EXPERTISECENTRUM O&O MONITORING (ECOOM)
SERVICE ACTIVITIES AND RESEARCH OUTPUT

WOLFGANG GLÄNZEL

ECOOM & Dept MSI, KU Leuven, Belgium

ECOOM

About ECOOM

The Expertisecentrum Onderzoek en Ontwikkelingsmonitoring (ECOOM) is an interuniversity consortium with participation of all Flemish universities (KU Leuven, UGent, UAntwerpen, VUB and UHasselt).

Its mission is to develop a consistent system of Research, Development & Innovation Indicators for the Flemish government.

This indicator system has to assist the Flemish government in mapping and monitoring the RD&I efforts in the Flemish region.

URL: http://www.ecoom.be

The ECOOM members

KU Leuven (Co-ordinator): Bibliometrics, Technometrics and Innovation

UAntwerpen: Flemish academic bibliographic database for social sciences and humanities (VABB-SHW)

UGent: Production of doctorate degrees, academic careers and mobility

VUB: Research in the arts, research excellence

1. Services for the Flemish government

- Bibliometric support in framework of university funding
- Coordination and edition of the biennial Flemish Indicator Book R&D and Innovation
- Bibliometric profiling and support for FWO
- Bibliometric-technometric studies of Strategic Research Centres in Flanders
- Domain studies
- Ad hoc tasks

2. Supporting activities

- Creation and maintenance of an appropriate IT platform
- Integration of multidisciplinary bibliographic databases
  - TR WoS (SCIE, SSCI, AHCI, Proceedings, JCR metrics)
  - Elsevier SCOPUS
- Integration of supplementary data sources

3. Research activities ...

- Methodological/theoretical
- Applied
- Policy relevant

... and fields

- Information science
- Computer science
- Economics
- Science policy

4. Research topics

1. Development and improvement of bibliometric indicators for the evaluation of research
2. Research performance at the institutional, regional, national and supranational level
3. Dynamic and structural studies of science
4. Exploration of bibliographic databases for bibliometric use and improvement of subject delineation and classification
5. Bibliometrics in the social sciences and humanities

Development and improvement of bibliometric indicators for the evaluation of research

- A priori and a posteriori normalisation of citation
  - Transformation of impact factor scores
  - *Characteristic Scores and Scales* in the evaluation and ranking of scientific journals
  - A priori normalisation of citation indicators
- Scientometrics analysis of scholarly communication behaviour (e.g., author self-citations, delayed recognition)
- Properties and application of Hirsch-type measures

A priori and a posteriori normalisation of citation

Publication-activity and citation-impact statistics are influenced by a various factors (subject, age, time, status, communication form, etc.).

Two paradigmatic approaches are under discussion.

- *A posteriori normalisation*: mathematical manipulation of (standard) indicators
- *A priori normalisation*: fractional counting prior to indicator calculation

The subject bias is one of the most common issue in evaluative bibliometrics (see example on the next slide).

A priori and a posteriori normalisation of citation

The "Aggregate Impact Factor" of selected disciplines (both JCR Editions 2009)
Source: Thomson Reuters – Web of Knowledge

Subject Category	AIF
cell biology	5.696
neurosciences	3.869
psychiatry	3.151
physics, particles & fields	3.165
pharmacology & pharmacy	2.934
chemistry, analytical	2.608
psychology, developmental	2.341
soil science	1.560
geography	1.465
information science & library science	1.300
engineering, civil	1.096
economics	1.059
sociology	0.873
political science	0.742
mathematics	0.695

Transformation of impact factor scores

Let X_i be citation-impact observations (e.g., Impact Factors) with rank R_i, where $i \leq n$. The van der Waerden scores are standard normally distributed.

$$X_i \rightarrow \Phi^{-1}\left(\frac{R_i}{n+1}\right),$$

Where Φ^{-1} denotes the normal quantile function. Transforming these scores with an exponential function results in scores with lognormal distribution for some base line value $a > 0$.

$$X_i \rightarrow a^{\Phi^{-1}}\left(\frac{R_i}{n+1}\right)$$

The second-stage score is defined by attributing score '1' to the top 10%.

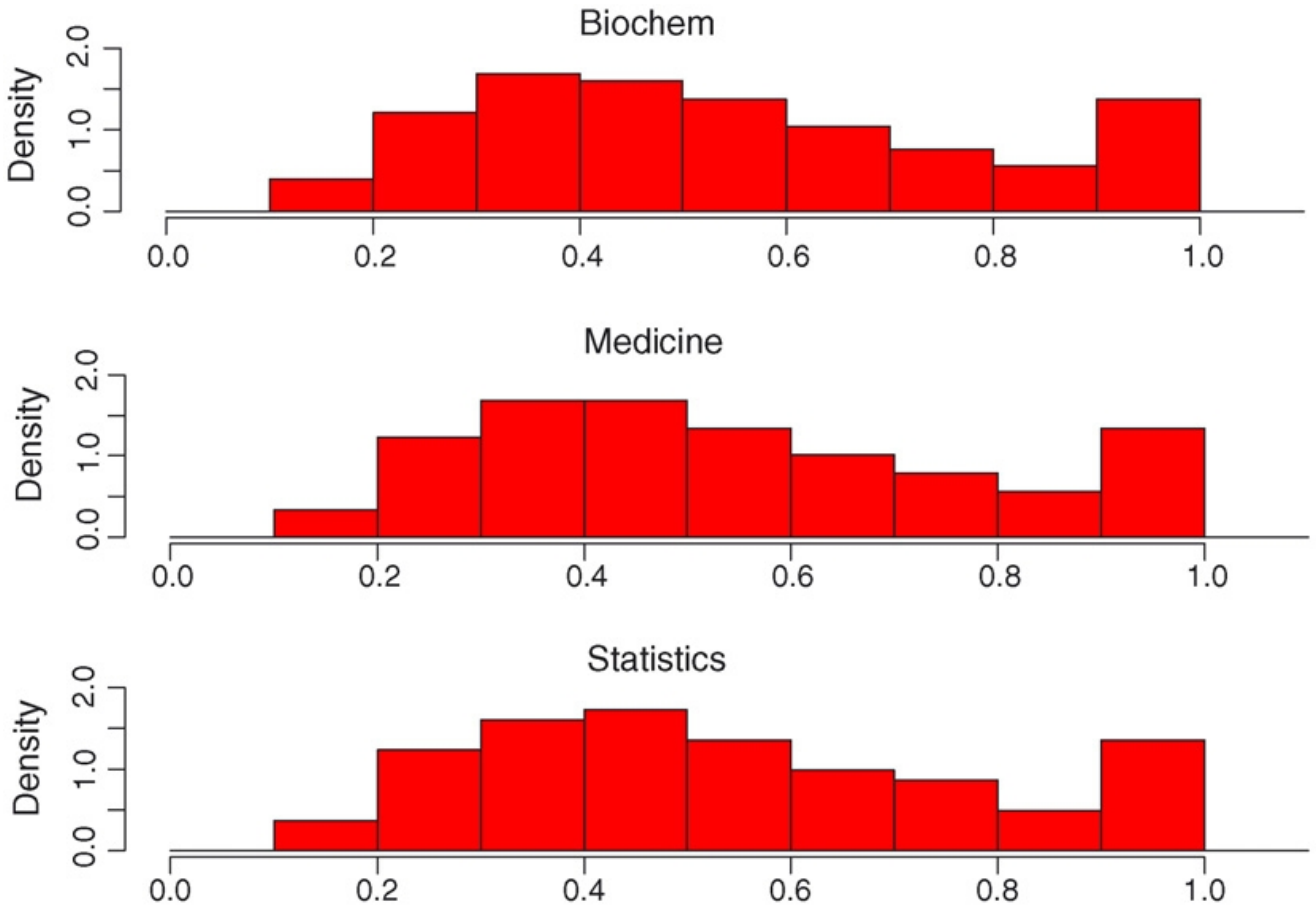
$$X_i \rightarrow a^{\Phi^{-1}(R_i/(n+1))-\Phi^{-1}(0.9)_-},$$

where $u_- = u$, if $u < 0$ and $u_- = 0$, if $u > 0$.

