Library and Information Centre of
the Hungarian Academy of Sciences
Department of Science Policy and Scientometrics

Scientometrics as network science

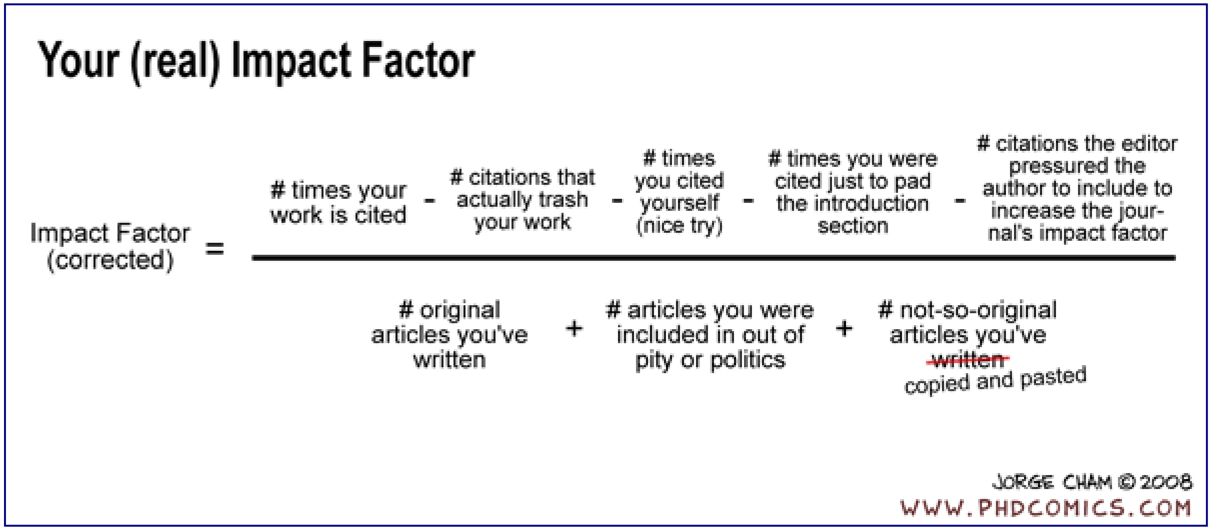
The hidden face of a misperceived research field

Sándor Soós, PhD
soossand@konyvtar.mta.hu

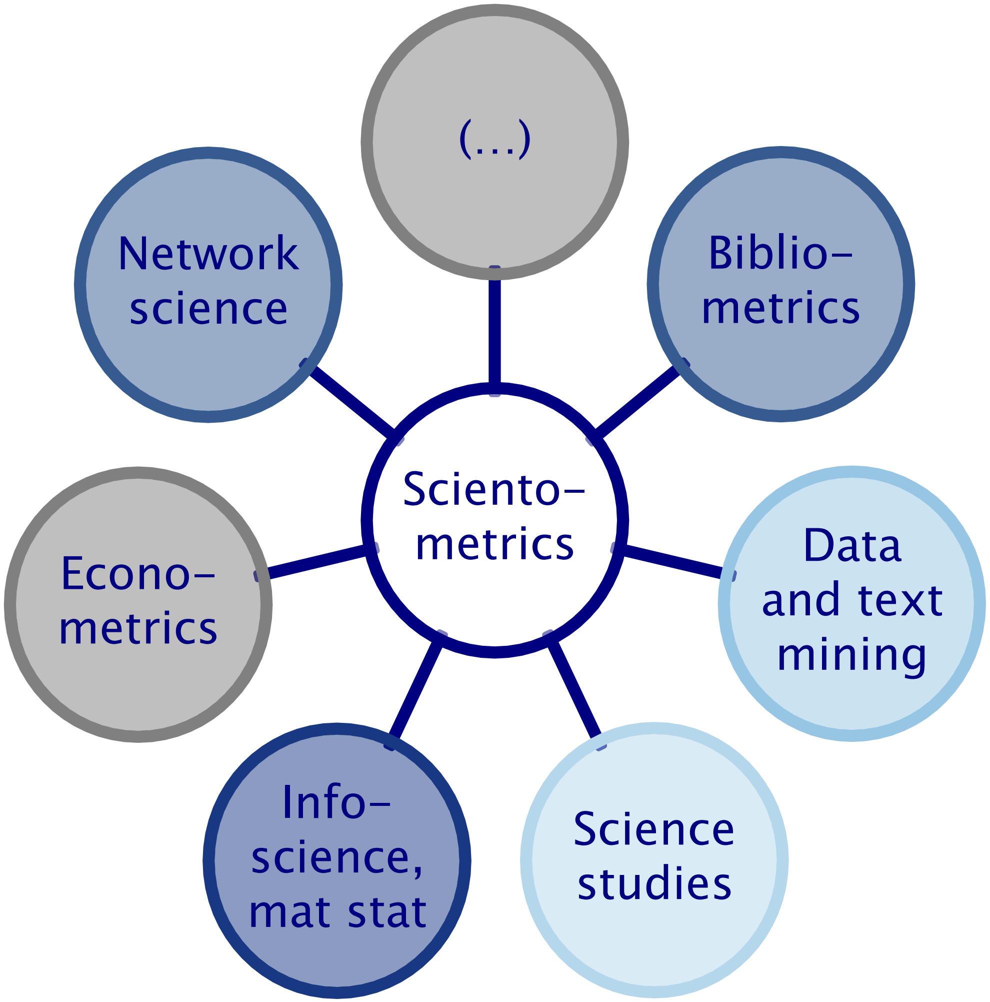
MTAK 1826

„Public understanding" of scientometrics

- Three common misperceptions:
  - Scientometrics is publication statistics (science administration's view)
  - Scientometrics is exclusively concerned with the measurement of scientific performance (researcher's view)
  - Scientometrics is a form of research evaluation (policy maker's view)



