Relevant Monograph on IR

- Melucci, M., and Baeza-Yates, R. (Eds): "Advanced Topics in Information Retrieval". The Information Retrieval Series, Vol. 33, Springer-Verlag, Berlin Heidelberg, 2011.

ESSIR 2009

Bibliometrics/Scientometrics and IR
A methodological bridge through visualization

Peter Ingwersen
Royal School of LIS, DK
Oslo University College, NO

Agenda

- Introduction: *IR in Scientometrics*
- Basic power laws of common interest:
  - Lotka; Bradford; Zippf; Garfield
- Scientometrics/bibliometrics make use of document representations and relationships
- Co-occurrence of terms .. authors .. Citations – **visual science mapping**

Agenda 2

- Citations vs. Inlinks on the Web: similar?
- Scientific maps – online visualization
  - Salton's cosine and other similarity measures
- Closing Remarks

Introduction: *IR in Scientometrics*

- What is
- **Scientometrics**
- **Bibliometrics**
- **Webometrics**
- ….?

Intro 2: Definitions - Bibliometrics

- The study of quantitative aspects of the production, dissemination, and use of **recorded** information
- It develops mathematical models and measures for these processes and then use the models and measures for prediction and decision making

(Tague-Sutcliffe, 1992)

Intro 3: Definitions - Scientometrics

- The study of quantitative aspects of *science* as a discipline or economic activity
- Part of *sociology of science* and has application to science policy-making
- It involves quantitative studies of scientific activities, among others, publication, and so overlaps bibliometrics to some extent

(Tague-Sutcliffe, 1992)