Beirlant et al., *Journal of Informetrics*, 2007

Transformation of impact factor scores

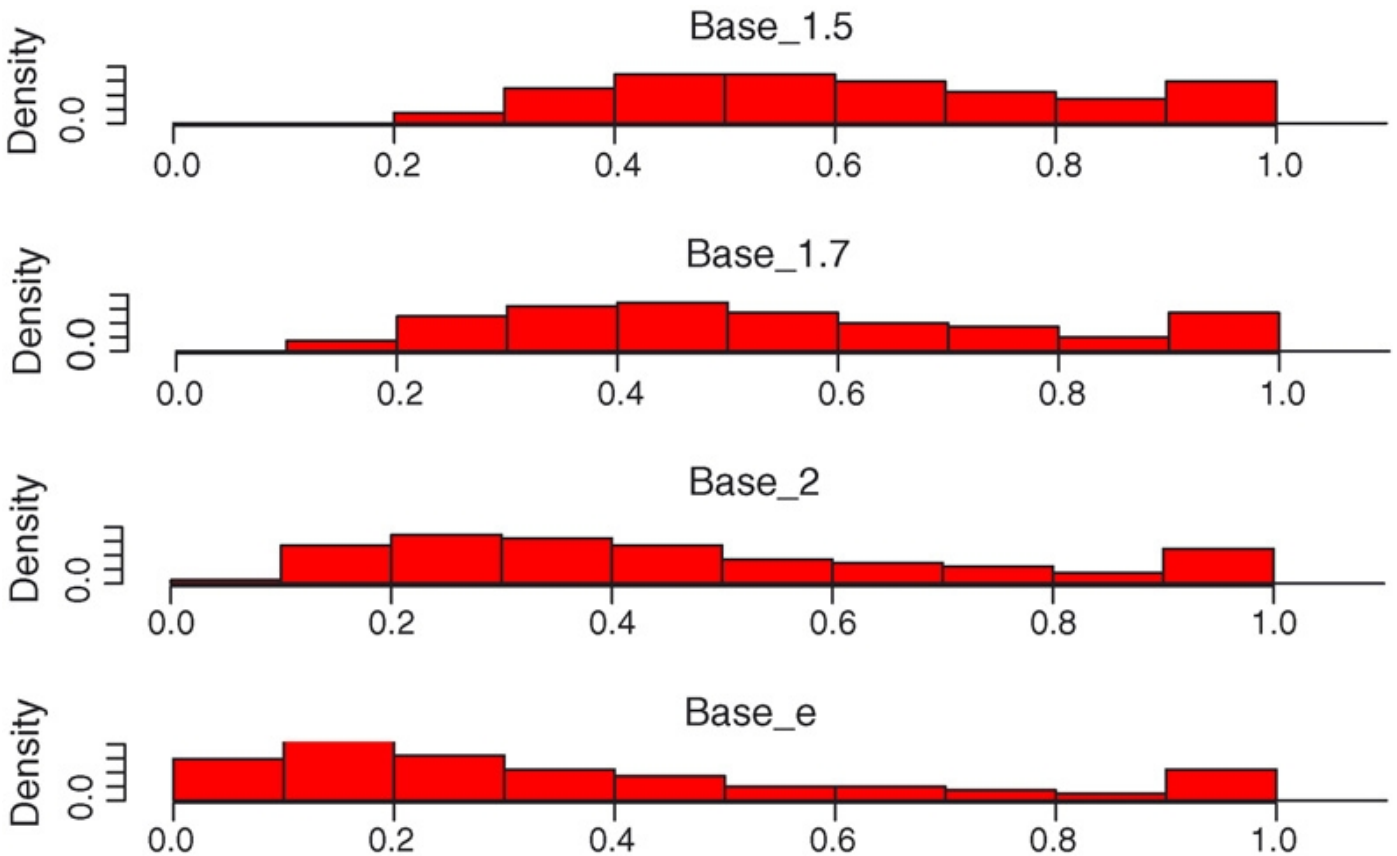
Distributions of second-stage scores for three JCR subject categories



Source: Beirlant et al., *Journal of Informetrics* (2007)

Transformation of impact factor scores

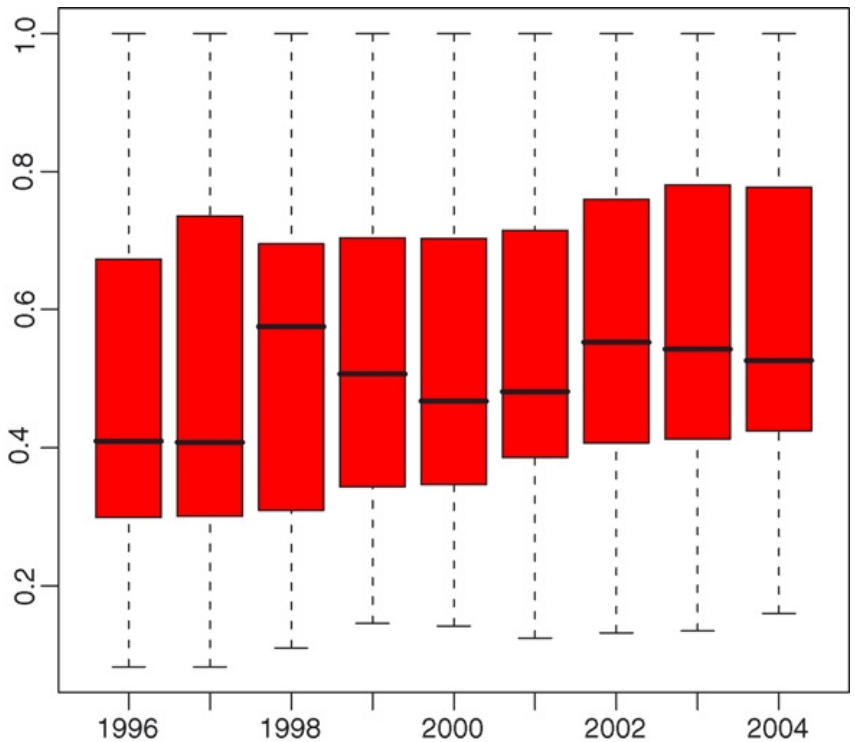
Distributions of second-stage scores for different values of *a*



Source: Beirlant et al., *Journal of Informetrics* (2007)

Transformation of impact factor scores

Boxplots of the transformed scores for the papers of Belgian mathematicians for the period 1996–2004



Source: Beirlant et al., *Journal of Informetrics* (2007)

Definition (Characteristic Scores and Scales)

Let X_i^* be n ranked observations, $\beta_0 = 0$ and $v_0 = n$. β_1 is defined as the mean

$$\beta_1 = \sum_{i=1}^{n} \frac{X_i^*}{v_0}.$$

The value v_1 is defined by $\quad X_{v_1}^* \geq \beta_1$ and $X_{v_1+1}^* < \beta_1$.

This procedure is repeated recurrently.

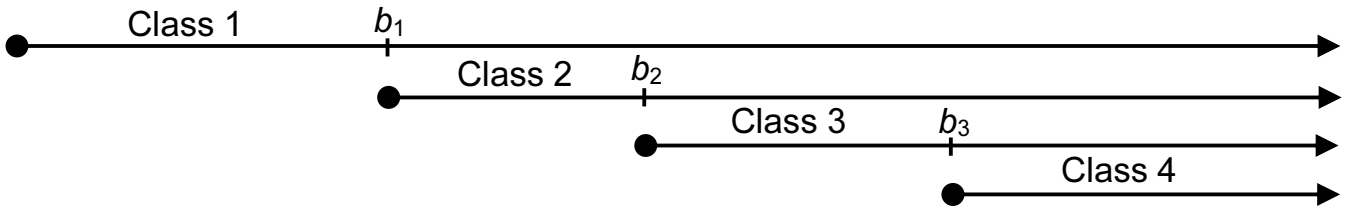
$$\beta_k = \sum_{i=1}^{v_{k-1}} \frac{X_i^*}{v_{k-1}}$$

and v_k is chosen so that $\quad X_{v_k}^* \geq \beta_k$ and $X_{v_k+1}^* < \beta_k\,, \quad k \geq 2$.

The properties $\beta_0 \leq \beta_1 \leq \ldots$ and $v_0 \geq v_1 \geq \ldots$ are obvious from the definition.