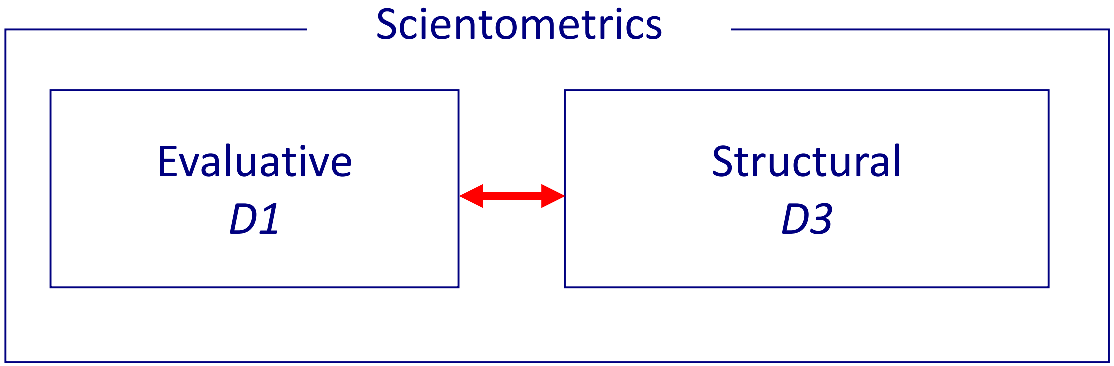
Disciplinary composition today



Research directions

- **(D1)** Development of [...] quantitative indicators on important aspects of S&T
- **(D2)** Development of information systems on S&T
- **(D3)** Study of cognitive and socio-organizatonal structures of scientific fields […] (and other aggregates - SS) in relation to societal factors

A.F.J. Van Raan, 1997



Structural scientometrics

- [D3] is the „old" sociological root of [scientometrics], makes it instrumental to [sociology of science].
    - *A.F.J. Van Raan, 1997*

- Instruments: formal models of the socio-cognitive organization of science: science maps

Network models

to be constructed and analysed via the rich toolbox of SNA

Social networks?

A typology of network models as science maps

- Dimensions: (1) types of relations and (2) level of aggregation (determinants of meaning)

- **1. Collaboration networks**
  - **Individual level:** co-author networks.
    - Meaning: cognitive structure. The community structure represents building blocks of current science (fields, schools, research directions etc. (Where appropriate.) Well-studied.

  - **Aggregated levels** (institutions, countries etc.):
    - Meaning: the institutional organization of science

A typology of network models as science maps

- Dimensions: (1) types of relations and (2) level of aggregation (determinants of meaning)

- **2. Information/Knowledge flow** networks, relation: citation
  - **Document level:** doc citation networks.
    - Meaning: knowledge flow, knowledge diffusion, historical relations of ideas („algorithmic historiography", E. Garfield). Type: Inverse, unweighted directed graphs.
  - **Aggregated levels:** nodes are document sets (individuals, journals etc.)
    - Meaning: cognitive organization of science, communities as buliding block. Type: weighted, undirected graphs.

A typology of network models as science maps

- Dimensions: (1) types of relations and (2) level of aggregation (determinants of meaning)

- **3. Proximity networks**, relation: induced proximities, not actual interactions („social networks”)
  - Indicator: textual descriptors→ *co-word networks*.
    - Meaning: cognitive, conceptual structure (e.g. research fronts). The community structure represents building blocks of current science (research problems, foci, fields, schools, research directions etc.
  - Indicator: references, citations → *bibliographic coupling, co-citation networks*
    - Meaning: the institutional organization of science

Global maps of science

- Demonstration of the interplay between evaluative scientometrics and science mapping
- A running example:
  - » construction and application of a global science map
  - » Development into an analytical framework informing sociology of sci and evaluative studies
  - » Own contributions to the model

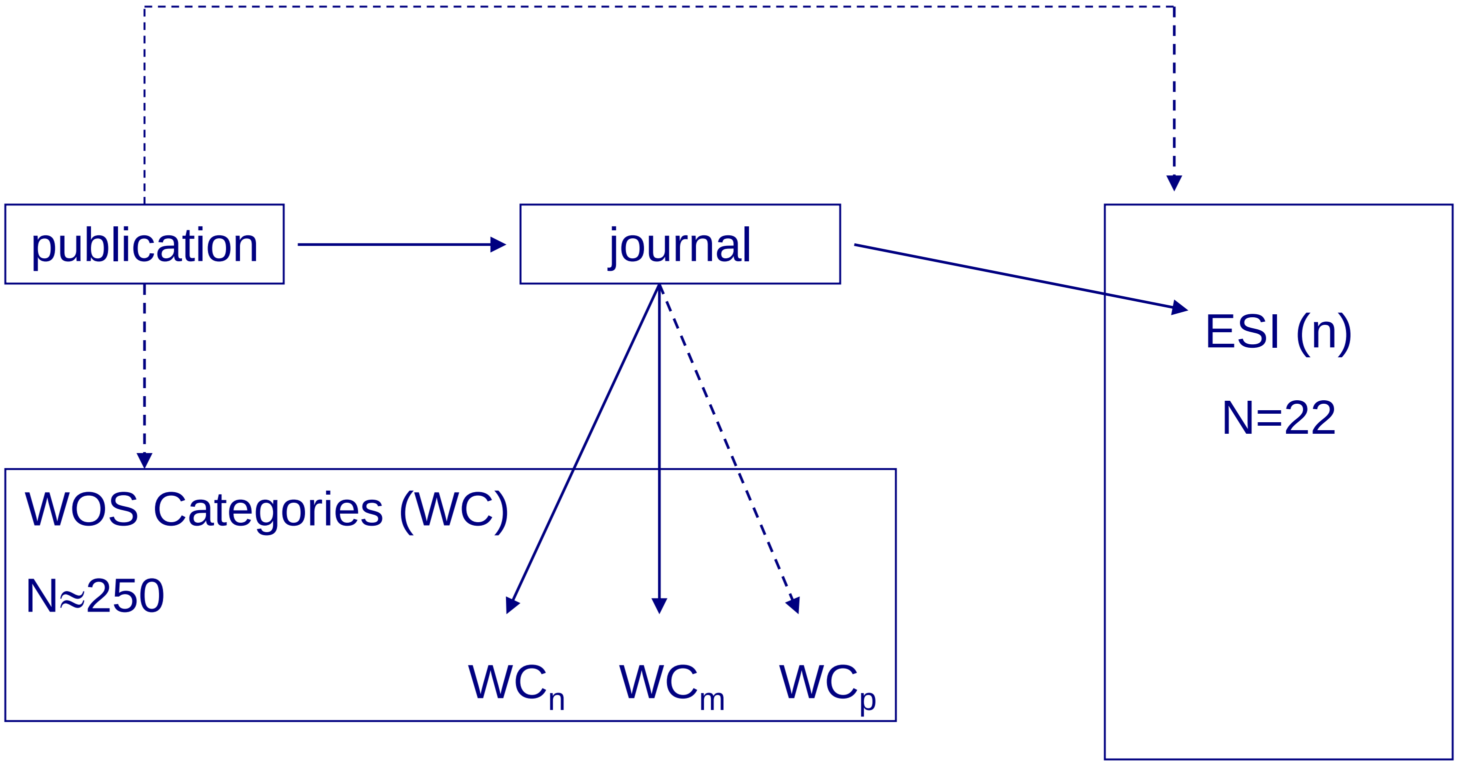
- Global science maps: proximity networks

