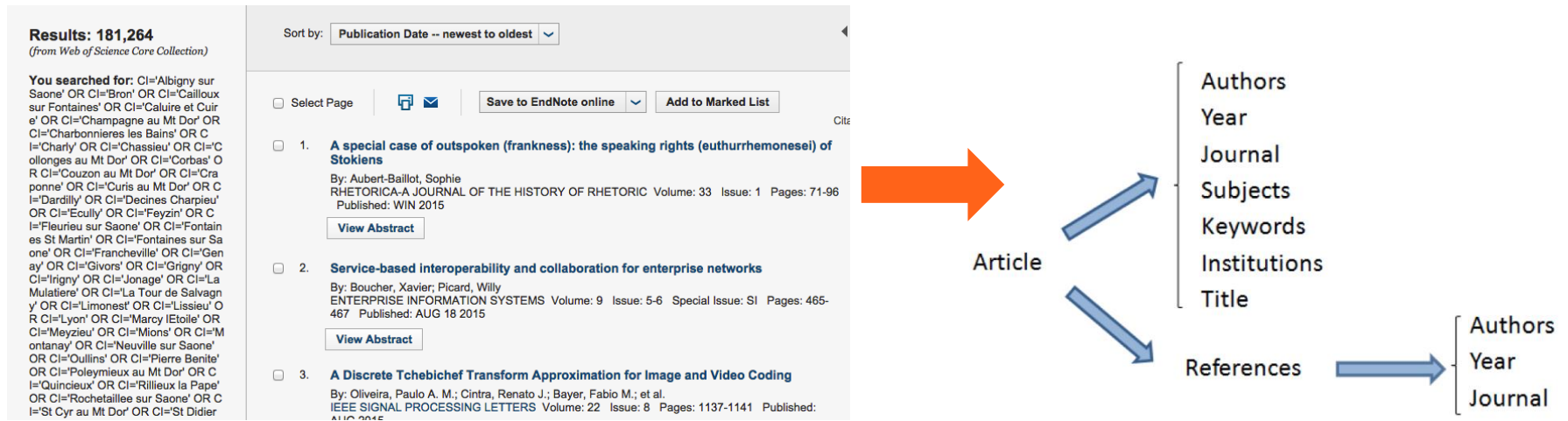
→ « Design »

DEFINING

scientific communities

WHICH DATA ?

BUILDING BLOCKS FOR A QUANTITATIVE DESCRIPTION OF SCIENCE



➔ **Bibliographical data**, typically extracted from WOS or Scopus

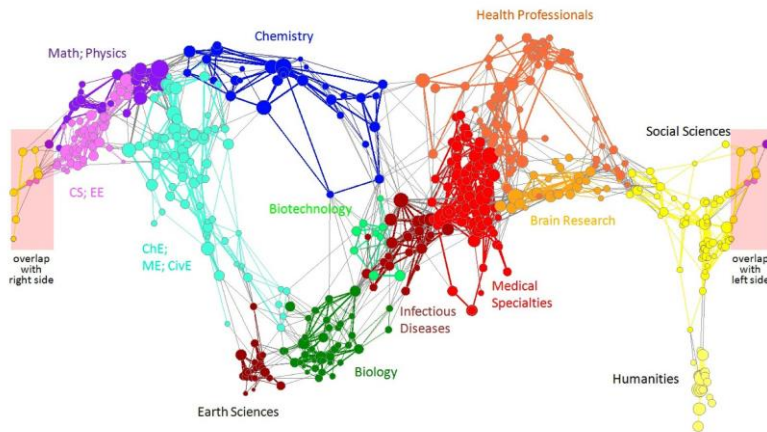
Alternatives data: patents, newspapers, social media, etc

WHICH DATA ?

CORPUS SELECTION

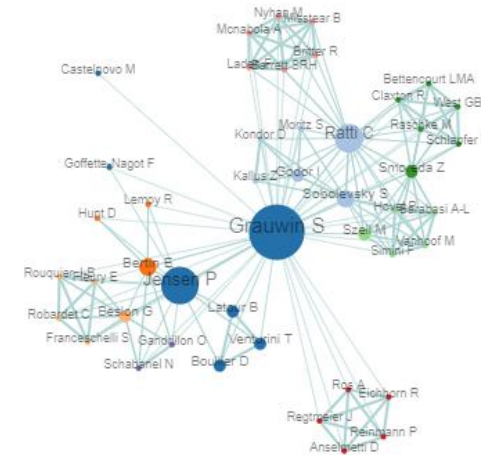
GLOBAL MAPS OF SCIENCE

Börner & al; Leydesdorff & al



MORE FOCUSED MAPS

on a discipline / field, an institution / authors, etc



- ➔ Corpus defined on a query on keywords, authors, journals, categories, etc
- ➔ **Corpus with different scales & ranges**

WHICH COMMUNITIES?

Communities = individuals

➔ Studies of individuals careers

Professional communities

- Co-occurrence of authors
- Co-occurrence of institutions / labs

Communities = Knowledge fields

- Co-occurrence of words/ keywords
- Citations maps
- Word / keywords coupling maps
- Co-citations maps
- **Bibliographic Coupling maps**

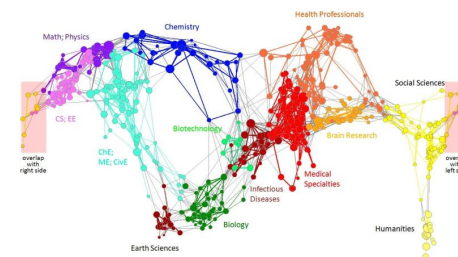
≠ aggregation scales (publis, journals, categories)



R Sinatra & al (2017),
Data-driven prediction in the science of science



O Beauchesne (2014),
Map of scientific collaboration



Börner & al (2012), The
UCSD Map of Science

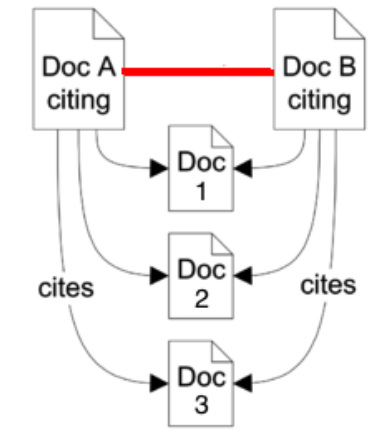
WHICH COMMUNITIES?

TIME DEPENDENCIES

Bibliographic coupling: Similarity measure based on shared references.

- Fixed once and for all (does not depend on future citations)
- Only use info chosen by the authors
- All papers can be taken into account in the map
- Recent papers can readily appear in the map

➔ « Fair » treatment of papers with \neq publication years

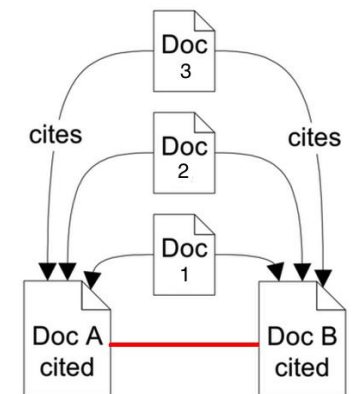


Bibliographic Coupling

Co-citation: Similarity measure based on shared citing publications.

- Depend on studied corpus
- Will evolve in time, with the accumulation of citations
- Only cited papers will be taken into account

➔ Biased treatment of papers with \neq publication years

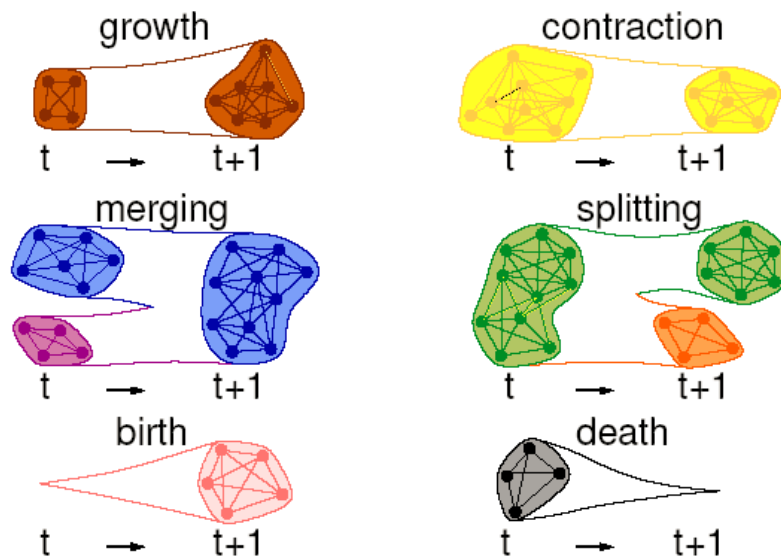


Co-citations

WHICH « HISTORICAL EVENTS »?

DYNAMICS AS « LIFE-CYCLE » STRUCTURAL CHANGES

Two-steps signatures



Palla et al.(2007). *Quantifying social groups evolution.*

N-steps signatures



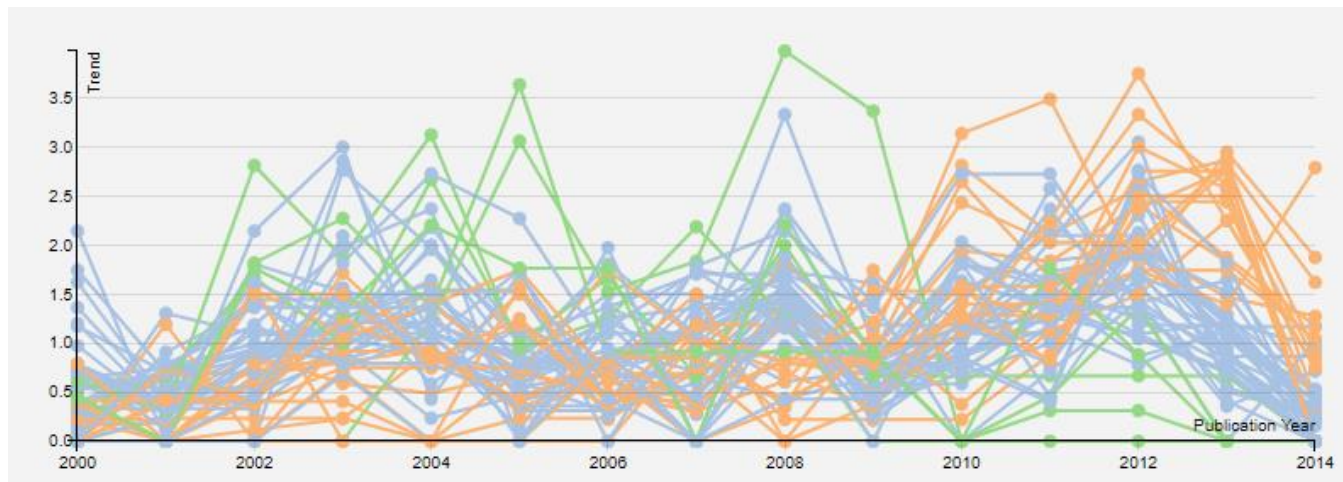
Vehlow et al. (2015) *Visualizing the Evolution of Communities in Dynamic Graphs*

WHICH « HISTORICAL EVENTS »?

DYNAMICS AS EVOLUTION OF CONTENT

Trending keywords, references, journals, etc within a community may change with time.

➔ These internal factors drive structural changes



Grauwin et al. (ongoing) Trends of top keywords in an “Educmap” cluster

DEFINING SCIENTIFIC COMMUNITIES

We are interested in **thematic communities**. Ideally, we want to be able to detect and visualize their **hierarchical structures**, as well as **their internal and structural dynamics**.