Glänzel & Schubert, *Journal of Information Science*, 1988

Visualisation of characteristic scores and scales for four classes



The transformation suggested by Schubert et al. (1989) is applied, however, without shifting the variable by the β_1.

$$u^* = \frac{x}{\beta_2 - \beta_1},$$

where x represents the actual citation statistic.

- Schubert et al., *Scientometrics*, 1989
- Glänzel, *Journal of Informetrics*, 2007
- Glänzel, *Journal of Information Science*, 2011

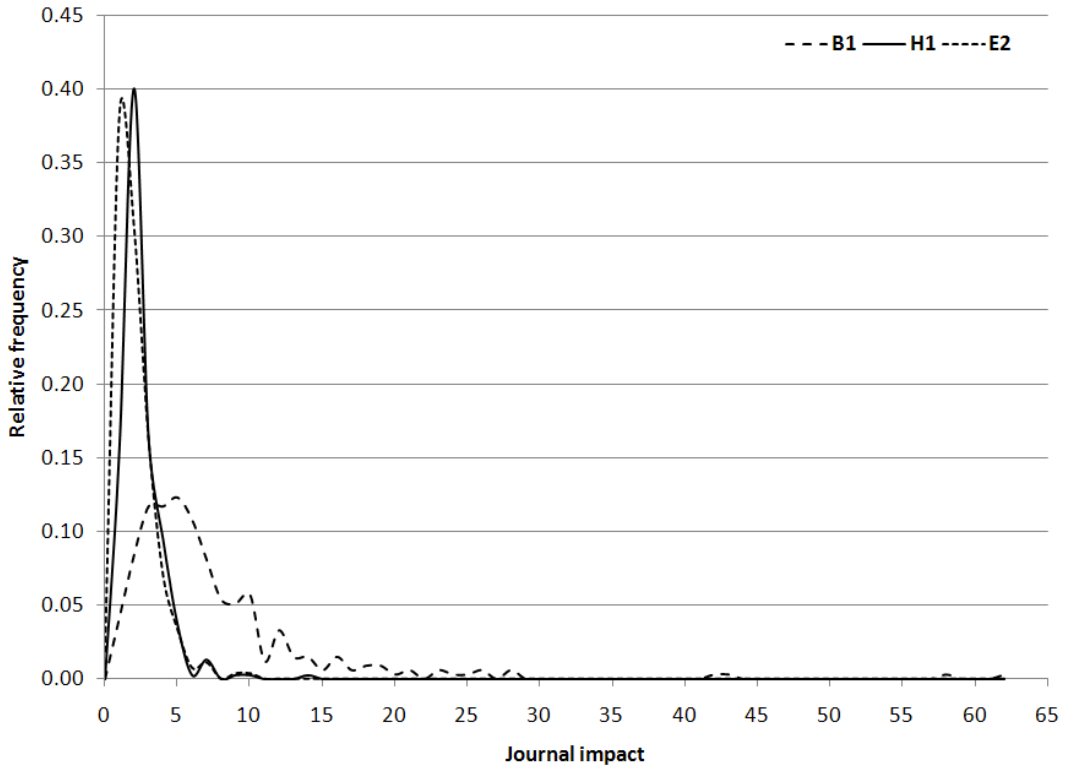
CSS values (β_k) according to the distribution of journal impact measures and their normalised versions (β_k^*) for three subfields (citation window: 2006–2008)

k	β_k			β_k^*		
	B1	**H1**	**E2**	**B1**	**H1**	**E2**
0	0.00	0.00	0.00	0.00	0.00	0.00
1	6.90	1.82	1.59	1.07	1.41	1.19
2	13.33	3.11	2.92	2.07	2.41	2.19
3	22.69	4.47	4.36	3.53	3.46	3.27

Legend B1: biochemistry/biophysics/molecular biology, H1: applied mathematics, E2: electrical & electronic engineering

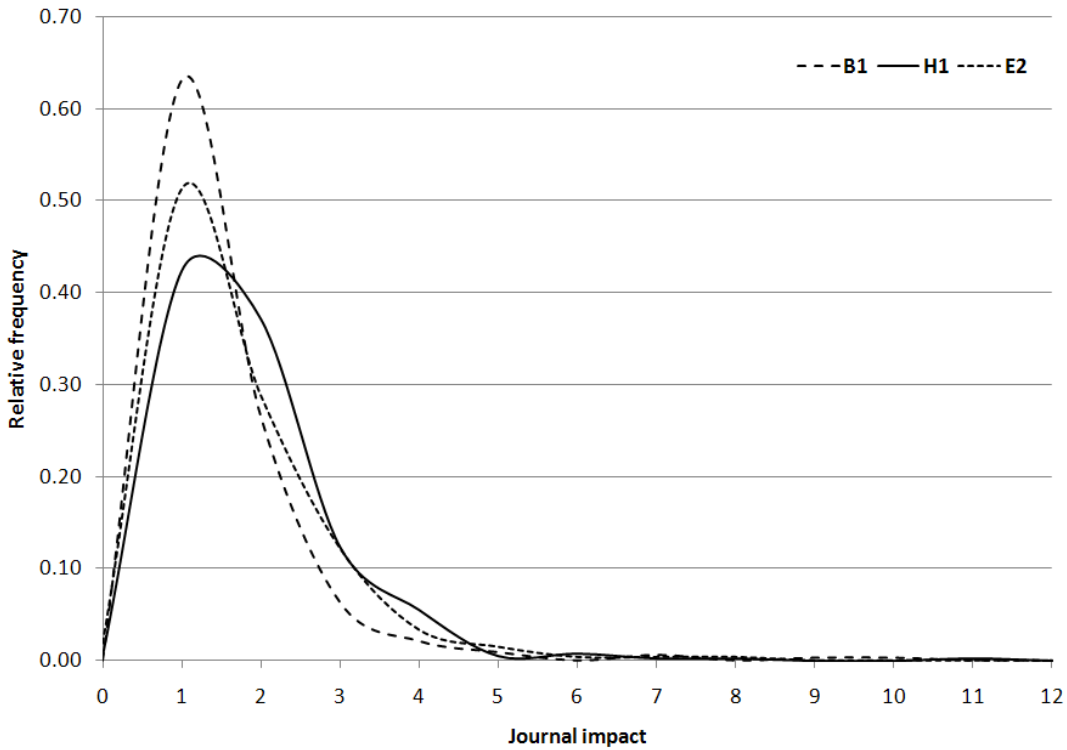
Source: Glänzel, *Journal of Information Science* (2011)

Distribution of mean citation rate over journals (citation window 2006–2008)



Source: Glänzel, *Journal of Information Science* (2011)

Distribution of mean citation rate over journals after the u^* normalisation



Source: Glänzel, *Journal of Information Science* (2011)

A priori normalisation:

- Fractional citation counts use references from indexed source items at the level of individual papers.
- Fractional 'citation value' amounts to $1/k$ if paper A is published in year y and paper B has k references to papers indexed in the year y
- The case $k = 0$ cannot occur once A is cited by B.
- *"Consistency" requirement*: The grand total over the the impact measure of all papers equals the impact measure of the total.

Glänzel et al., *Scientometrics*, 2011

Citation indicators of science fields based on integer and fractional counts

Field	MOCR	$MOCR\|_{+}$	$MOCR\|_{F}$	$MOCR\|_{F+}$	f_o
A	3.18	4.24	1.26	1.68	25.0%
Z	4.60	5.84	1.51	1.92	21.3%
B	7.93	8.91	2.05	2.31	11.0%
R	5.55	6.66	1.61	1.93	16.7%
I	7.18	8.93	2.03	2.52	19.6%
M	4.28	5.69	1.47	1.95	24.9%
N	5.68	6.88	1.74	2.10	17.5%
C	4.35	5.82	1.42	1.89	25.3%
P	3.90	5.30	1.47	2.00	26.4%
G	4.57	6.09	1.53	2.04	25.0%
E	1.71	3.40	0.81	1.60	49.7%
H	1.85	3.21	0.98	1.69	42.4%

Source: Glänzel et al., *Scientometrics* (2011)

Main topics

- General and mathematical properties of the h-index
  - Glänzel, *Science Focus*, 2006
  - Glänzel, *Scientometrics*, 2006
  - Glänzel, *Scientometrics*, 2008a,b
- Application of the h-index to scientific journals
  - Braun et al., *Scientometrics*, 2006
  - Schubert & Glänzel, *Journal of Informetrics*, 2007
- Hirsch-type indices for characerisation and testing the tail properties of scientometric distributions
  - Schubert & Glänzel, *Journal of Informetrics*, 2010
  - Glänzel, *Scientometrics*, 2010

Research performance

E Glänzel, W., Leta, J. Thijs, B., Science in Brazil. Part 1: A macro-level comparative study. *Scientometrics*, 67 (1), 2006, 67–85.