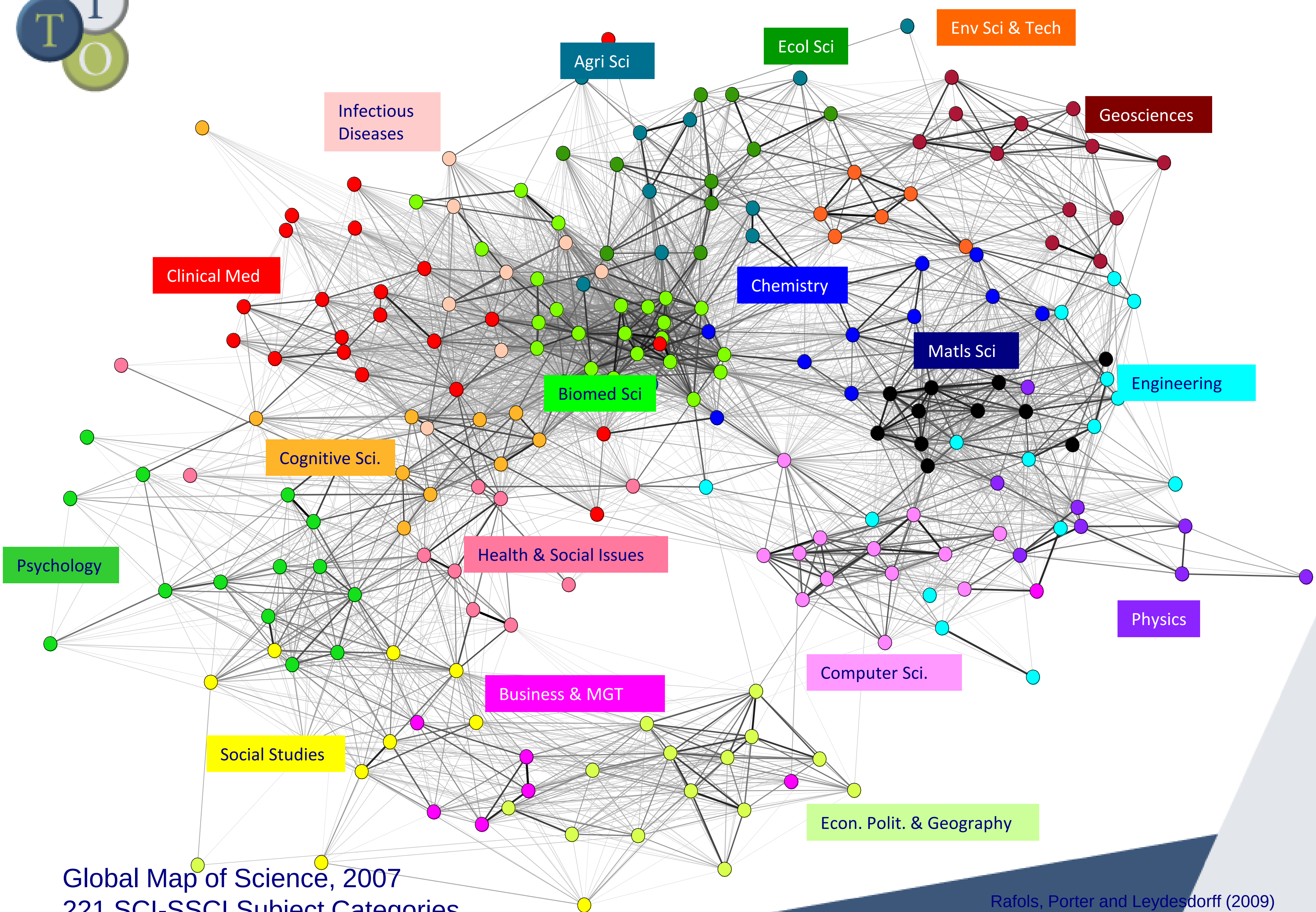
Example chosen:

global science map based on WoS Subject Categories (Rafols-Leydesdorff, 2007)

A global map constructed on WoS Subjects

- Based on journal categorization in the Web of Science





Global Map of Science, 2007
221 SCI-SSCI Subject Categories

Rafols, Porter and Leydesdorff (2009)