Let's focus on Bibliographic Coupling:

What can we do?

DETECTING

scientific communities

STATIC CLUSTERING

BUILDING THE BIBLIOGRAPHIC NETWORK

Standard approach

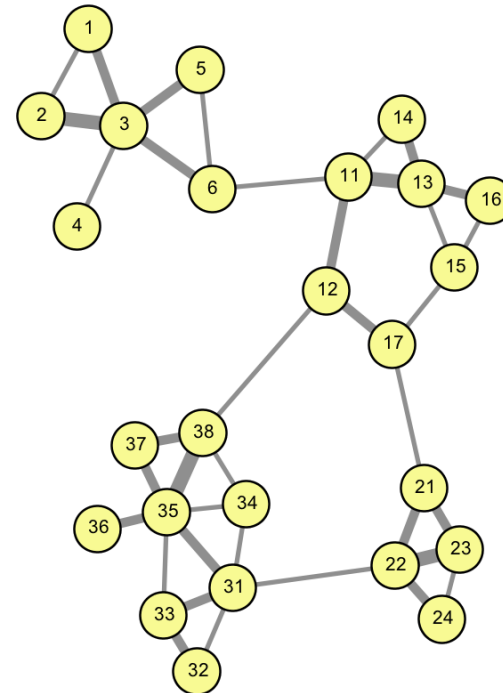
- Nodes = papers
- Links between papers sharing references, weighted by Kessler (1963)'s cosine similarity:

$$w_{ij} = \frac{|R_i \cap R_j|}{\sqrt{|R_i| |R_j|}}$$

R_i being the set of references of paper i .

Other approaches

- $w_{ij} = 1$
- $w_{ij} \times \Theta(|R_i \cap R_j| - NC^*)$
- $w_{ij} \times f(|y_i - y_j|)$
- Other variants...



STATIC CLUSTERING

DETECTING CLUSTERS

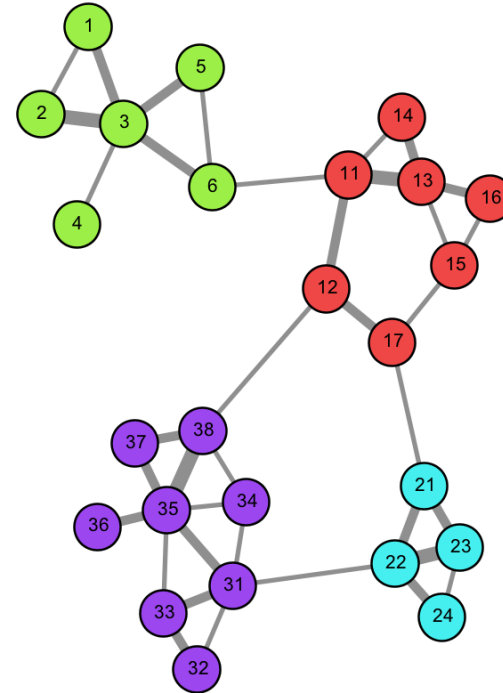
Popular approach

Community detection by modularity maximization and fast Louvain Algorithm (Blondel, 2008):

$$Q = \frac{1}{2\Omega} \sum_{ij} \left[\omega_{ij} - \frac{\omega_i \omega_j}{2\Omega} \right] \delta(c_i, c_j)$$

Other approaches

- Random walks / InfoMap (Rosvall, 2008)
- CPM (for overlapping communities)



STATIC CLUSTERING

CHARACTERIZING CLUSTERS

Their aggregate characteristics

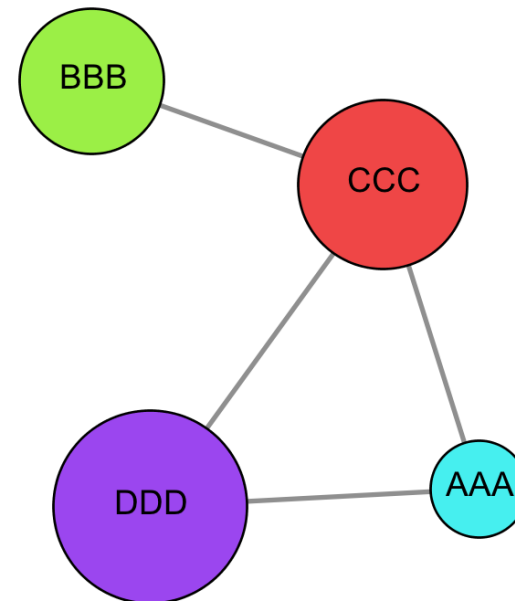
- Size of cluster \approx number of articles
- Labels are based on frequent /significant keywords
- Clusters \approx research areas / subfields, with specific shared references

Their inner structure

- Some are cohesive, with a strong core
- Some can be split in several sub-clusters

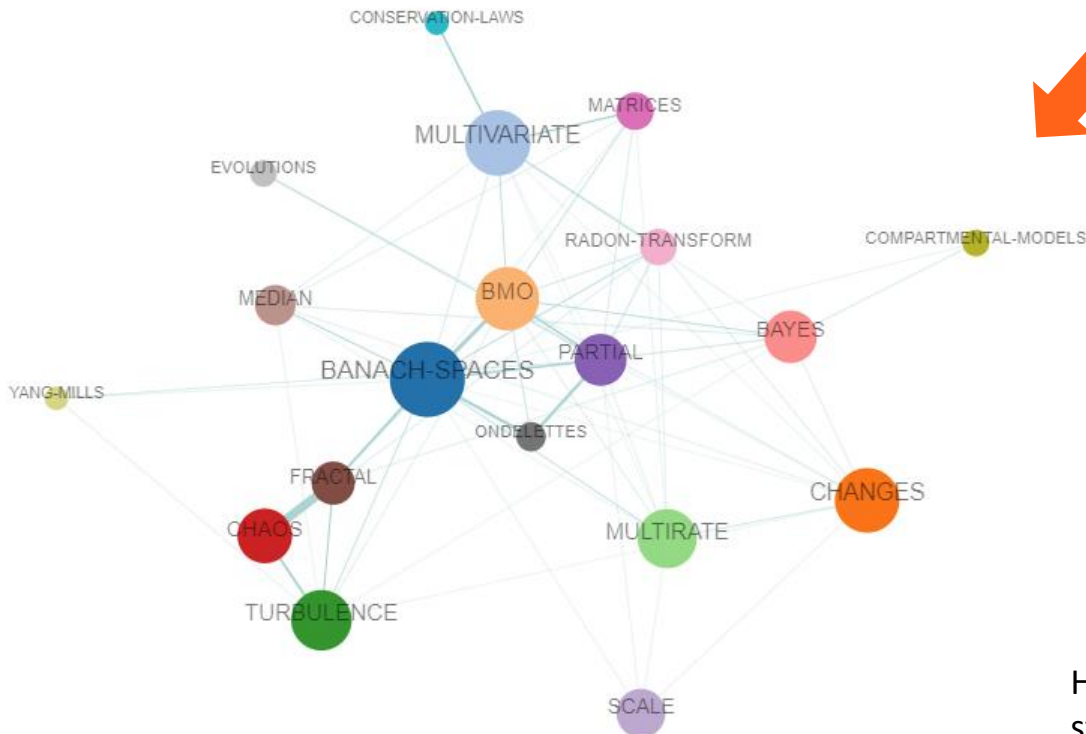
Their relationships

- What references do they share?

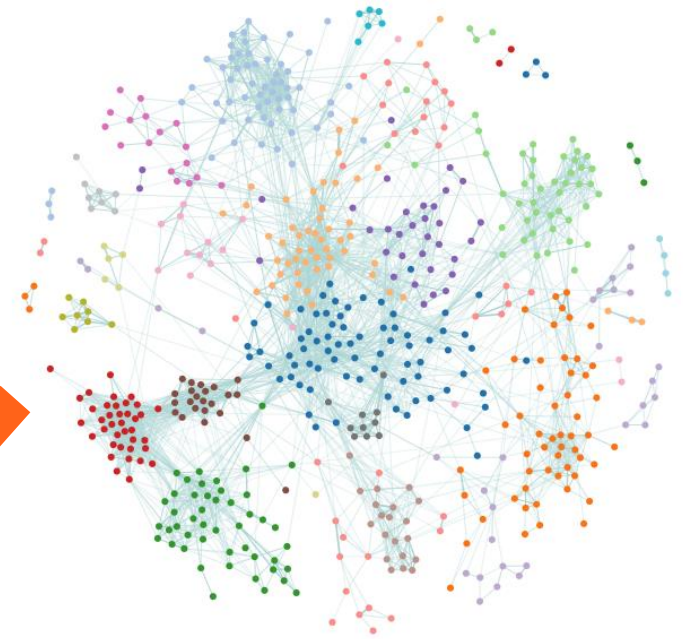


STATIC CLUSTERING

EXAMPLE: WAVELETS 80-89 (605 PUBLIS)



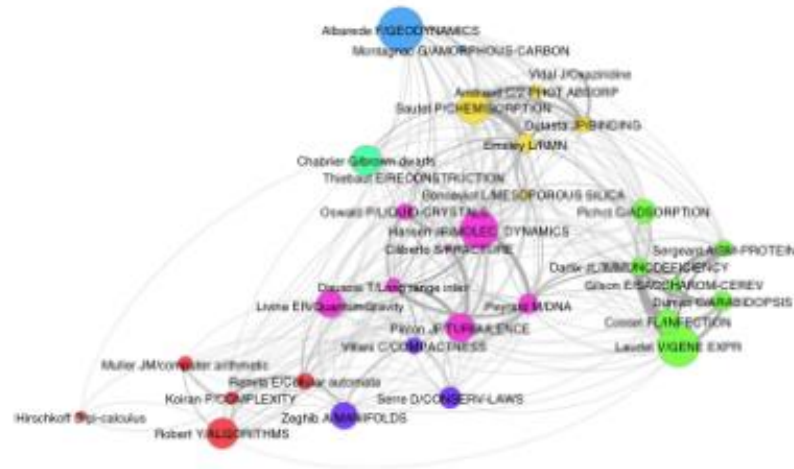
Here, labels = most significant title word