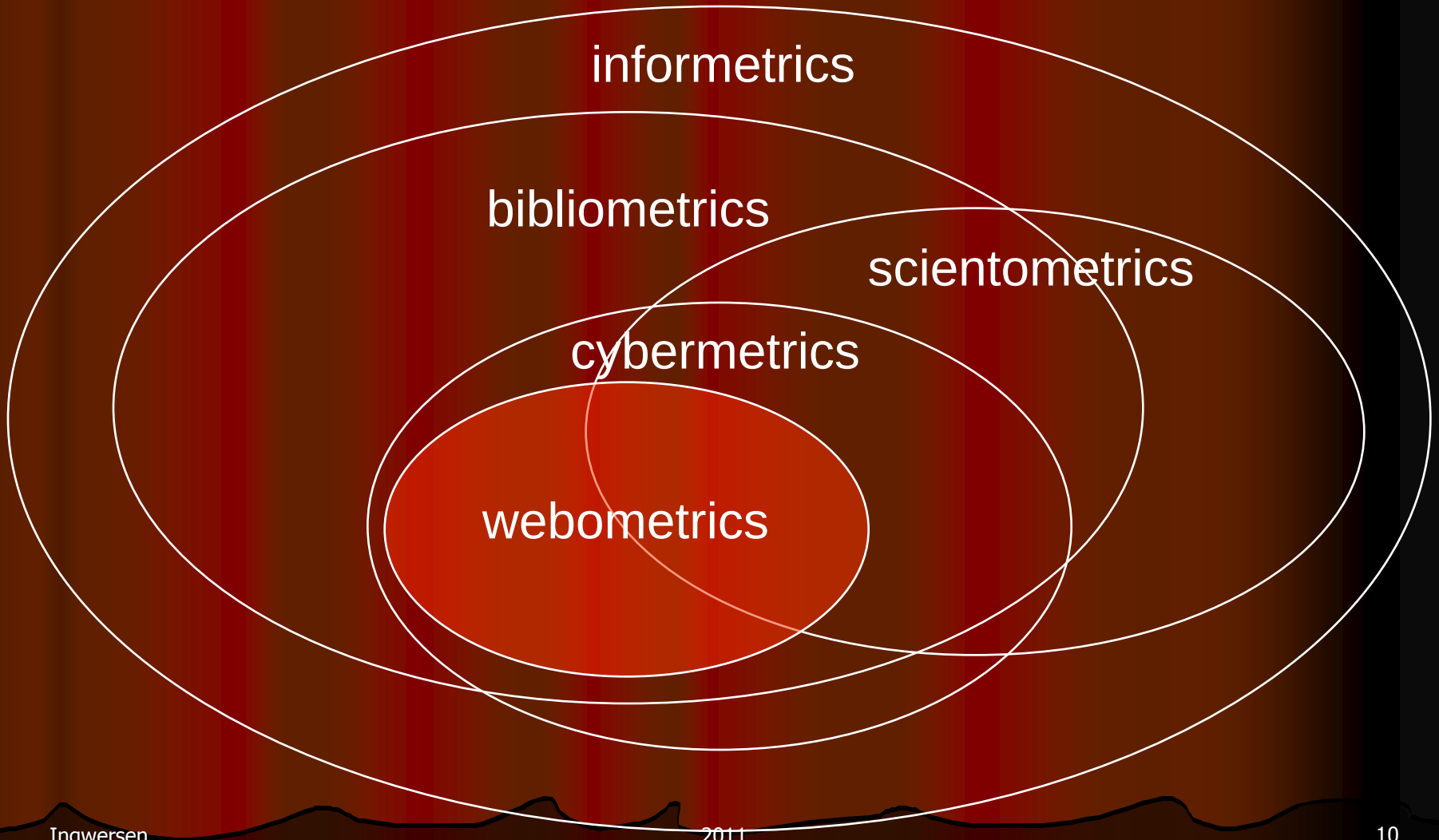
Webometrics

- The study of quantitative aspects of the construction and use of information resources, structures and technologies on the Web, drawing on bibliometric and informetric methods
  - web page contents
  - link structures, e.g., WIFs, cohesiveness of link topologies, etc.
  - users' information behaviour (searching, browsing, etc.)
  - search engine performance