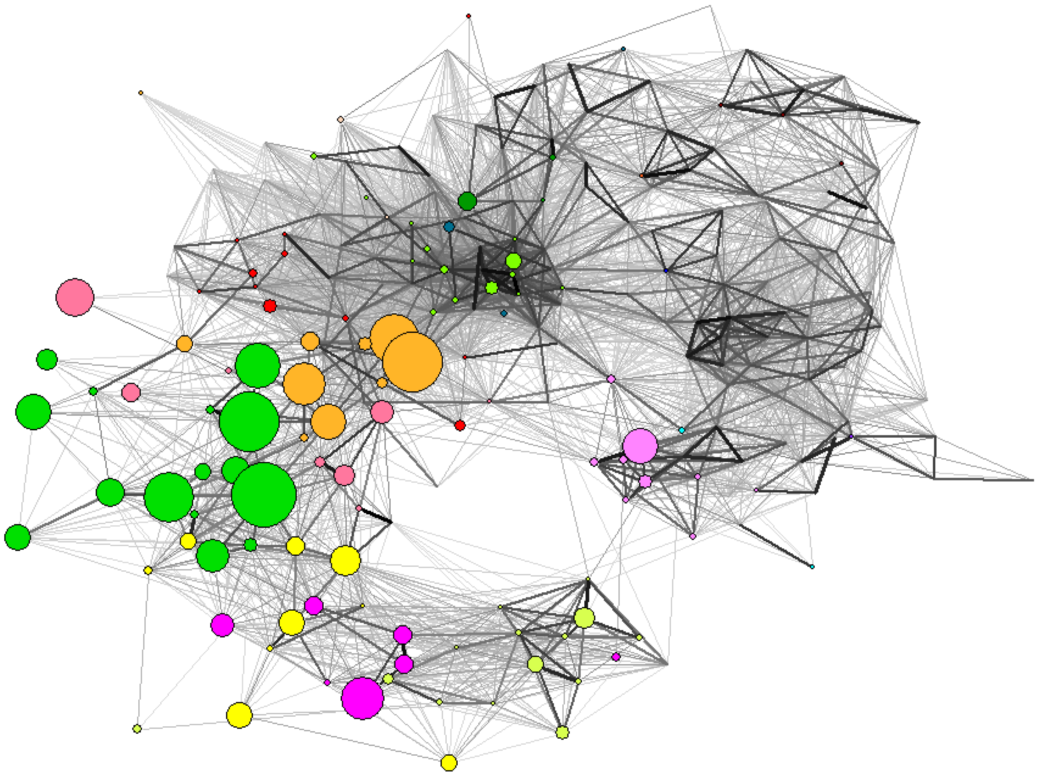
Construction of the map

- **Unit of analysis:** ISI Subject Category (SC)
- **The map:** the proximity or genealogy-based network of Subjects
- **Method:** „bibliometric coupling" of SCs
  - Principle: shared intellectual background (or inherited body of scientific knowledge)
  - The more references two subjects share, the more closer they are within the system of science (proximity in terms of citing the same SCs)
  - Techically: references are compared in terms of SCs (SC-SC references)
- **Disciplines**: clusters (factors) in the proximity network
  - PCA on the the proximity matrix for identifying coherent subject sets

Modelling research profiles

The science overlay technique

- Position of an actor within the scientific landscape=
  - Structure of its research profile
- Method: Mapping a set of publications onto the global map (basemap)
- SCs related to the publication record are highlighted, indicating their respective weights



Structural measures

- Measuring multi- and interdisciplinarity (IDR) upon this model: the Stirling index
- Novelty: Three structural features accounted for:
  - Number of SCs („variety”)
  - Distribution of pubs over SCs („balance”)
  - **Proximity/distance** of constituent SCs („disparity”)

Table 1 *Typology of the Stirling index in measuring research diversity*

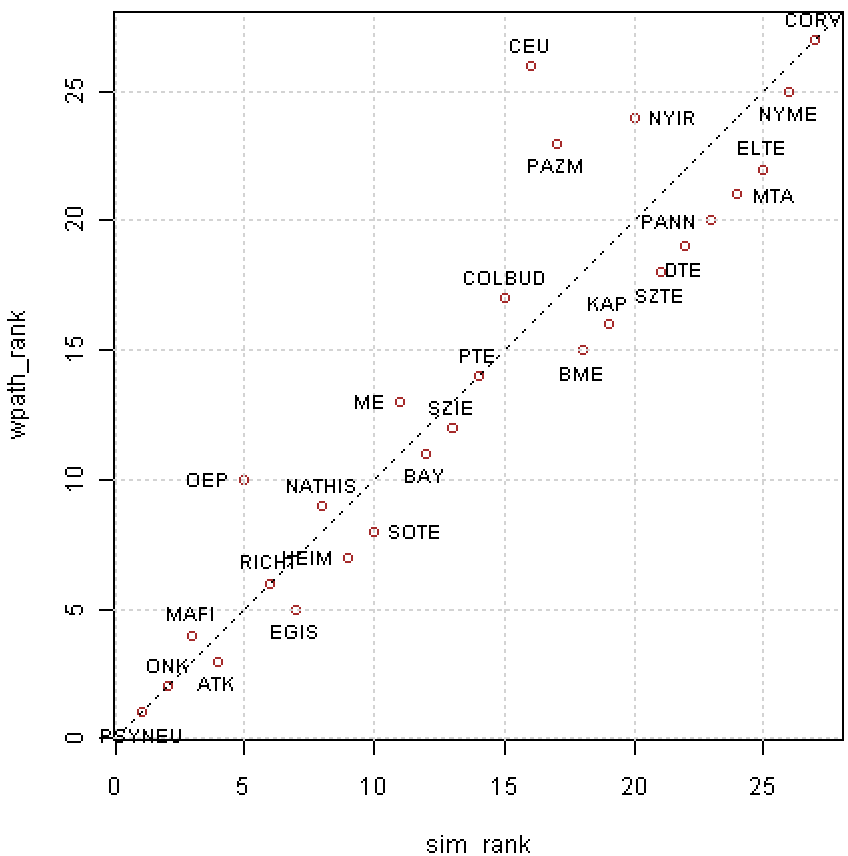
	Formula (versions of the generalized Stirling index)	d_{ij}	Underlying science map (level of aggregation)	Measuring diversity of…
1	$\sum_{ij(i \neq j)} d_{ij} p_i p_j$	$1 - s_{ij}$, where sij=cos(i,j)	Similarity network of (1) journals (2) ISI Subject Categories (based on the cited and citing dimension) Rafols, Meyer, Porter, Leydesdorff	(1) journals, (2) work of researchers, (3) output of organizations
2	$\sum_{ij(i \neq j)} d_{ij}$	g_{ij} shortest path from i to j (# edges)	Similarity network of papers (based on bibliographic coupling) Rafols, Meyer	particular research area

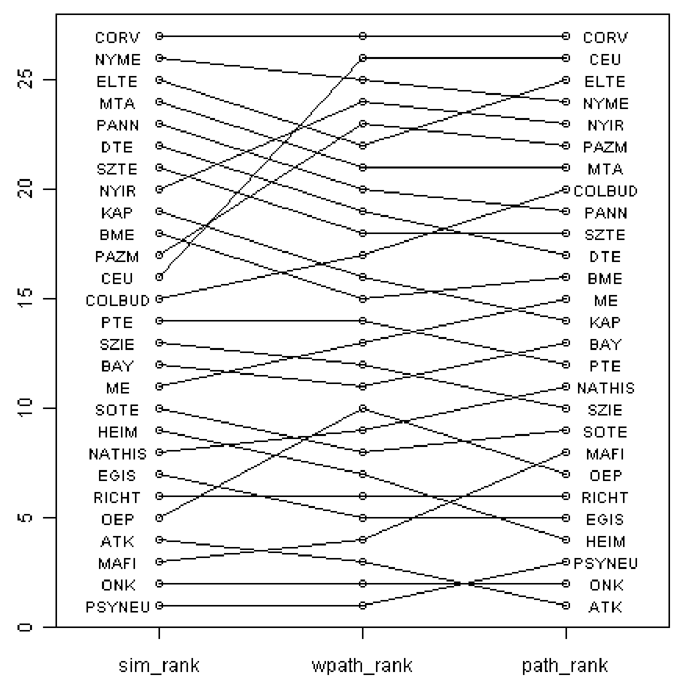
Structural measures

- „Polarity index”