E Glänzel, W., Debackere, K., Meyer, M., 'Triad' or 'Tetrad'? On global changes in a dynamic world. *Scientometrics*, 74 (1), 2008, 71–88.

E Zhou, P., Thijs, B., Glänzel, W., Is China also becoming a giant in social sciences? *Scientometrics*, 79 (3), 2009, 593–621.

E Zhou, P., Thijs, B., Glänzel, W., Regional analysis on Chinese scientific profile. *Scientometrics*, 81 (3), 2009, 839–857.

E Zhou, P., Glänzel, W., In-depth analysis on China's international cooperation in science. *Scientometrics*, 82 (3), 2010, 597–612.

N Schubert, A., Glänzel, W., Cross-national preference in co-authorship, references and citations. *Scientometrics*, 69 (2), 2006, 409–428.

N Glänzel, W., Schlemmer, B., Schubert A., Thijs, B., Proceedings literature as additional data source for bibliometric analysis. *Scientometrics*, 68 (3), 2006, 457–473.

N Bolaños-Pizarro, M., Thijs, B., Glänzel, W., Cardiovascular research in Spain. A comparative scientometric study. *Scientometrics*, 2010, 85 (2), 509–526.

N Zimmerman, E., Glänzel, W., Bar-Ilan, J., Scholarly collaboration between Europe and Israel: A scientometric examination of a changing landscape. *Scientometrics*, 78 (3), 2009, 427–446.

N Zhang, L., Rousseau, R., Glänzel, W., Document-type country profiles. *Journal of the American Society for Information Science and Technology*, 2011, 62 (7), 1403–1411.

E: Emerging economies; N: National research performance

Research performance

I Leta, J., Glänzel, W., Thijs, B., Science in Brazil. Part 2: Sectoral and institutional research profiles. *Scientometrics*, 67 (1), 2006, 87–105.

I Thijs, B., Zimmerman, E., Bar-Ilan, J., Glänzel, W., Israeli research institutes: a dynamic perspective. *Research Evaluation*, 18 (3), 2009, 251–260.

I Thijs, B., Glänzel, W., A structural analysis of benchmarks on different bibliometrical indicators for European research institutes based on their research profile. *Scientometrics*, 79 (2), 2009, 377–388.

I Thijs, B., Glänzel, W., A structural analysis of collaboration between European research institutes. *Research Evaluation*, 2010, 19 (1), 55–56.

D Glänzel, W., Veugelers, R., Science for wine: A bibliometric assessment of wine and grape research for wine producing and consuming countries. *American Journal of Enology and Viticulture*, 57 (1), 2006, 23–32.

D Glänzel, W., Janssens, F., Thijs, B., A comparative analysis of publication activity and citation impact based on the core literature in bioinformatics. *Scientometrics*, 79 (1), 2009, 109–129.

D Meyer, M., Debackere, K., Glänzel, W., Can Applied Science Be 'Good Science'? Exploring the Relationship Between Patent Citations and Citation Impact in Nanoscience. *Scientometrics*, 2010, 85 (2), 527–539.

D Glänzel, W., Zhou, P., Publication activity, citation impact and bi-directional links between publications and patents in biotechnology. *Scientometrics*, 2011, 86 (2), 505–525.

C Czarnitzki, D., Glänzel, W., Hussinger, K., Patent and publication activities of German professors: an empirical assessment of their co-activity. *Research Evaluation*, 16 (4), 2007, 311–319.

C Czarnitzki, D., Glänzel, W., Hussinger, K., Heterogeneity of patenting activity and its implication for scientific research. *Research Policy*, 38 (1), 2009, 26–34.

I: Institutional research performance; N: Domain studies; C: Co-activity studies

The hybrid approach for dynamic and structural analysis

The integration of citation-based and text-based methods has three important fields of application in scientometrics.

- Mapping the cognitive structure of science
- Subject-classification issues
- Bibliometrics-aided retrieval

☛ **NB: The objective defines the method!**

Methods: Co-word (CW); Term frequency (TF)

Advantages

- Works with traditional abstract databases.
- Labelling outcomes (e.g., by using the best TF-IDF terms).

Disadvantages

- Preferably applied to full text ⇒ however, full text contains links.
- Dimensionality; homonyms and synonyms ⇒ solution: SVD.
- "smooth" approach ⇒ tends to overestimate links
- sensitive to the peculiarites of "global" and "local" clustering
- less suited for longitudinal studies (changing vocabulary)

The link approach

Methods: Co-citation (CC); Bibliographic coupling (BC); direct link (DL)

Advantages

- Added values: BC with "retrieval effect", CC for research front, DL with information flow
- Suited for longitudinal studies, if properly applied

Disadvantages

- Works only with citation indexes or full-text databases.
- Sparse matrices (can be solved ⇒ smoothing the 'singularity' but decreases efficiency).
- Tends to polarise relationship (underestimates relationship)
- Requires large citation window for DL, CC; critical mass needed for CC (Hicks, 1987)

Combining full-text with link-based clustering

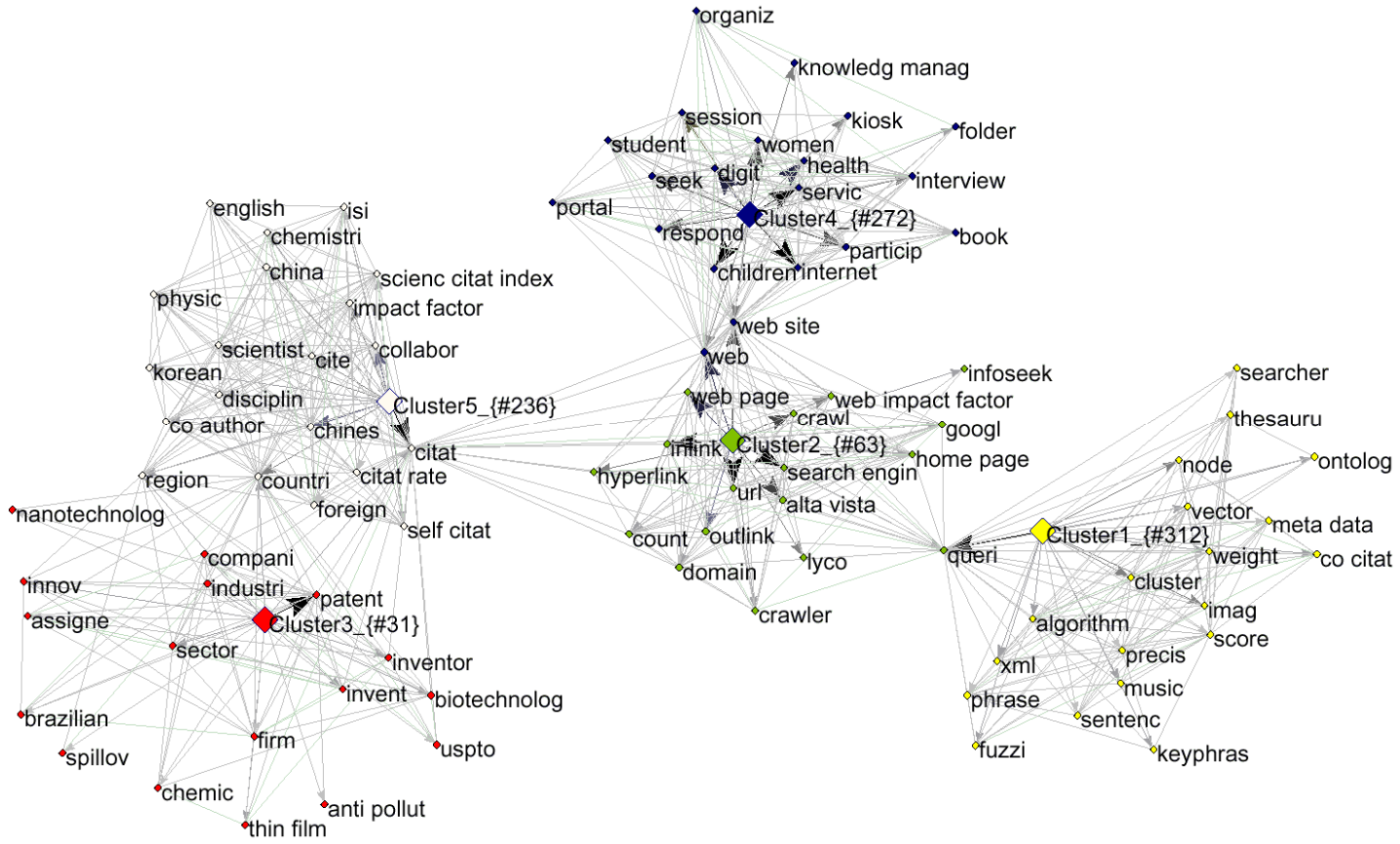
The following solutions have been tested:

1. Vector-space model
   - Concatenation of vectors
   - Linear combination of distances
   - Linear combination of angles
   - Fisher's inverse chi-square method
2. Graph model
   - Graph integration
   - Graph coupling

Hybrid clustering makes the best of the two worlds and allows labelling based on the textual component.

Cognitive structure of LIS based on hybrid approach

Cognitive structure of LIS using the best TF-IDF terms for five clusters



Janssens et al., *Scientometrics*, 2008

Improving subject classification based on clustering

- Not merely visualising the structure of science by presenting yet another map using alternative approaches,
- Validating and improving existing (journal-based) subject classifications used for research evaluation,
- Using an existing subject classification scheme as a “control structure”.

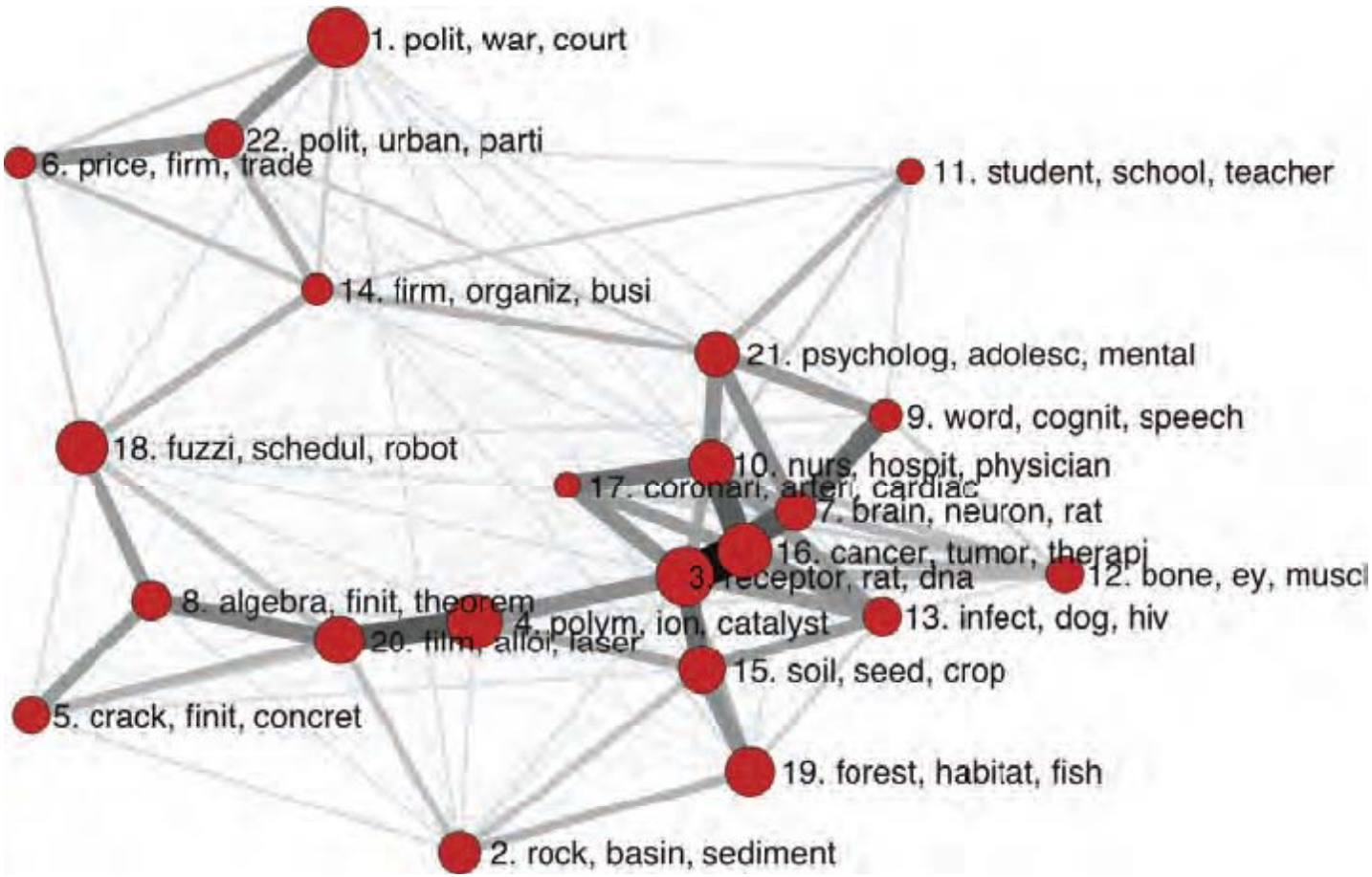
- All papers of a database have to be assigned (not only a representative set of papers based on cited or retrieved documents)
- Evaluation of existing schemes as if those were results of clustering.
- Evaluative comparison of mapping and reference structure

Cognitive mapping vs. subject classification

Methods

- Evaluation based on Silhouette values, Modularity, Jaccard index, Rand Index and F-scores
- Three clustering approaches: cross-citation, textual and hybrid
- Labelling subject fields based on best TF-IDF terms for both reference and cluster structure
- Studying concordance between clustering solution and the reference classification scheme
- Migration of journals among subject fields and clusters
- Application to 22 ESI fields (partition), 15-field Leuven scheme (fuzzy) (and 7 fields suggested by the algorithm)