Hierarchical structure

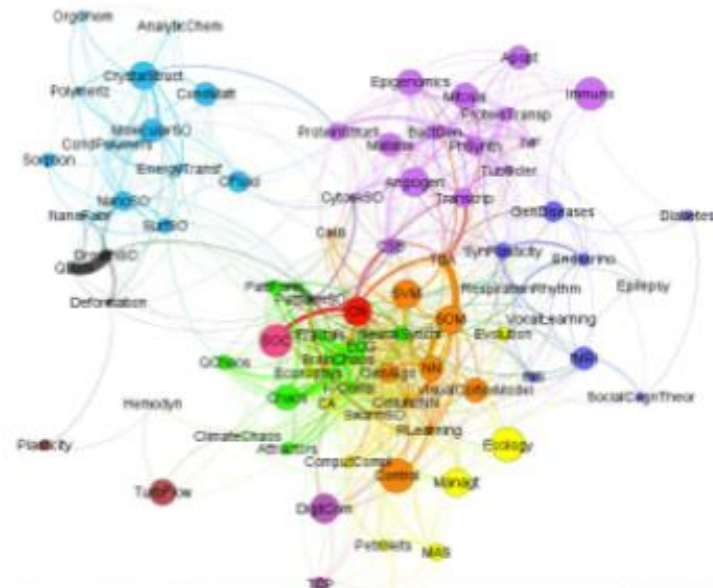


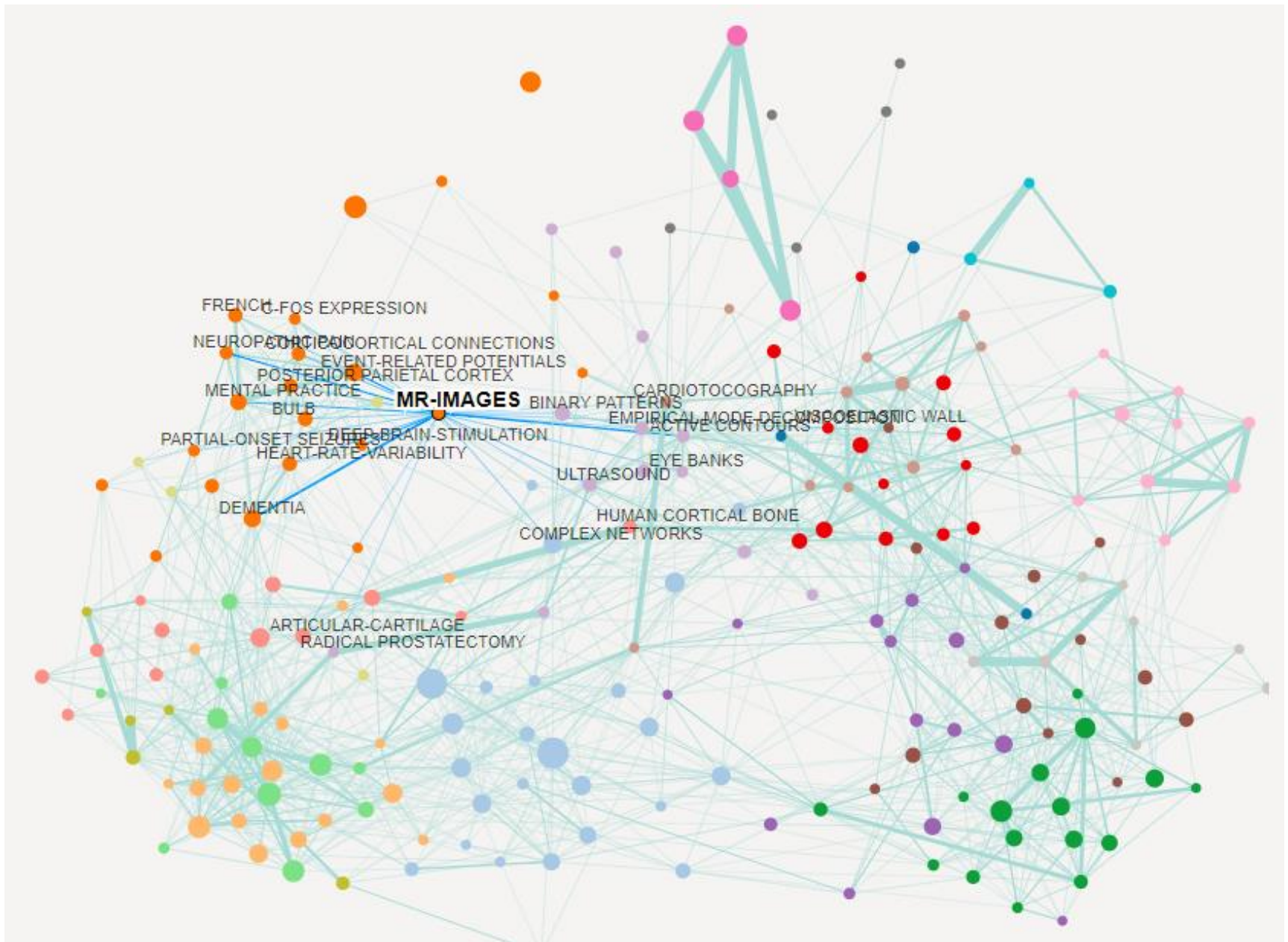
STATIC CLUSTERING



the research at ENS LYON
S Grauwin, P Jensen (2011), Mapping scientific
Institutions, *Scientometrics* 89(3)

the Complex Systems field
S Grauwin et al (2012), Complex systems
science: dreams of universality, reality of
interdisciplinarity, *JASIST* 63(7)

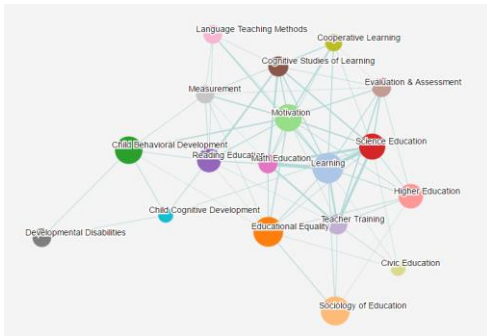




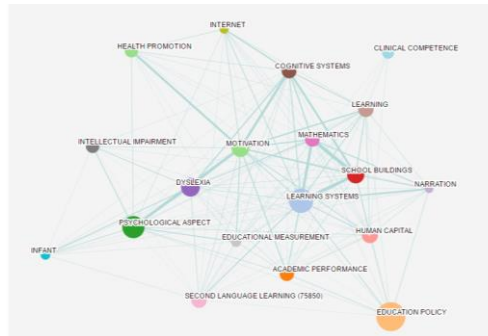
DYNAMIC CLUSTERING

COMMON APPROACH: SUCCESSIVE SNAPSHOTS

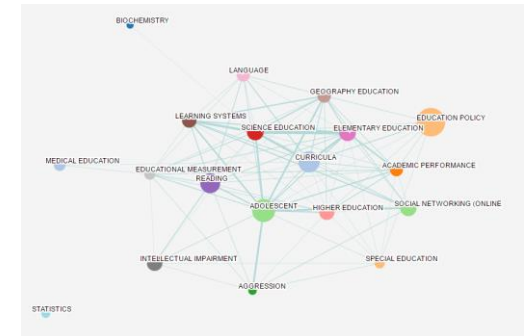
2000-2004



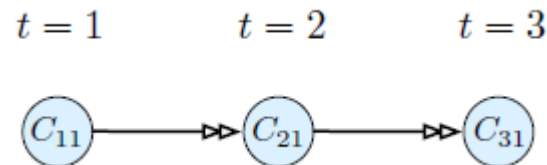
2005-2009



2010-2014



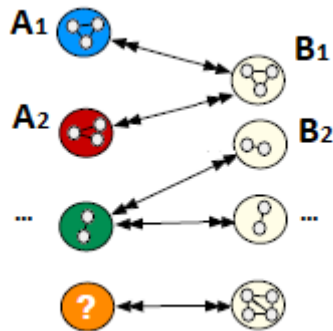
- (Independent or informed) static detections on a series of snapshots focusing on a given time window
- Dynamic community = chain of related communities observed over several time windows



DYNAMIC CLUSTERING

MATCHING COMMUNITIES FROM SUCCESSIVE SNAPSHOTS

→ Matching between communities with overlapping time windows based on a Jaccard coefficient

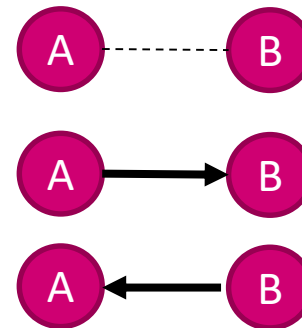


$$J_{ij} = \frac{|A_i \cap B_j|}{|A_i \cup B_j|} > \theta$$

A_i and B_j are related if $J_{ij} > \theta$

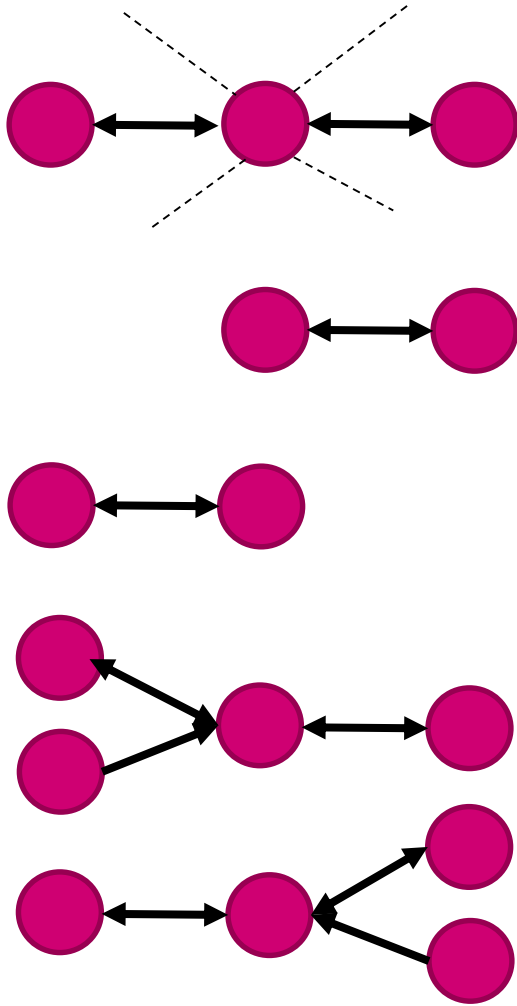
B_j is A_i 's **successor/child** if $J_{ij} = \max_k (J_{ik})$

A_i is B_j 's **predecessor/parent** if $J_{ij} = \max_k (J_{kj})$



DYNAMIC CLUSTERING

ONE EXAMPLE OF RULES SET



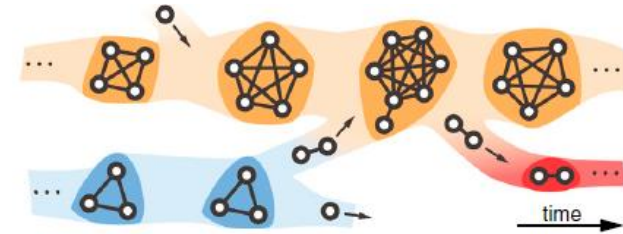
- **Continuity:** when two communities are predecessor / successor of each other.
- **Birth:** when there is no predecessor.
- **Death:** when there is no successor.
- **Merge:** when a community is the successor of two (or more) communities.
- **Split:** when a community is the predecessor of two (or more) communities

DYNAMIC CLUSTERING

SPECIFICITY OF BC NETWORK

Dynamic clustering techniques are developed for network with nodes lasting in time (e.g. human in a social network, references in a co-citation network), but publications in a BC network only exist on one time, their publication year:

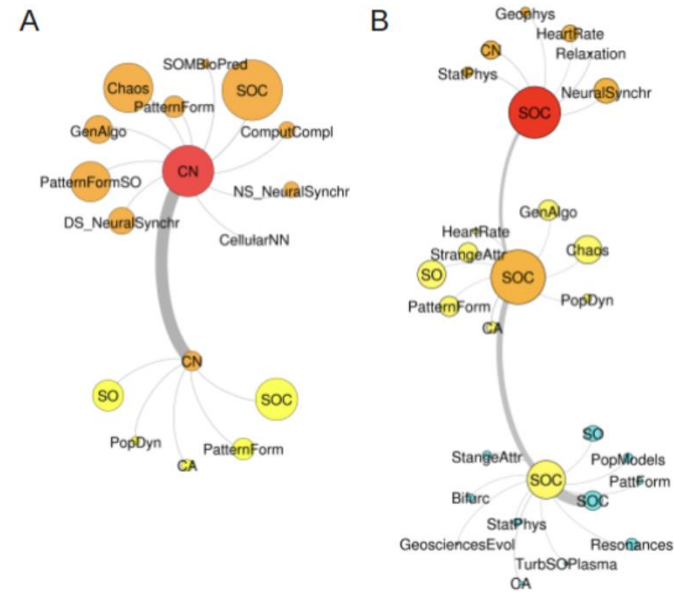
➔ Use of overlapping time window



Alternative: instead on the Jaccard index, use a similarity measure based on shared references $\Omega_{IJ} = \langle w_{ij} \rangle_{i \in I, j \in J}$