Intro 4

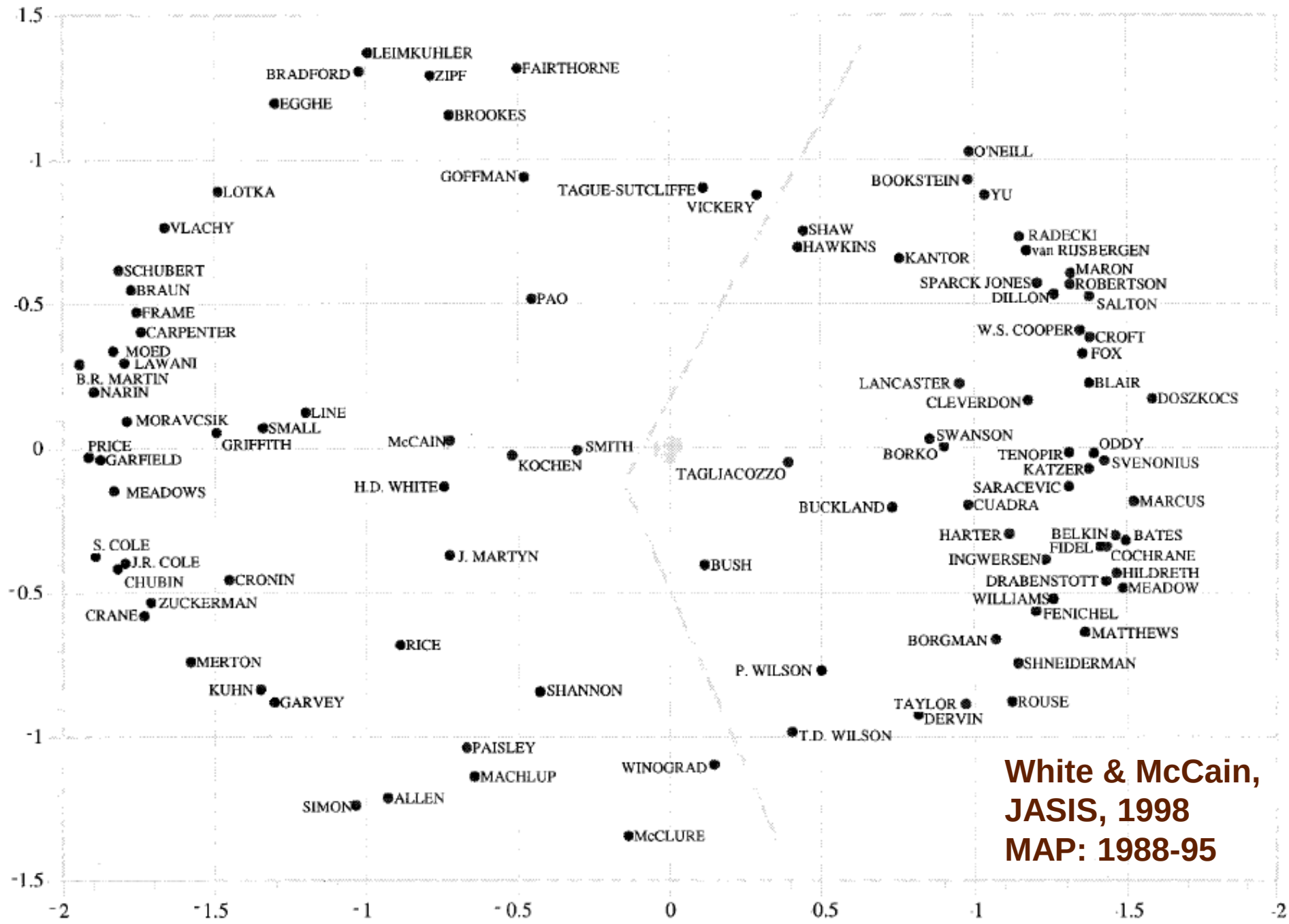
- **IR is necessary** for doing scientometric & bibliometric data collection & analyses
- Commonly domain databases (e.g. Medline; Inspec) or citation databases (Science Citation Index (SCI); Social SCI) or Web services (webometrics) are used as data sources online (or offline), combined with demographic data.
- Boolean logic has been preferred (exactness)

infor-/biblio-/sciento-/cyber-/webo-/metrics



Definitions – the classic … metrics

- Bibliometrics/Scientometrics/Webometrics **are unique types of empirical research methods developed by Library and Information Science** and **Sociology of Science**;
- Informetrics utilizes quantitative analysis, statistics, and data visualization to investigate patterns of:
  - References, citations, authors, journals, institutions, words, keywords, classification codes etc.



White & McCain, JASIS, 1998
MAP: 1988-95

Three Bibliometric power "laws"

- Lotka's law (1926): www.lis.uiuc.edu/~jdownie/biblio/lotka.html
  - Productivity of authors (researchers); how many researchers have written 1, 2, 3… articles?
  - Frequency distribution (probability)
- Bradford's law (1934): www.lis.uiuc.edu/~jdownie/biblio/bradford.html
  - Scatter of the articles over journals in given subject (graphic)
  - *Equal-production groups* (verbal version)
- Zipf's lov (1949): www.lis.uiuc.edu/~jdownie/biblio/zipf.html
  - The frequency of words in text
  - Rank/Frequency distribution