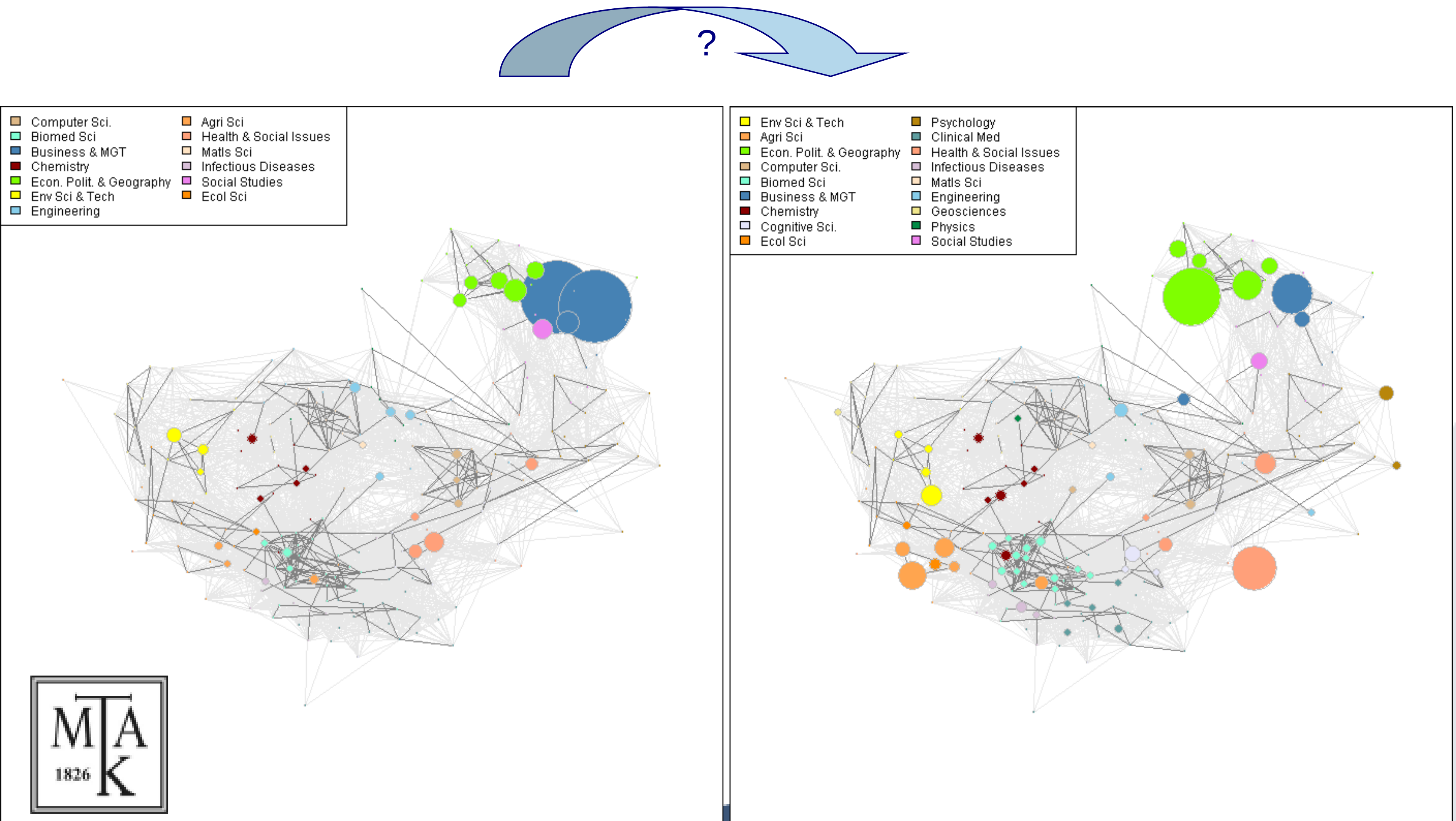
(Soós-Kampis, 2011, *Scientometrics)*

$$\sum_{ij(i\neq j)} g_{ij}^{W} p_i p_j \text{ , where } g_{ij}^{W} = \text{sum of the weights of edges in the shortest path form } i \text{ to } j,$$





Knowledge dynamics



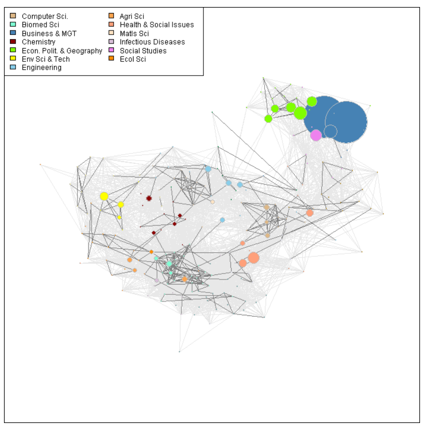
A flexible proposal

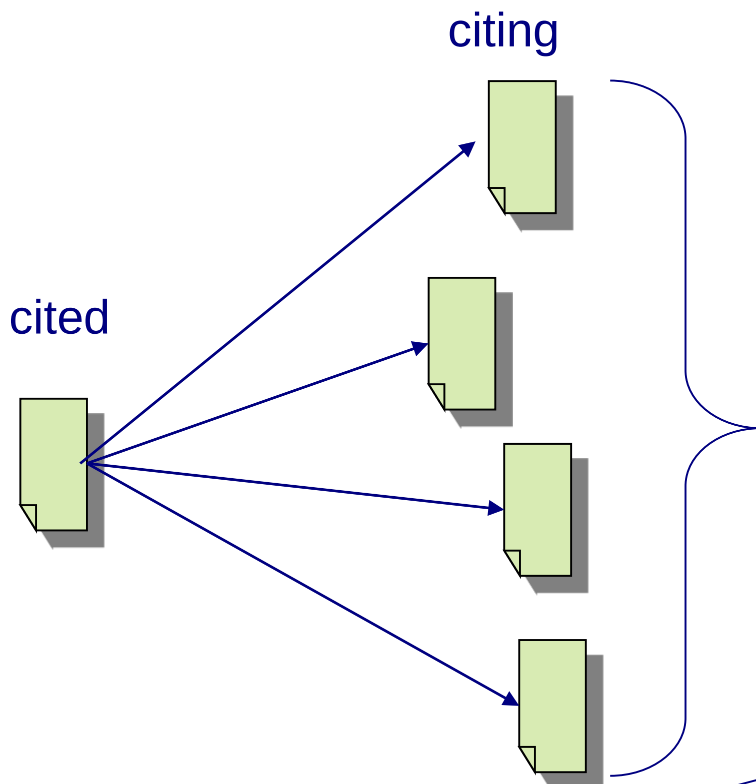
Mean Overlay Distance (MOD) = $\frac{1}{n*m}\sum_{i=1,j=1}^{n,m} p_i p_j d_{ij}$