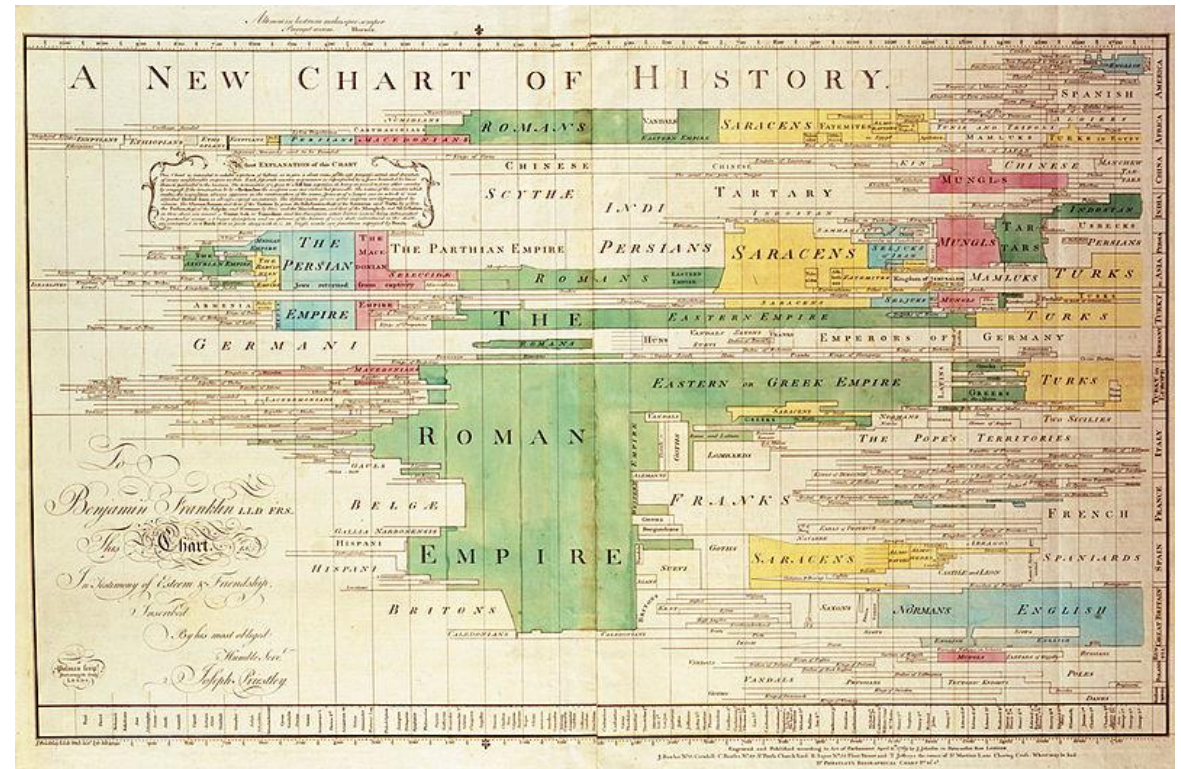
➔ No need of overlapping window

➔ Allow to compare thematic / temporal similarities



S Grauwin, PhD thesis (2011)

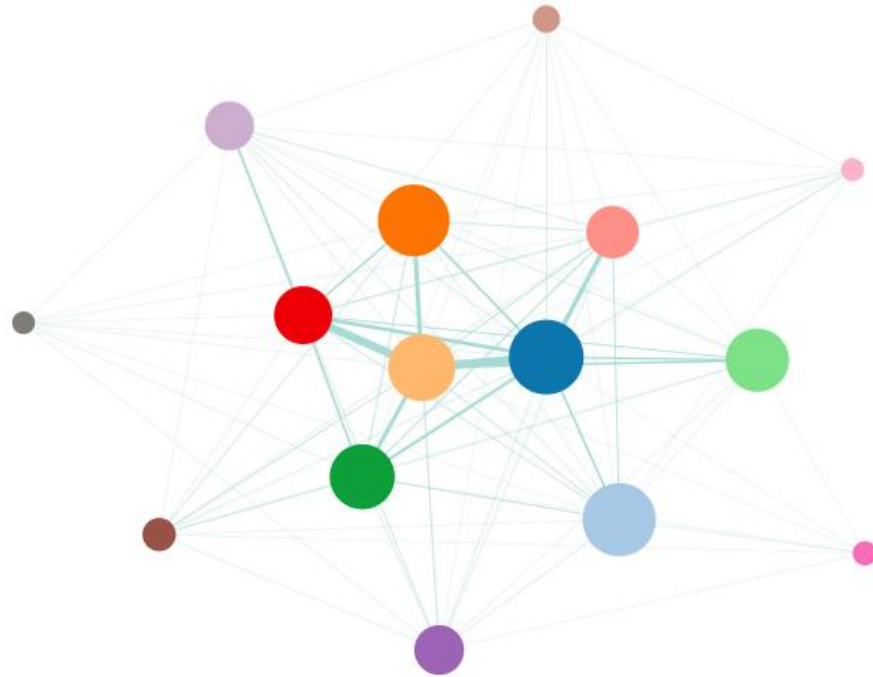
VISUALISING scientific communities



Joseph Priestley's *A New Chart of History*, 1765.

VISUALISING STATIC COMMUNITIES

NETWORK SPATIALISATION



Force-based spatialisation layout: more similar clusters are \approx closer to each other.

VISUALISING STATIC COMMUNITIES

DEDICATED INTERACTIVE TOOL: BIBLIOMAP™

DYNCOMM PROJECT
Corpus Description BC Network BC Map Notes

Thematic Map

ENS Lyon 2000-2015

Based on the references they share, thematically close publications of the studied corpus are gathered into 26 topics and 160 subtopics.

Legend
Search
Layout
Export

Circles represent research topics. Circle size is proportional to the number of publications in a topic.

Lines represent connections between topics. Line thickness is proportional to the similarity of two topics.

Colors represent inclusions within a research topic.

XY Labels correspond to:

the most frequent subjects

Select Level: Topics Subtopics

Select Layout: Static Dynamic

X

Physics, Fluids & Plasmas

The topic Physics, Fluids & Plasmas gathers 87 publications.

Compared to the average trend, the number of publication in a given year can be: ■ significantly less than expected, ■ as expected, or ■ significantly more than expected.

Jump to:

Most Frequent References

Most Frequent Keywords

| Keyword | f(%) | σ |
|--------------|-------|----|
| FLOW | 11.49 | + |
| DYNAMICS | 8.05 | = |
| LAYERS | 8.05 | ++ |
| CONTACT | 6.9 | ++ |
| FILMS | 6.9 | + |
| FLUCTUATIONS | 6.9 | + |
| MEDIA | 6.9 | + |
| SAND | 6.9 | ++ |

Necessary to explore / get a sense of the nature of the communities, go beyond a few keywords!

VISUALISING DYNAMIC COMMUNITIES

TOO MANY SNAPSHOTS?

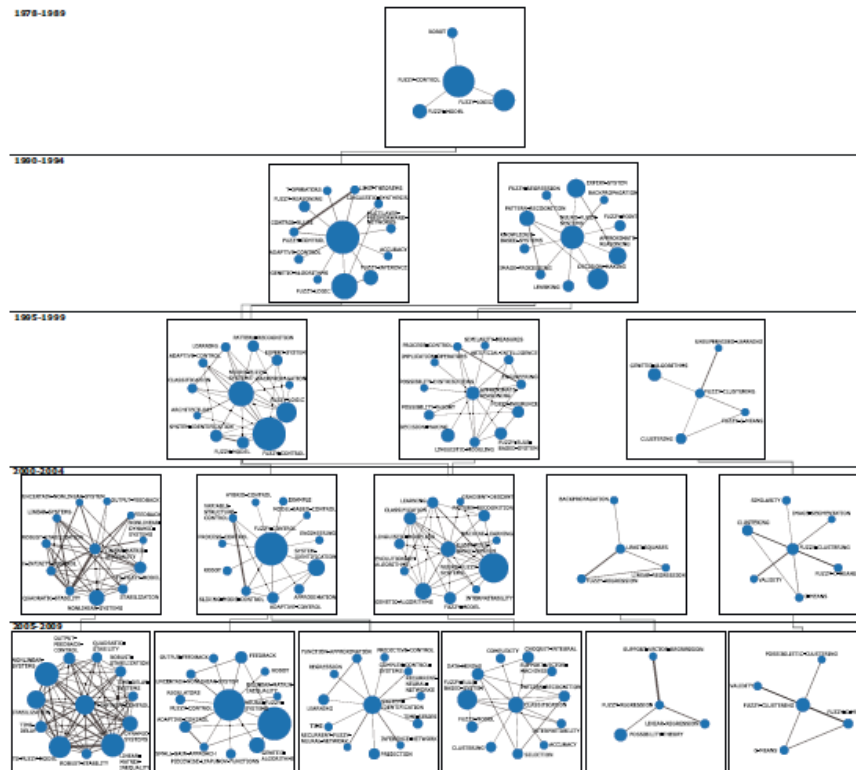
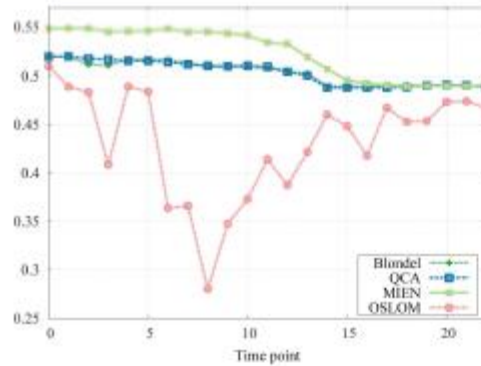


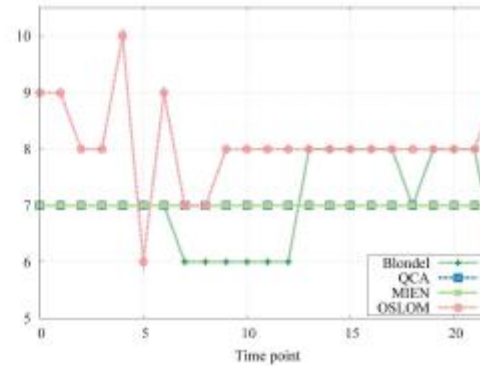
Fig. 12. The FUZZY-CONTROL thematic area (1978-2009).

VISUALISING DYNAMIC COMMUNITIES

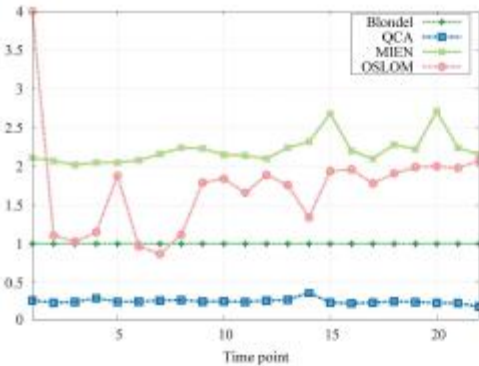
« STATISTICS » VIEW: TOO... STATISTICAL?