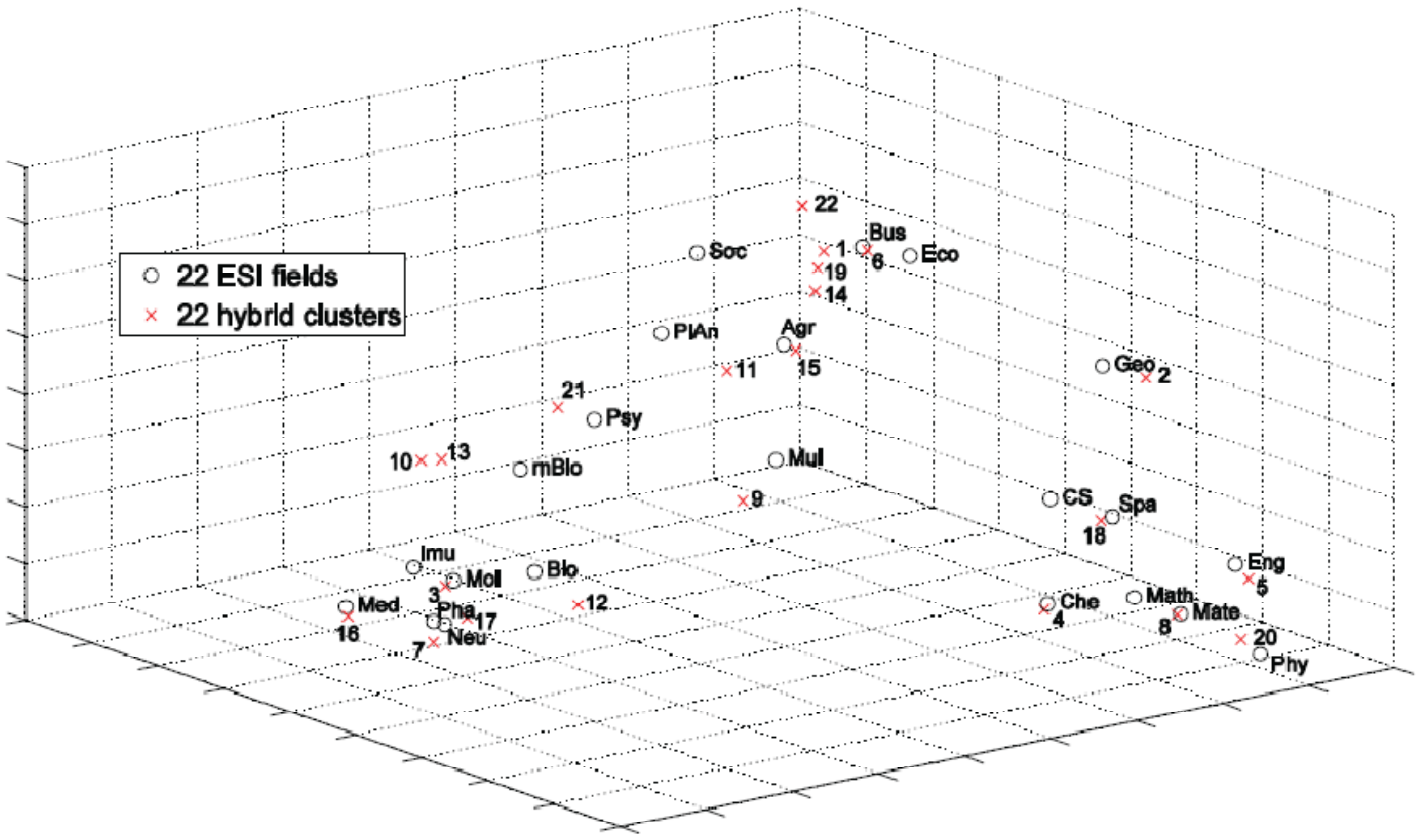
Network structure of hybrid clusters with 3 most important TF-IDF terms



Source: Janssens et al., *Information Processing & Management* (2009)

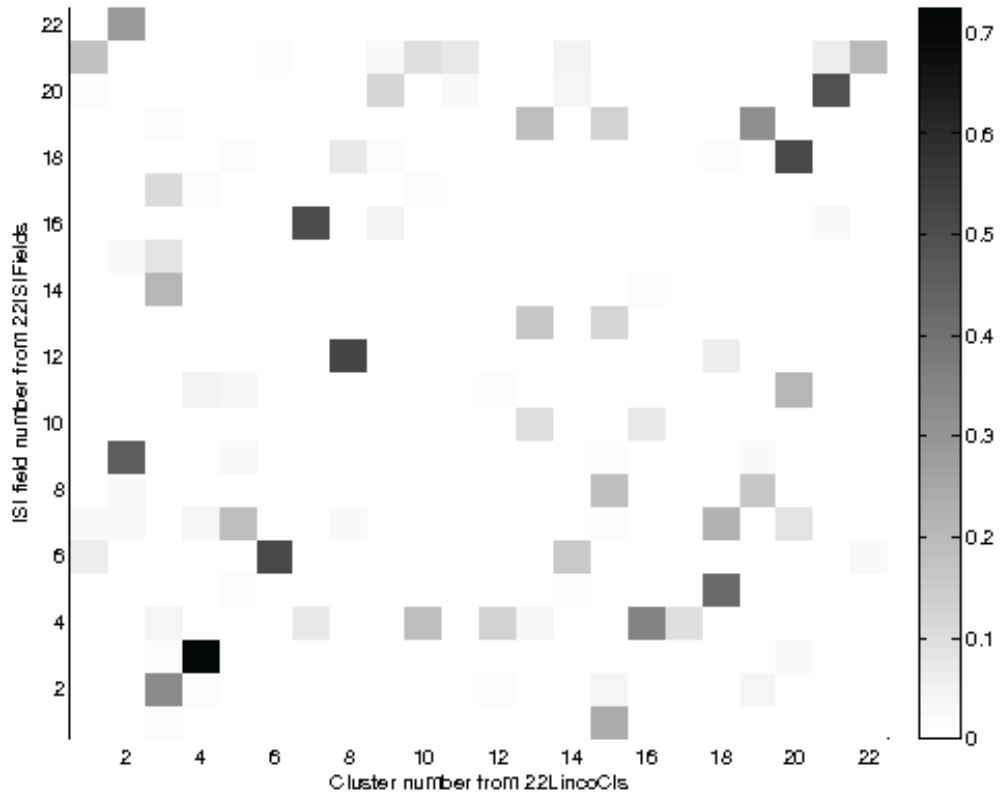
Cognitive mapping vs. subject classification

3-dimensional MDS map visualising distances between the centroids of the 22 fields and clusters



Source: Janssens et al., *Information Processing & Management* (2009)

Concordance between clustering solution and the ESI Scheme based on the Jaccard index



Source: Janssens et al., *Information Processing & Management* (2009)

Some more results

A trivial 3- and an interesting 7-cluster scheme has been found.

1. social sciences
   - 1.1 economics, business and political science
   - 1.2 psychology, sociology, education,
2. life-sciences & medical sciences
   - 2.1 biosciences and biomedical research
   - 2.2 clinical, experimental medicine and neurosciences
3. natural & technical sciences
   - 3.1 biology, agriculture and environmental sciences
   - 3.2 physics, chemistry and engineering
   - 3.3 mathematics and computer science

Important steps in the bibliometric analysis of emerging topics

1. Structural analysis of the discipline
   Preferably based on hybrid methods
2. Dynamic analysis of the discipline
   Synchronistic approach required
3. Identification of emerging topics
   Example: ERACEP (2010-2012)
4. Delineation of the topic (optional)
   Requires sophisticated search strategies. *Example*: Bioinformatics (2006-2009)
5. Network analysis of the topic
   Concerns both internal structure and links to other field (environment).
   Example: Ongoing project on 'entrepreneurship research' with Univ Sussex and UMKC
6. Bibliometric study of the topic
   Identification of main actors, co-publications, citation-impact analysis.
   Example: Bioinformatics (2006-2009)

Specific problems

- At this level of aggregation (topics within the same discipline), terms and phrases might become less specific since they express common knowledge base and vocabulary. Others might gain more 'information value'.
- Keywords and terms proved not specific enough for topic description and labelling.

Solution

- Depending on the level of aggregation *and* the discipline under study, the weight of the two components can be adjusted.
- Instead of the best TF-IDF terms *core documents* can be used to describe and label clusters.

The notion of a "core" of literature goes back to *co-citation analysis*.

- Small, *JASIS*, 1973.

Core documents were defined as papers, which have at least *n* links of at least a given strength *r* according to a given similarity measure based on bibliographic coupling.

- Glänzel & Czerwon, *Scientometrics*, 1996

This notion can be extended to any *hybrid* method, e.g., combining bibliographic coupling with text-based methods and or co-citation links.