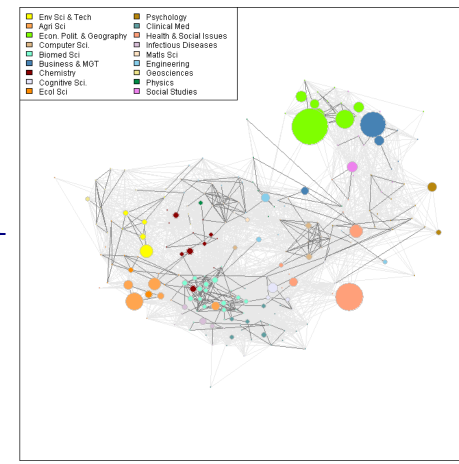
- p_i is the relative frequency of the i-th Subject Category within the **source** SC-profile, $i = 1, \ldots, n$,
- p_j is the relative frequency of the j-th Subject Category within the **target** SC-profile, $j = 1, \ldots, m$,
- d_{ij} is the distance of the i-th (source) and the j-th (target) Subject Category as determined by the (common) basemap for the (both) overlays.

The (average) distance between two overlay maps
based on pairwise (weighted) cognitive distances between constituent SCs

App1: development of science







$$\frac{1}{n*m}\sum_{i=1,j=1}^{n,m} p_i p_j d_{ij}$$

App1: development of science

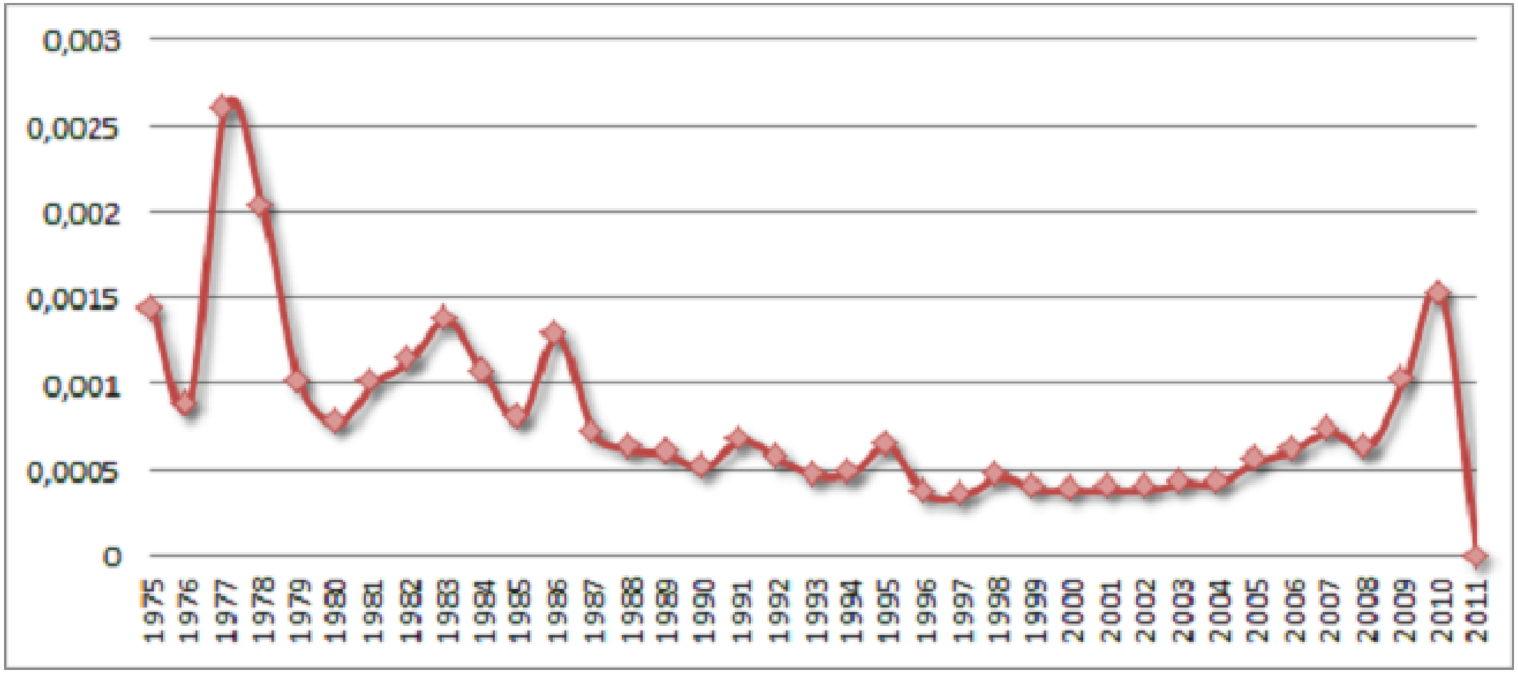
- MOD: **measuring knowledge diffusion/integration** through citation networks (evolution of a scholarly discourse)
- A detailed, large-scale case study: the species problem

Table 2. *Statistics of iterative corpus collection on the Species Problem based on WoS databases*

Iteration	*No. of source documents*	*No. of references*	*No. of unique references*	*Threshold value*	*No. of relevant references (retrievable)*
Initial corpus	1605	93 943	50 668	3	3223
2. generation	3223	155 742	62 574	10	851
3. generation	851	14 991	5305	10	2
Total	**5679**				