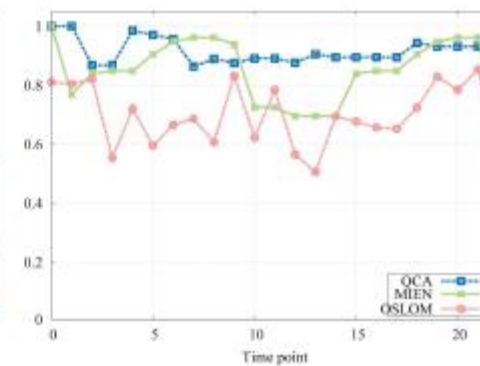
(a) Modularity



(b) Number of Communities



(c) Running Time(s)

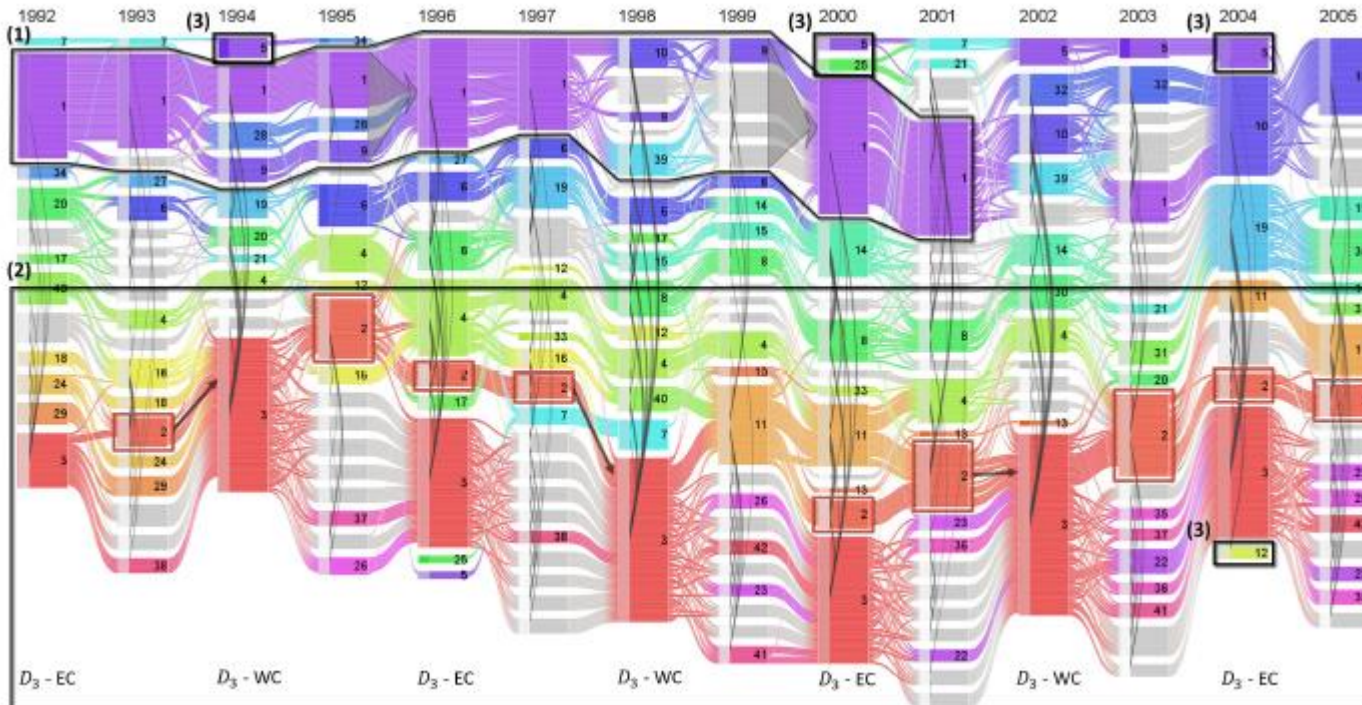


(d) NMI

Claveau & Gingras (2016), *Macrodynamics of Economics: A Bibliometric History*

VISUALISING DYNAMIC COMMUNITIES

STREAMGRAPHS

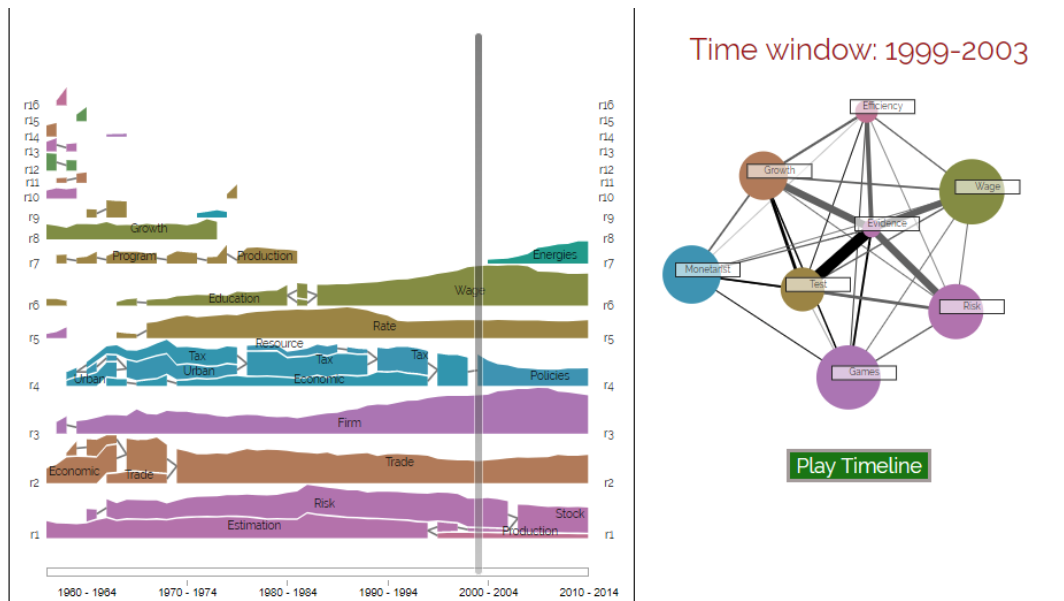


Soccer teams interactions, from C. Vehlow et al. (2014) *Visualizing the Evolution of Communities in Dynamic Graphs*.

- Initiated by Rosvall & Bergstrom (2010) for dynamic communities visualization.
- Specific challenges: ordering, coloring.

VISUALISING DYNAMIC COMMUNITIES

STREAMGRAPHS: CONCRETE REALISATIONS



Claveau & Gingras (2016), *Macrodynamics of Economics: A Bibliometric History*

- Room for improvement in terms of viz?
- Streamgraphs are well adapted for a relatively small number of communities...

REFLEXIONS UPON SOME ONGOING WORK

THE GOAL

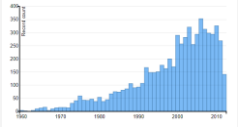
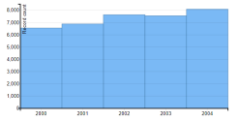
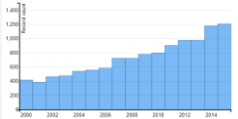
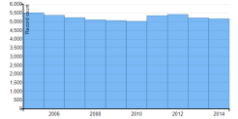
Press-button tool to create an interactive visualisation of dynamic thematic communities. We want to be able to detect and visualize their **hierarchical structures**, as well as **their internal and structural dynamics**.

Let's examine 3 sets of challenge:

- Interpretation of the communities
- Stability issues in static communities
- Evaluation of what makes a « good » history

TEST CORPUS

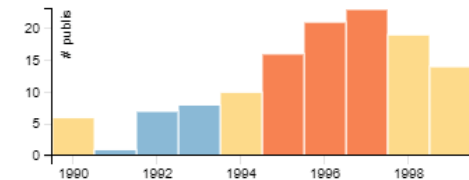
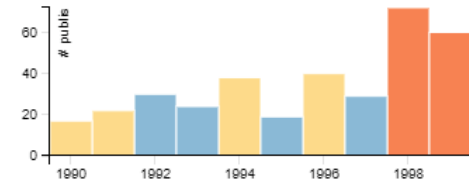
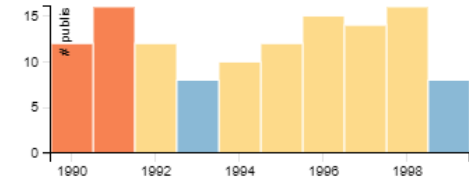
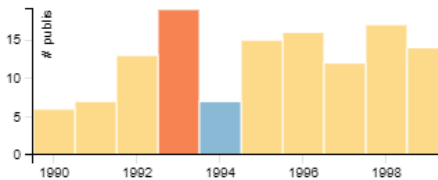
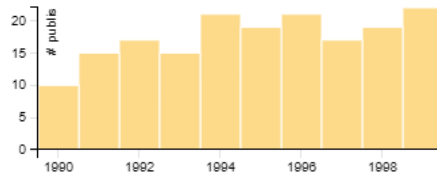
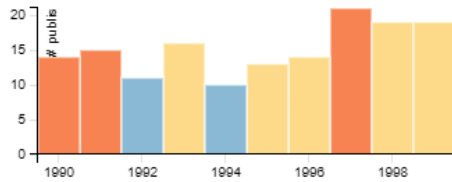
→ 4 corpus with various temporal & topical ranges

| Name | Type | Period | # publis | # publis / year |
|------------------|------------------------|-----------|----------|---|
| Wavelets | Thematic (~Specialty) | 1960-2012 | 6355 |  |
| Educmap | Thematic (~Discipline) | 2000-2004 | 36715 |  |
| ENS Lyon | Institution | 2000-2015 | 11699 |  |
| Nature & Science | Journals | 2005-2014 | 52406 |  |