- Glänzel & Thijs, *Scientometrics*, 2011

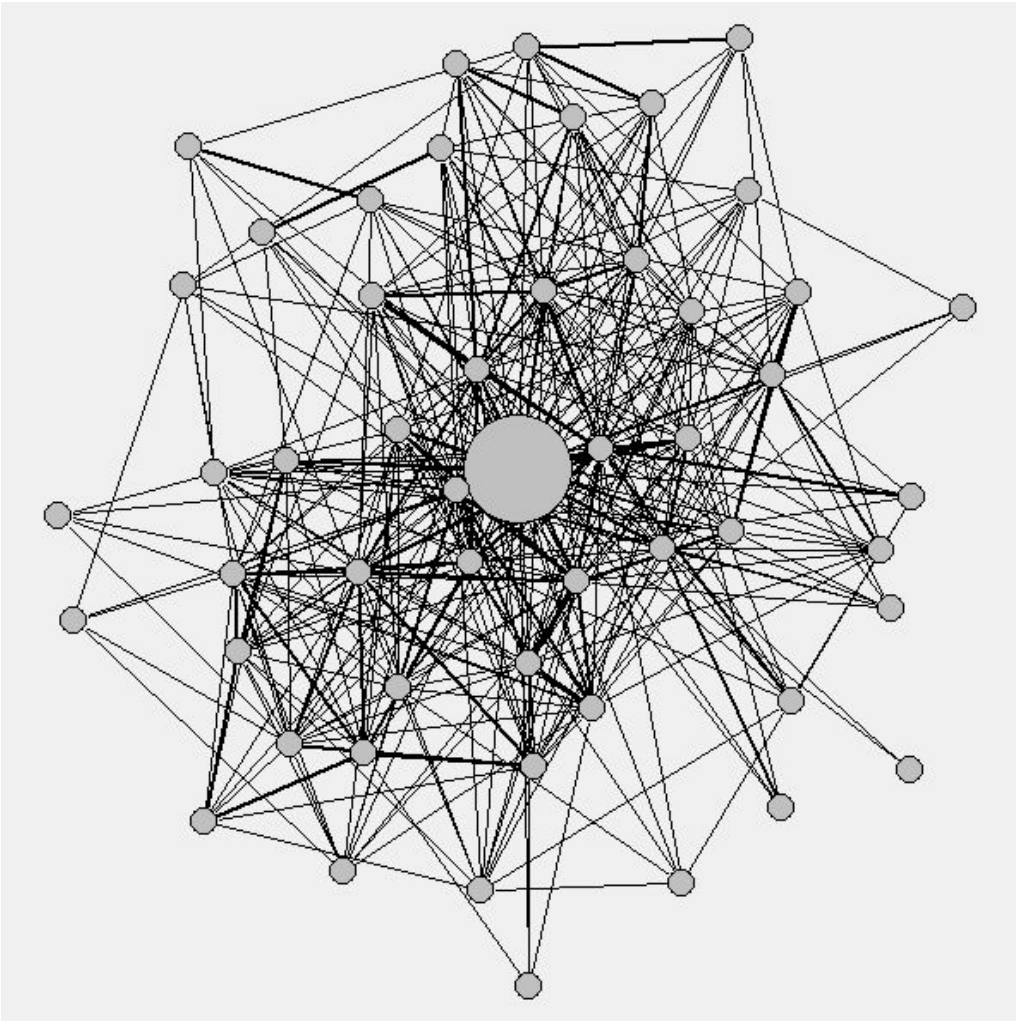
Definition A network has a degree h-index is *h* if not more than *h* of its nodes have a degree not less than *h*.

- Schubert et al., *Scientometrics*, 2009

Definition: Core vertices are vertices with at least *h* degrees each, where *h* is the h-index of the graph.

- Glänzel, *Scientometrics*, 2012

Visualisation of the link environment of a ‘core document’



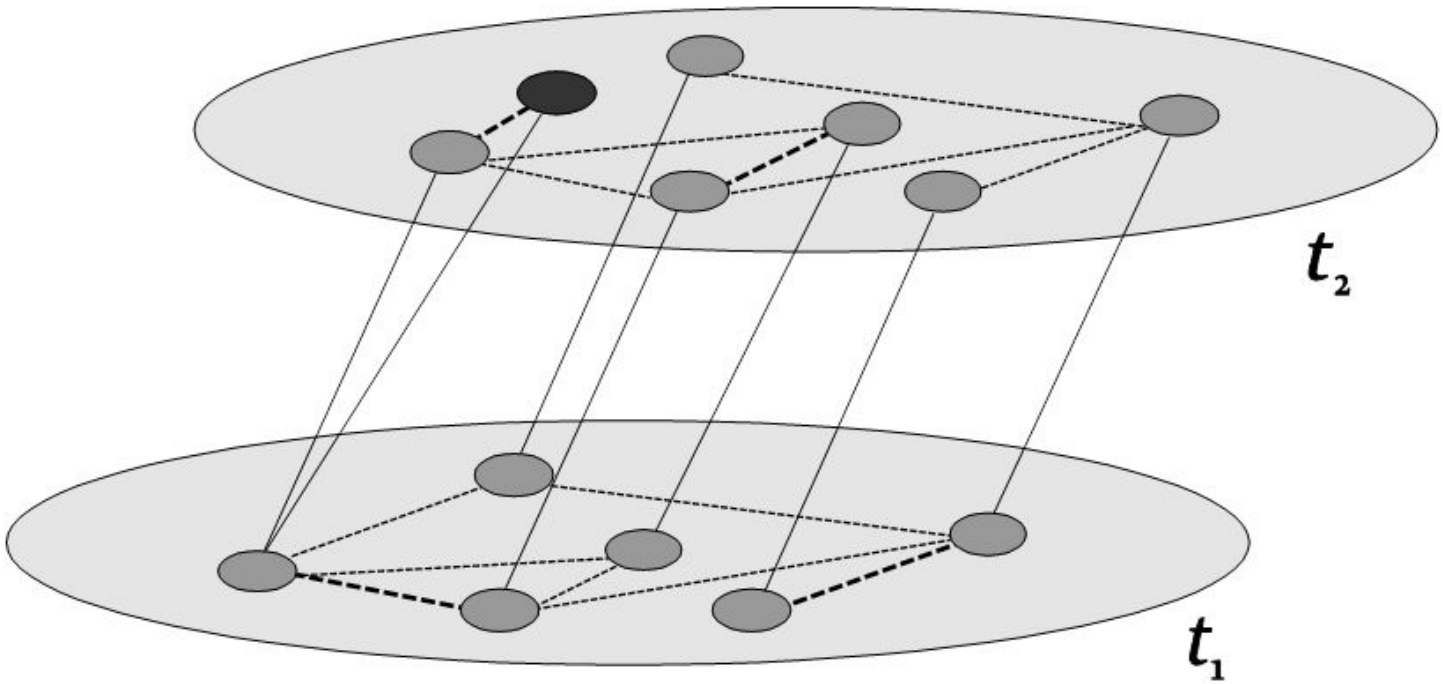
Source: Glänzel & Thijs, *Scientometrics*, 2012

Dynamic aspects

Techniques for detecting new emerging yet coherent structures

- Topics are searched in the mirror of their scholarly literature.
- Lexical approach: Growing frequency of specific terms within a given research area
- Extracted terms can be used for labelling and describing the obtained clusters.
  - Lamirel et al., *IASTED – AIA*, 2008
- Textual similarity based on shared terms is also related to strong citation links
  - Jo et al., *ACM SIGKDD – KDD*, 2007
- Topological measures to determine the role of each paper in the citation network can be used to decide whether there are emerging clusters.
  - Shibata et al., *Technovation*, 2008
- After clustering a discipline in disjoint periods a link analysis among papers in clusters of the different periods is conducted.
  - Glänzel & Thijs, *Scientometrics*, 2012

Sketch of a research field's changing topic structure over time
(dotted lines: internal links, solid lines: links between the time slides t_1 and t_2)



Source: Glänzel & Thijs, *Scientometrics*, 2012

Three paradigmatic cases of cluster evolution can be distinguished.