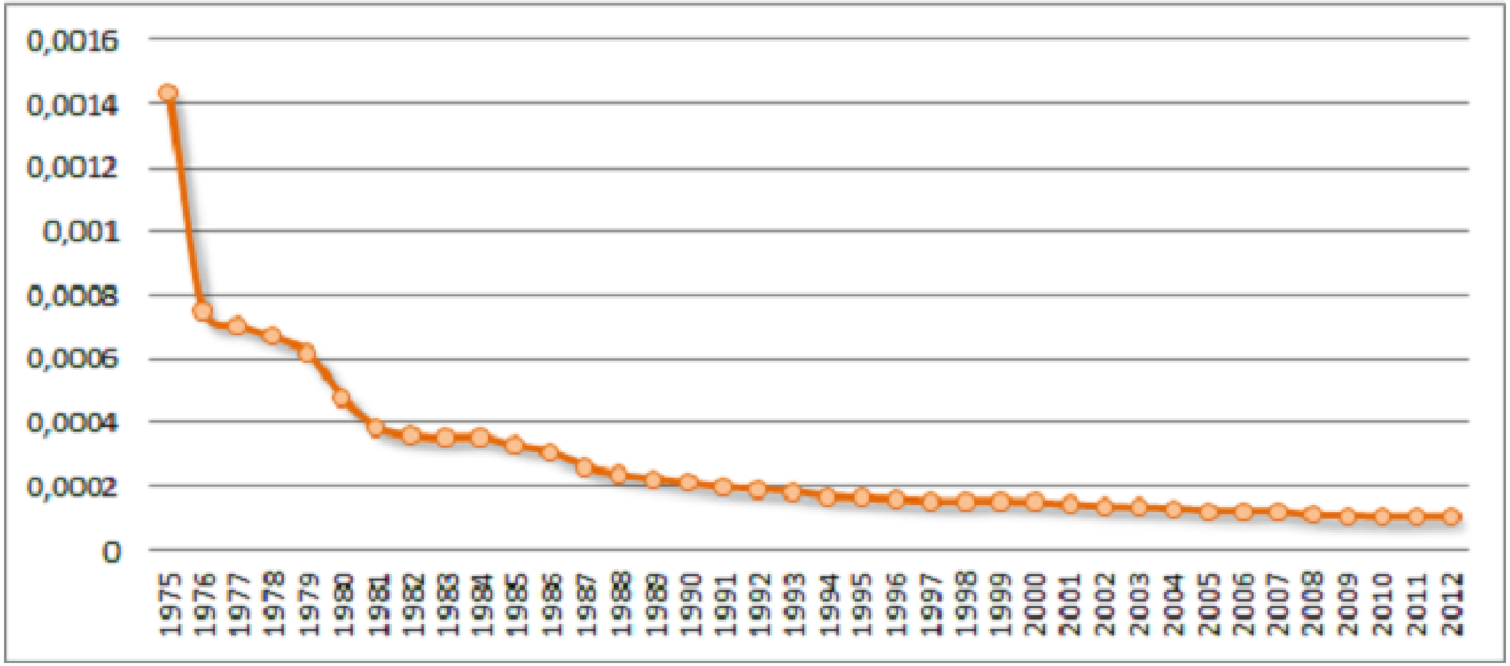
App1: development of science

Fig. 2 *Development of the MOD index comparing annual sections and their citing environment*



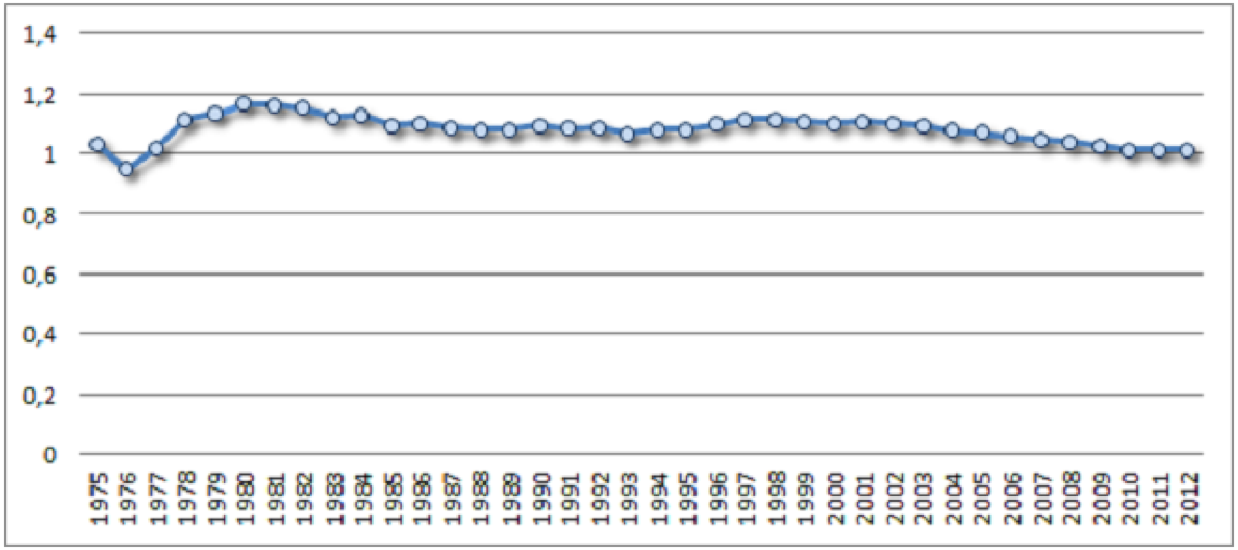
App1: development of science

Fig. 3 *Development of the MOD index comparing accumulated papers wit their citing environment*



App1: development of science

Fig. 4. *Development of the ODR index comparing the diversity of accumulated papers up to each year with the diversity of their citing environment*



$$\mathrm{ODR} = \frac{OD_{target}}{OD_{source}}, \text{ whereby}$$

- OD_{target} is the Overlay Diversity of the target set (as measured by the Stirling index),
- OD_{source} is the Overlay Diversity of the source set (as measured by the Stirling index).

App2: research evaluation

- MOD as an evaluative/impact measure

- Usual impact measures: based on quantity
  - Absolute (number of cits)
  - Normalized (field-normalized relative impact)
  - Weighted (eigenfactor)

MOD in this context: **scope of citation impact**

- MOD as an impact measure:
  - How far (distance) a publication gets from its own research field, i.e. what effect it bears on the scientific landscape

App2: research evaluation

Table 1 Mean Statistics for 1995 Benchmark SCs

Subject category	Sample size	Cited refs. (mean)	Times cited (mean)	Integration score (mean)	Diffusion score (mean)	Integration versus cited refs. (Pearson correlation)	Diffusion versus times cited (Pearson correlation)
Neuroscience	1,910	42.53	43.46	0.43	0.46	−0.05	0.04
Med-R&E	664	33.65	59.72	0.42	0.47	−0.07	0.10
Physics-AMC	1,017	33.40	32.52	0.40	0.38	−0.10	0.09
Biotech	840	31.23	27.37	0.37	0.44	−0.07	0.15
EE	1,719	18.40	13.51	0.35	0.37	0.24	0.14
Math	658	17.90	9.11	0.19	0.19	0.22	0.13
Total	6,808	30.43	30.54	0.37	0.40	0.20	0.13