TEMPORAL vs THEMATIC

WAVELETS 1990-1999

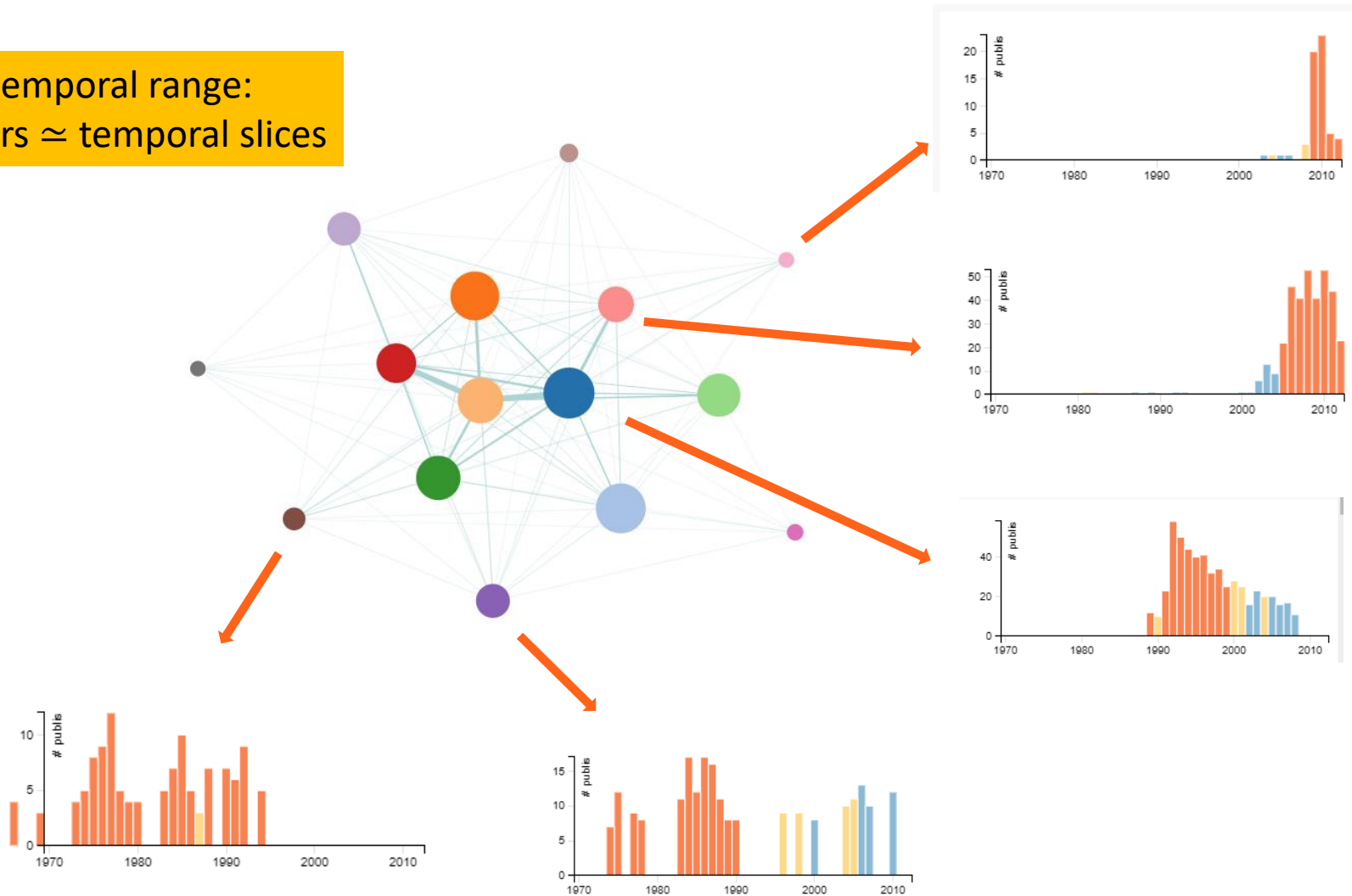
Short temporal range:
Clusters \approx thematic groups



TEMPORAL vs THEMATIC

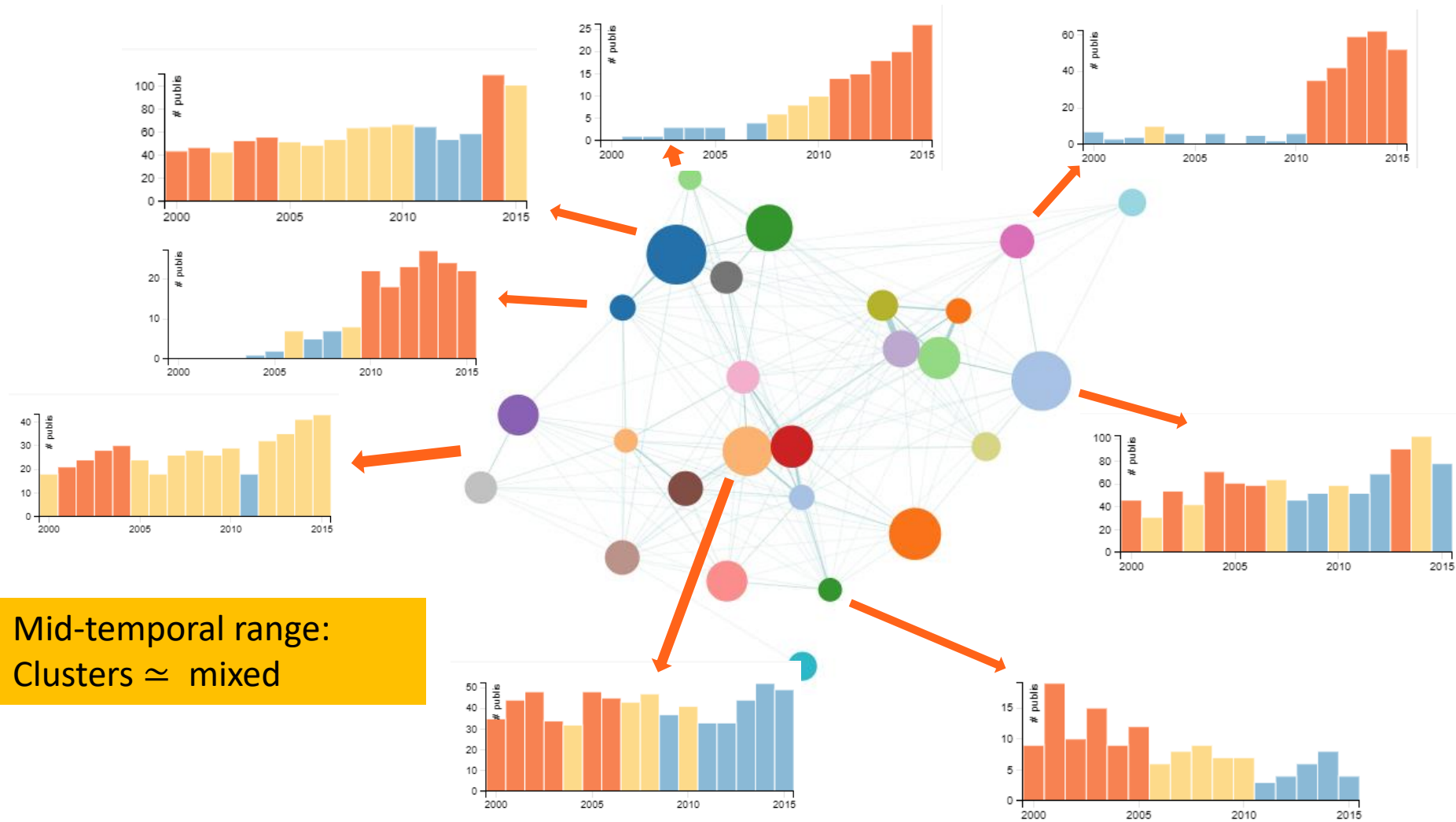
WAVELETS 1960-2012

Long temporal range:
Clusters \approx temporal slices

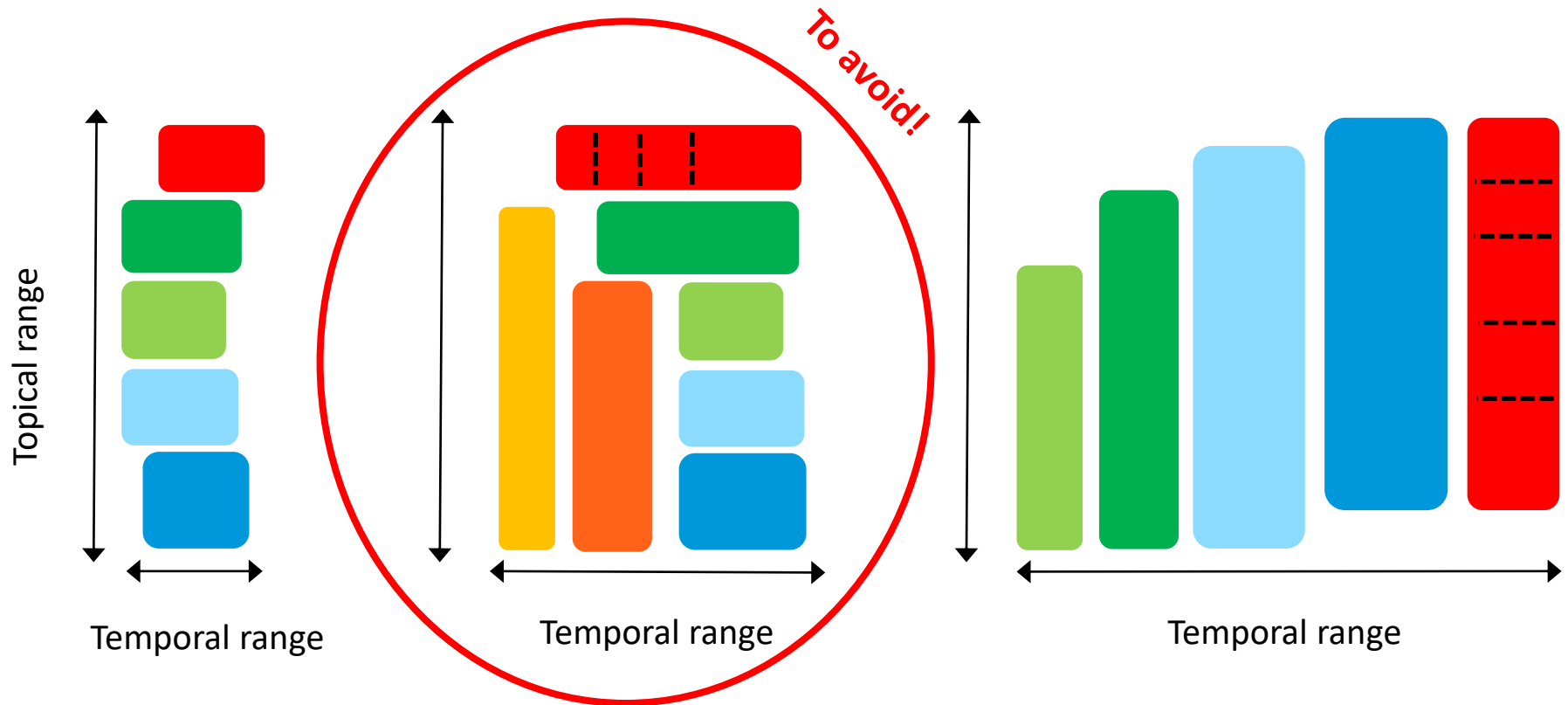


TEMPORAL vs THEMATIC

ENS-LYON 2000-2015



TEMPORAL vs THEMATIC



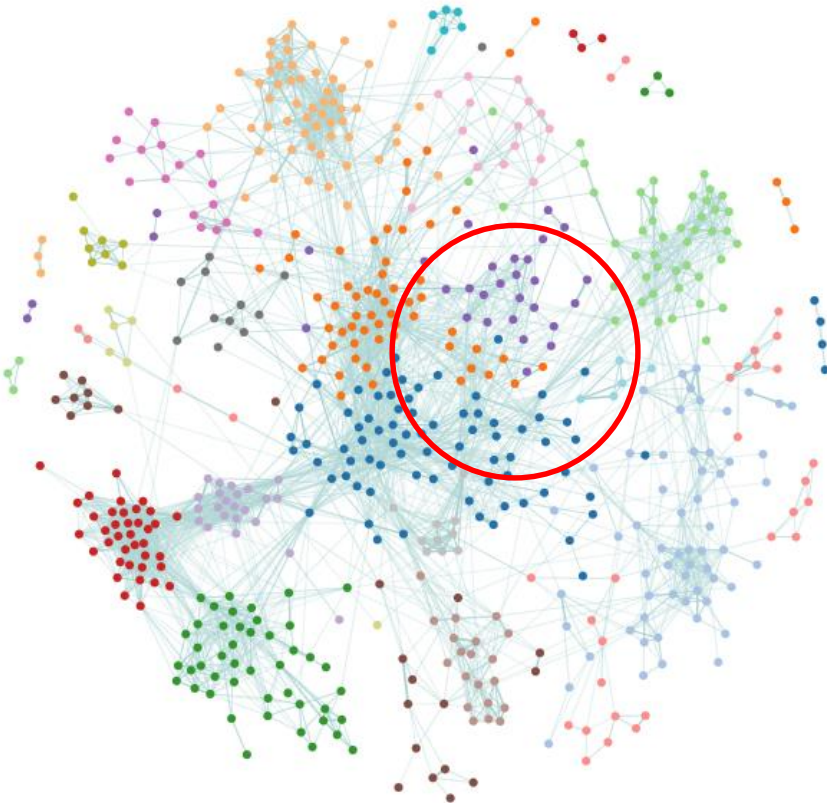
Depending on the temporal / topical ranges of the studied corpus, the algorithms may detect communities which are thematically and/or temporally different.