(1) Existing cluster with an exceptional growth,

(2) Completely new cluster with its root in other clusters and

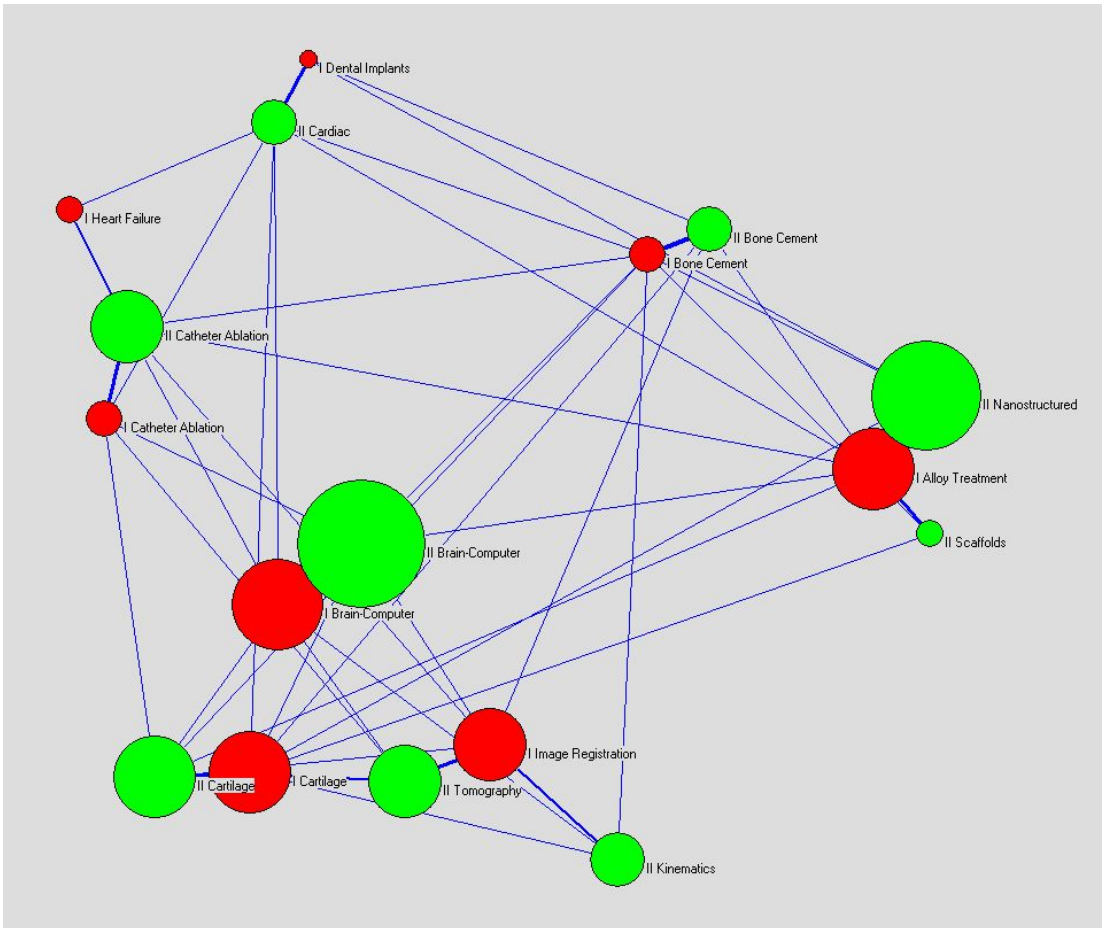
(3) Existing cluster with a topic shift.

☛ NB: It should be mentioned that evolution also works in the opposite direction in case (1) and (2), say, as declining or vanishing topics.

When can we speak about a 'new emerging topic'?

- A rapidly growing number of publications and scientists dealing with this topic is required.
- Its literature has reached a critical mass.
- The topic must be coherent, have a certain independence of its "mother topic" and other disciplines.
- It must be largely self-sustaining.

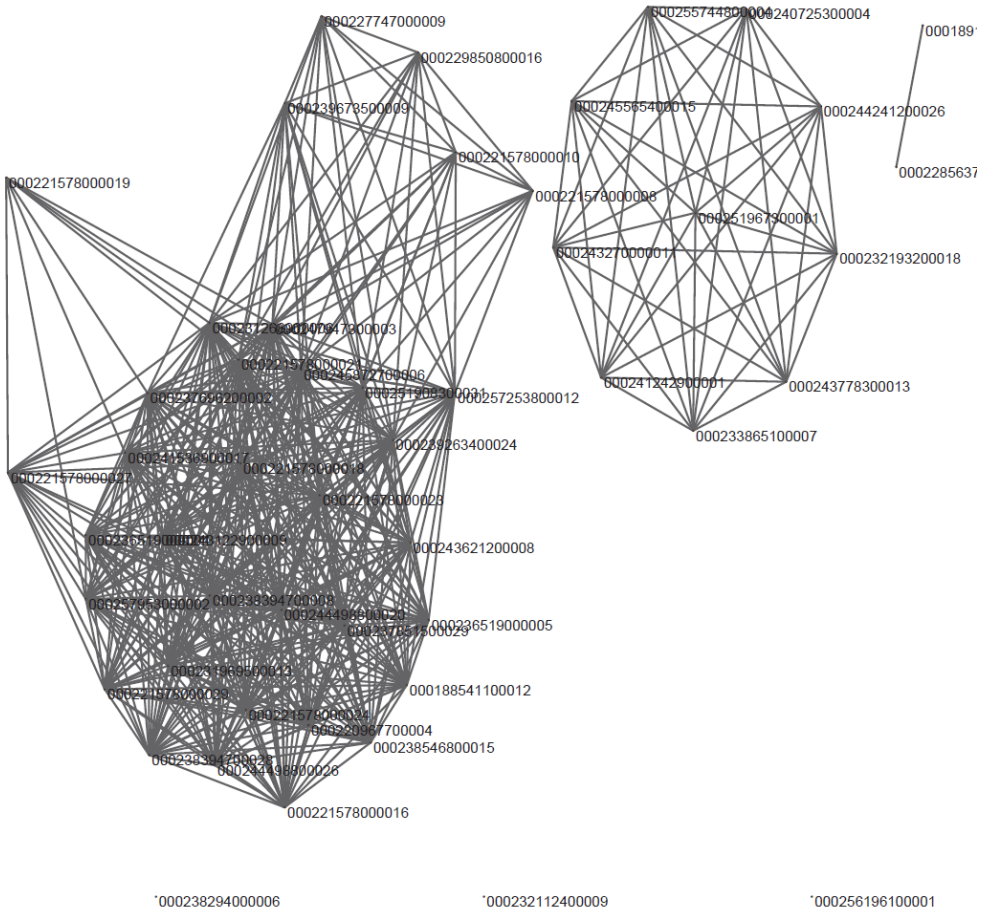
Cluster representation of the Subject Category 'Biomedical engineering'
(Kamada-Kawai layout; Red: 1999–2003, Green: 2004–2008)



Source: Glänzel & Thijs, *Scientometrics*, 2012

Cluster representation

Internal structure of the topic “brain-machine interface” (2004-2008) within the subject category “engineering, biomedical” (Pajek; Kamada-Kawai layout)



Source: Glänzel, *Scientometrics*, 2012

Selected core documents representing the topic "brain-machine interface" (2004-2008) within the subject category "engineering, biomedical"

ISI UT-code	Document title
000188541100012	"Virtual keyboard" controlled by spontaneous EEG activity
000189183300028	Planar gradiometer for magnetic induction tomography (MIT): theoretical and experimental sensitivity maps for a low-contrast phantom
000220967700004	Adaptive BCI based on variational Bayesian Kalman filtering: An empirical evaluation
000221578000008	Model-based neural decoding of reaching movements: A maximum likelihood approach
000221578000010	Ascertaining the importance of neurons to develop better brain-machine interfaces
000221578000016	Anasynchronously controlled EEG-based virtual keyboard: Improvement of the spelling rate
000221578000018	Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms
000221578000019	Support vector channel selection in BCI
000221578000021	Classification of single-trial electroencephalogram during finger movement
000221578000023	BCI2000: A general-purpose, brain-computer interface (BCI) system
000221578000024	The BCI competition 2003: Progress and perspectives in detection and discrimination of EEG single trials
000221578000027	BCI competition 2003 - Data set IIa: Spatial patterns of self-controlled brain rhythm modulations
000221578000029	BCI competition 2003 - Data set IIb: Support vector machines for the P300 speller paradigm
000227747000009	Closed-loop cortical control of direction using support vector machines
000228563700029	A new type of gradiometer for the receiving circuit of magnetic induction tomography
000229850800016	Interpreting spatial and temporal neural activity through. a recurrent neural network brain-machine interface
000231268900006	Spatio-spectral filters for improving the classification of single trial EEG
000231969500013	Sensorimotor rhythm-based brain-computer interface (BCI): Feature selection by regression improves performance
000232112400009	Resonance behaviour of whole-body averaged specific energy absorption rate (SAR) in the female voxel model, NAOMI
000232193200018	Automated methodology for determination of stress distribution in human abdominal aortic aneurysm
000233865100007	A patient-specific computational model of fluid-structure interaction in abdominal aortic aneurysms
000236519000004	Robust classification of EEG signal for brain-computer interface
000236519000005	Steady-state somatosensory evoked potentials: Suitable brain signals for brain-computer interfaces?
...	...

Source: Thomson Reuters – Web of Knowledge, 2011; ERC–ERACEP 2011

Thank you very much for your attention.
Vielen Dank für Ihre Aufmerksamkeit!
Hartelijk dank voor uw aandacht!
Köszönöm szépen a figyelmüket!
Molte grazie per la vostra attenzione.