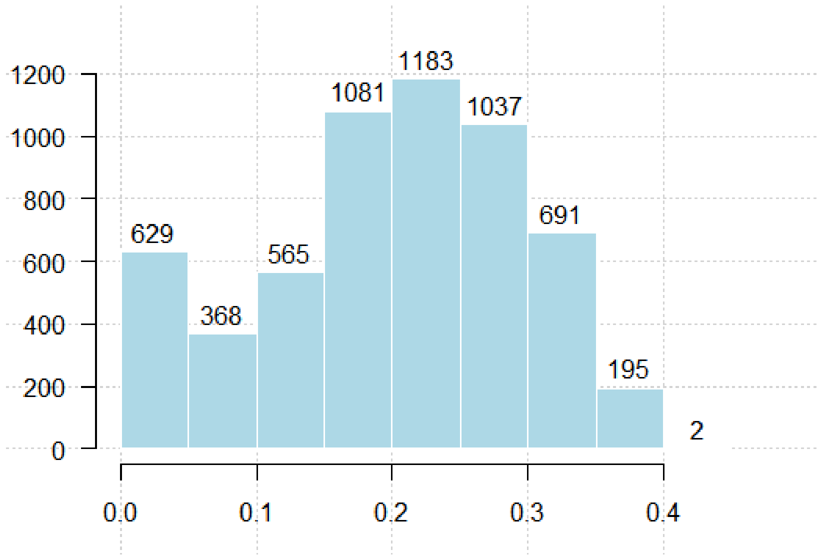
Carley, S., & Porter, A. L. (2012). A forward diversity index. *Scientometrics*, 90(2), 407–427.

App3: career and mobility studies

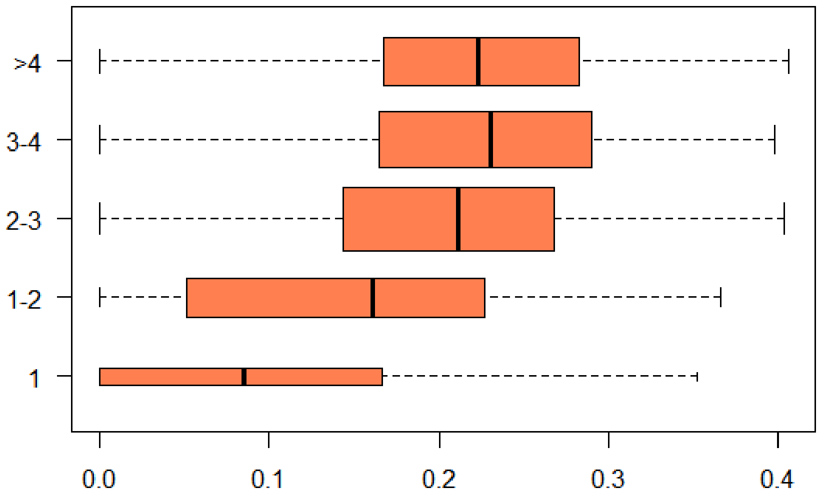
- Seldom addressed dimension of scientific careers and mobility: development of a research profile
- Important variable of econometric models on mobility:
  - Effect of profile dynamics on productivity or vice versa (generalist or specialist strategies)
  - Effect of various mobility dimensions on a research profile and vice versa
- SISOB (Science in Society Observatorium) program, FP7, *Mobility* use case
- The Stirling index as an aggregated/static measure of research profile development: thematic mobility for a large sample of engineers (SISOB case study) provided by SISOB partner Fondazione Rosselli (U Turin)

App3: career and mobility studies

Sample distribution of thematic mobility



Sample distribution by average number of coauthors



Science maps in quantitative assessments

- State-of-the art measures of scientific impact: field-normalized citation counts $\rightarrow$ **context sensitivity**

- **Background:**
  - Goal: comparing aggregates acting on different fields
  - The citation behavior of scholarly fields show large variation (citation densities, cf. mathematics vs. clinical medicine)
  - Solution: raw citation counts are corrected for field differences

Cnorm(**P**)=raw cit count (**P**) / expected cit count (**C, Y, T**)

Y= pubyear of **P**,

T=doctype of **P**,

C= Subject Category/Field of **P**

Rescaling citation distributions by research fields

- Rescaling cit. distributions by field average (Radicchi-Castellano, 2012, PLOS)

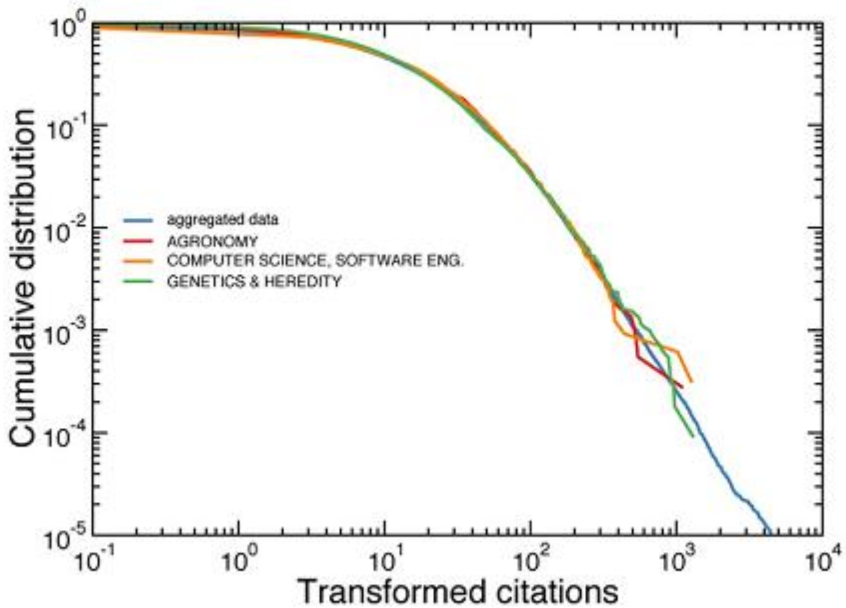


Figure 3. Cumulative distribution of the transformed citation counts. When raw citation numbers are transformed according to Eq. 2, the cumulative distributions of different subject-categories become very similar. All citation distributions are mapped on top of the cumulative distribution obtained by aggregating all subject-categories together (the common reference curve in the transformation). We consider here the same subject-categories as those considered in Figs. 1 and 2. The complete analysis of all subject-categories and years of publication is reported in the Supporting Information S2, S3, S4, S5, S6, and S7.

doi:10.1371/journal.pone.0033833.g003

Summary: Evaluative and structural

- Network perspective is inherent in professional scientometrics, which entertains rich SNA models not only on social networks.
- Network-based, structural measures reveal deep features of scientific performance and impact (diversification, inter-, and multidisciplinarity, scope and breadth of citation-based recognition or knowledge transfer etc).
- Network analytic methods are fundamental to establish reference sets for timely context-sensitive performance indicators.

Scientometrics as network science

- *Thank you for your attention!*