- ➔ Snapshots with tuned time windows
- ➔ $w_{ij} \rightarrow w_{ij} \times f(|y_i - y_j|)$ to remove links between papers published years apart.

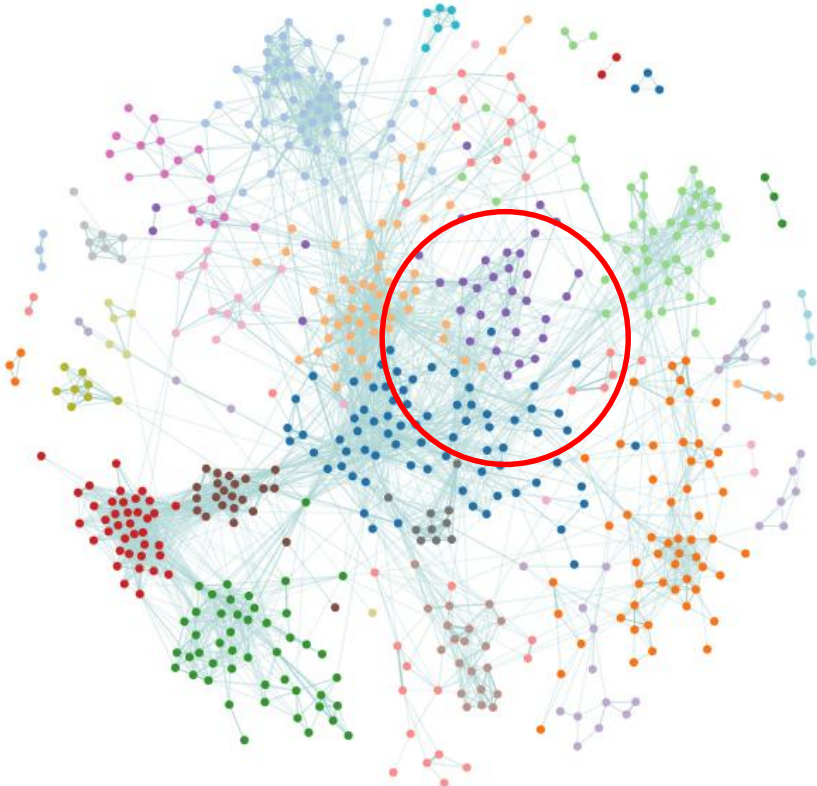
STATIC CLUSTERING - STABILITY

STABILITY PROBLEMS – WAVELETS 80-89

Run 1



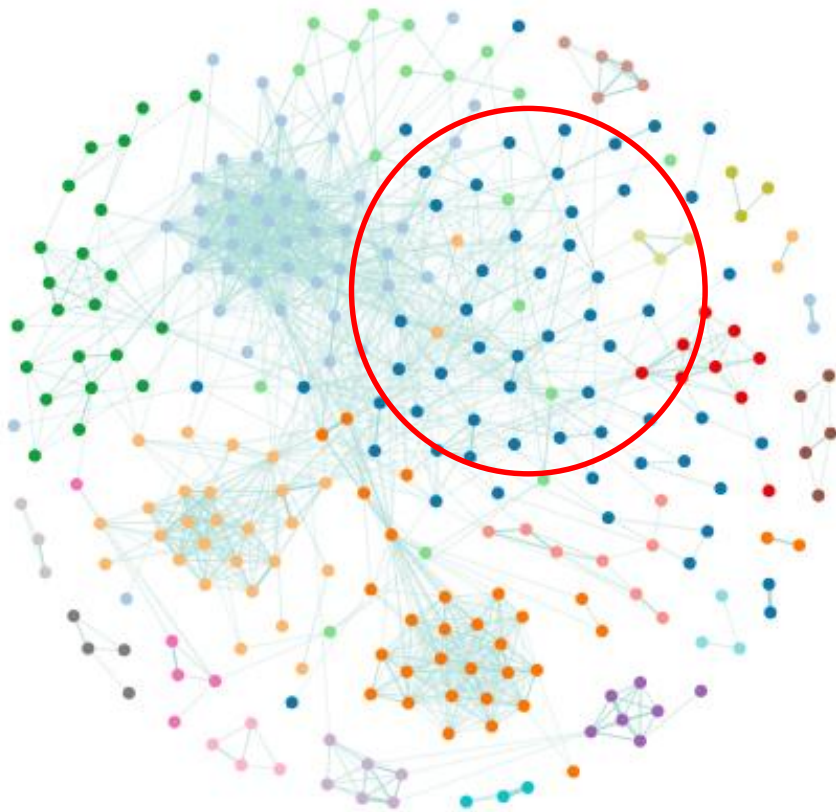
Run 2



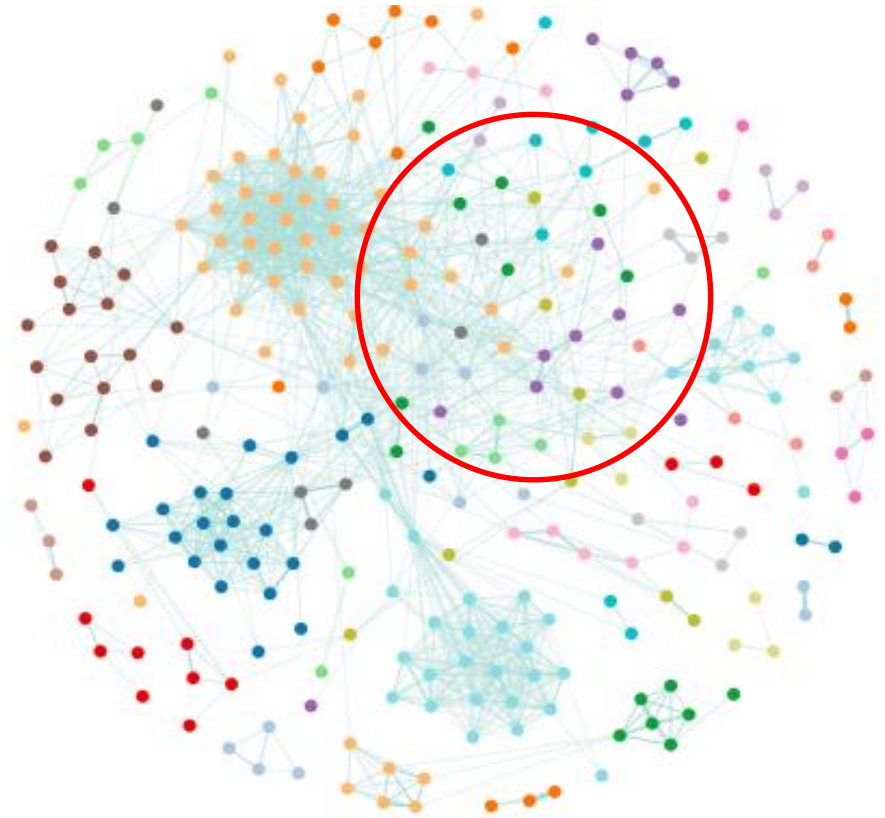
STATIC CLUSTERING - STABILITY

STABILITY PROBLEMS – WAVELETS 2010

Run 1



Run 2

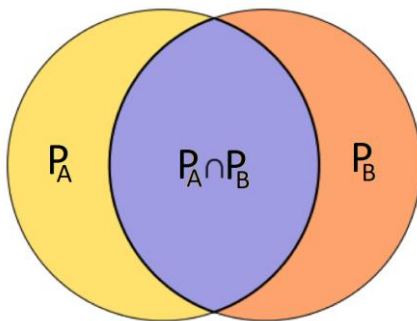


STATIC CLUSTERING - STABILITY

STABILITY MEASURES

Jaccard index

$$J(A, B) = \frac{|P_A \cap P_B|}{|P_A \cup P_B|}$$



P_X = pairs of nodes in same cluster in X.

J_w : takes into account the link weight between pairs of nodes.

(Other measures: NMI, F1score, etc)

| Corpus | Q | J | J_w |
|-------------------------------|------|------|-------|
| Wavelets 1960-2012 | 0.51 | 0.56 | 0.84 |
| Wavelets 1980-1989 | 0.82 | 0.89 | 0.98 |
| Educmap 2000-2004 | 0.53 | 0.52 | 0.87 |
| ENS Lyon 2000-2015 | 0.87 | 0.79 | 0.98 |
| Nature & Science 2005-2014 | 0.78 | 0.57 | 0.94 |

- « good » partition (high modularity Q) are more stable
- depends on thematic and temporal ranges of the corpus

STATIC CLUSTERING - STABILITY

SEARCH FOR (PERTINENT) STABLE « CORES »

Instabilities can make the matches between communities from successive snapshots less pertinent (eg, how to judge between true split or separation due to noise?). How could we reduce them?

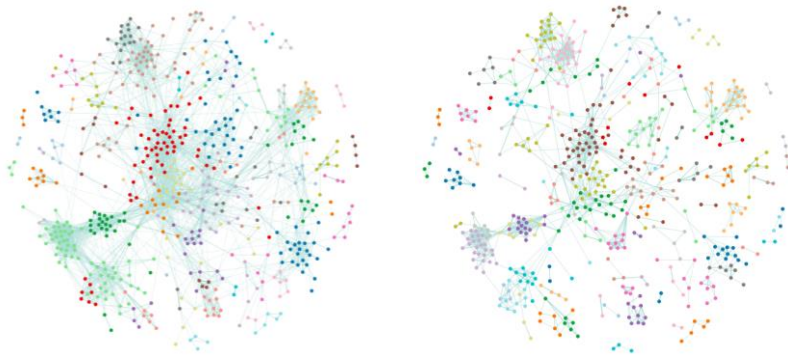
Impact of different part on the BC link definition?

$$w_{ij} = \frac{|R_i \cap R_j|}{\sqrt{|R_i| |R_j|}}$$

- $w_{ij} \rightarrow w_{ij} \times \Theta(w_{ij} - w^*)$ weight larger than threshold
- $w_{ij} \rightarrow w_{ij} \times \Theta(|R_i \cap R_j| - NC^*)$ # of shared refs larger than threshold
- $w_{ij} \rightarrow |R_i \cap R_j \cap R_{TU > TU^*}| / \sqrt{|R_i| |R_j|}$ only count shared ref used more than threshold

STATIC CLUSTERING - STABILITY

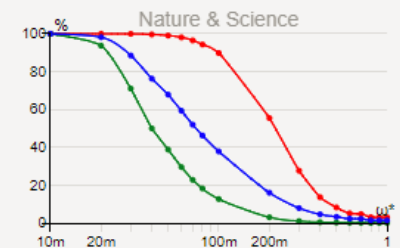
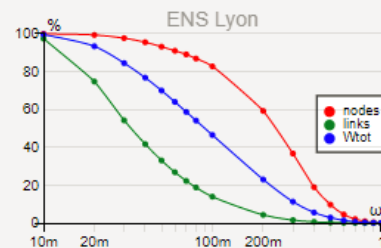
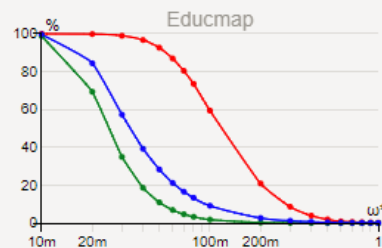
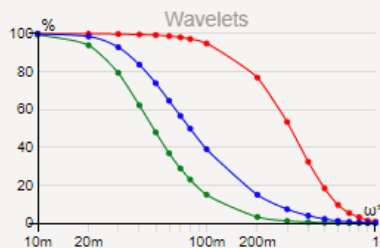
SEARCH FOR (PERTINENT) STABLE « CORES »



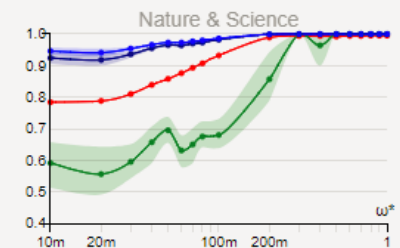
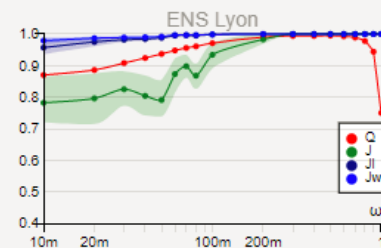
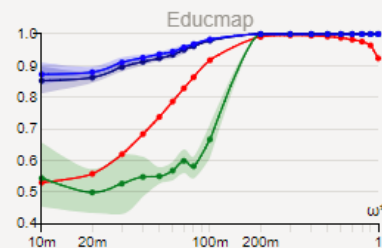
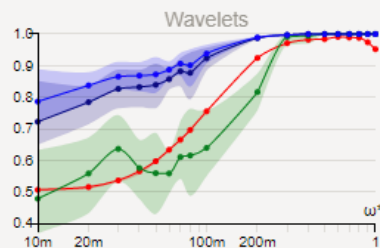
Filtering on link weight:

- Remove links then nodes
- Improvement on Q or J only when network is well truncated

What remains in the BC network with $\omega_{ij} \geq \omega^*$?



Stability of the BC partitions with $\omega_{ij} \geq \omega^*$?



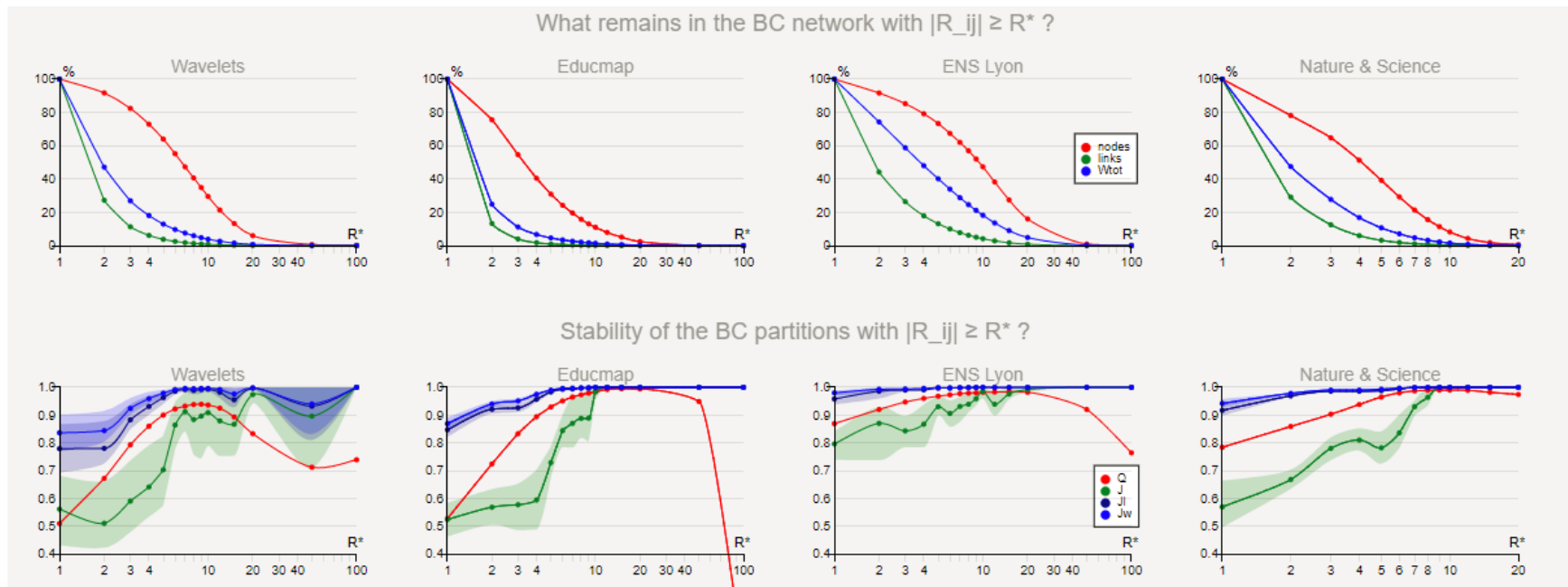
STATIC CLUSTERING - STABILITY

SEARCH FOR (PERTINENT) STABLE « CORES »



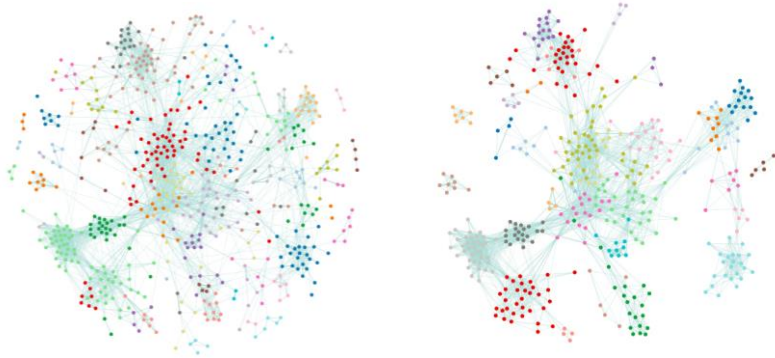
Filtering on # of shared reference:

- Remove links then nodes
- Improvement on Q or J costly
- Use $R^*=1$ for pluri-disciplinary corpus, $R^*=2$ or 3 for mono-disciplinary corpus



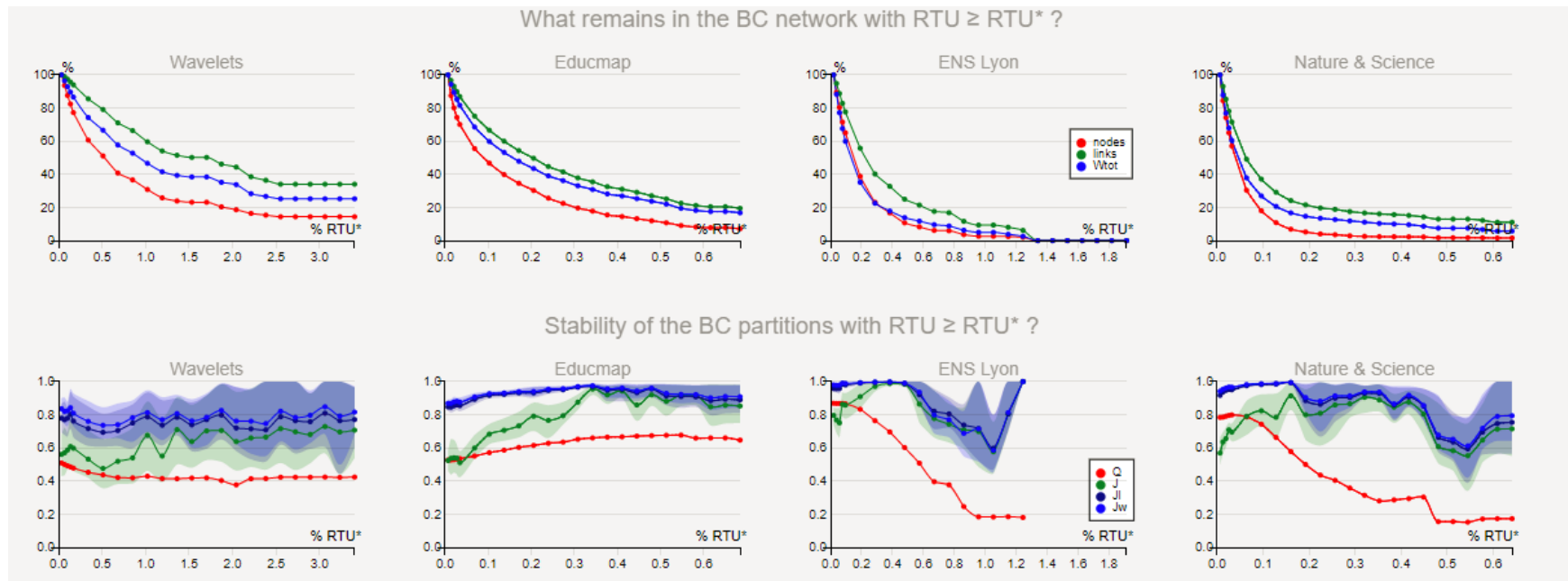
STATIC CLUSTERING - STABILITY

SEARCH FOR (PERTINENT) STABLE « CORES »



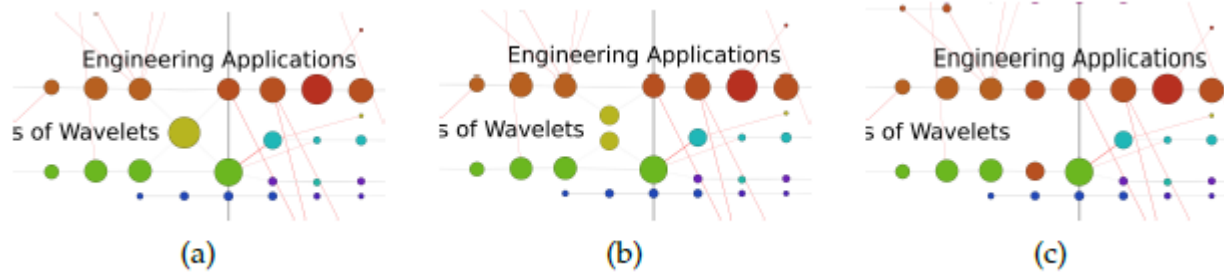
Filtering Ref Times Used:

- Remove nodes then links
 - Improvement on stability, costly
 - Corpus more thematically focused: slower decrease (plateau?)
- ➔ To adapt to detect communities' cores?



HOW TO COMPARE / EVALUATE HISTORIES?

SEARCH FOR AN OBJECTIVE FUNCTION



We want to avoid « unwanted » splits / merges

→ On which criterion?

HOW TO COMPARE / EVALUATE HISTORIES?

SEARCH FOR AN OBJECTIVE FUNCTION

For each snapshot t (set of publis P_t), we maximize in an independant manner the modularity

$$Q_t = \frac{1}{2\Omega_t} \sum_{i \in P_t, j \in P_t} \left[\omega_{ij} - \frac{\omega_i^t \omega_j^t}{2\Omega_t} \right] \delta(c_i, c_j)$$

In a « good » history, the partition at each time step should be chosen among those with quasi-max value of Q_t to best match the other steps.

Specificity of BC: links weight ω_{ij} can assess the thematic similarity as well as the temporal distance.

Dubbing h_i the reconstructed dynamic communities , one can evaluate the history with

$$Q_{history} = \frac{1}{2\Omega} \sum_{i \in P_t, j \in P_{t'}, t \neq t'} \left[\omega_{ij} - \frac{\omega_i \omega_j}{2\Omega} \right] \delta(h_i, h_j)$$