The Basic Power laws

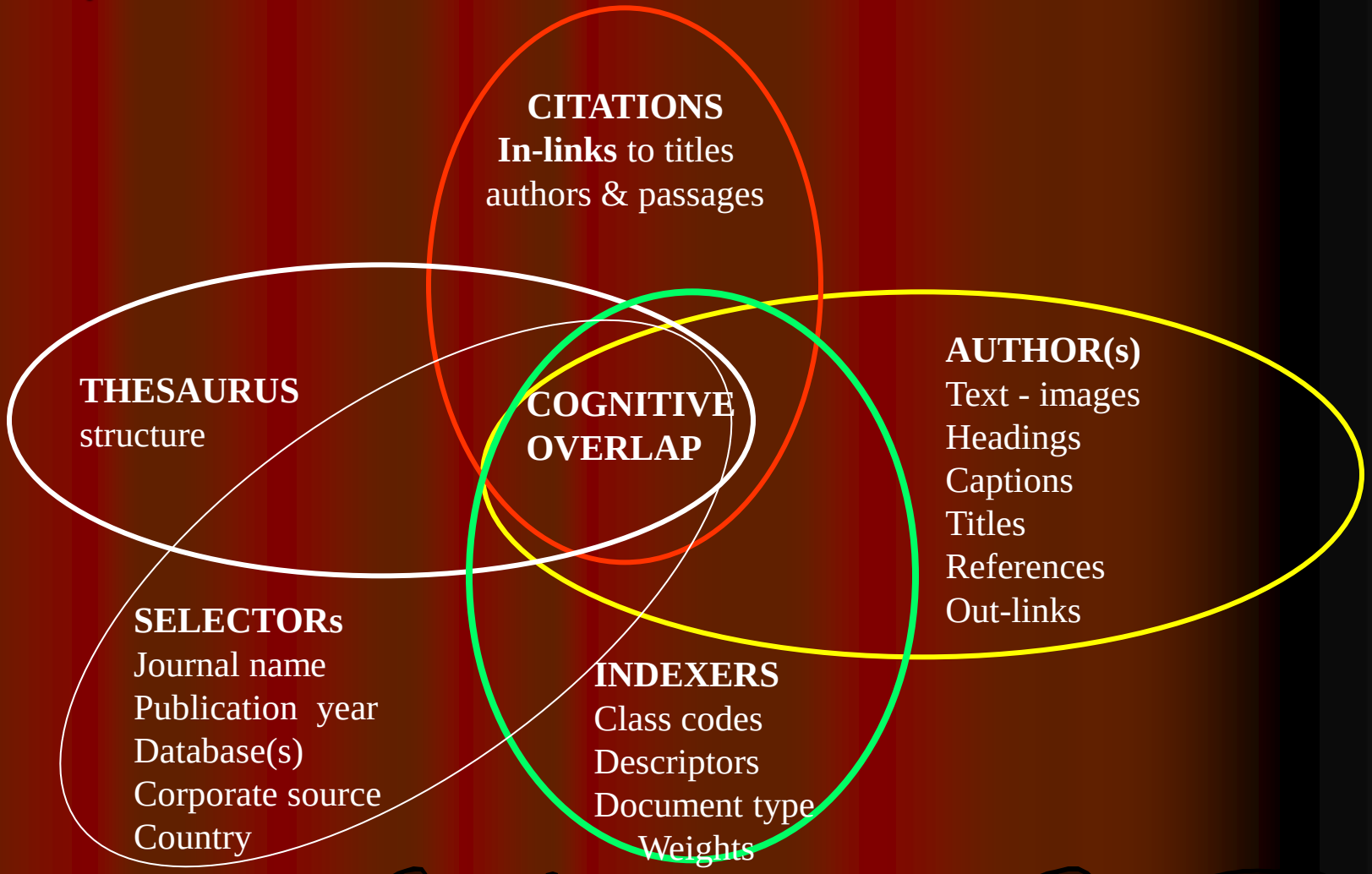
One bridge between IR and Informetrics

- Rank distributions that are useful for:
- Research evaluation
- *H*-index calculations (top-researchers)
- Collection development (e.g. which journals not to buy!)
- Authoritative institutions
- Weighting purposes
- Background for science mapping

Document representations: another common bridge

- Media and genres consist of documents having different kinds of representations available:
- **Contents related representations** (e.g. terms, music frequencies, figure color & shapes …..) – aboutness;
- Production related representations (e.g. publisher, authors, publication dates ….) - isness

cognitively & **typologically** different representations in scholarly documents



Representations and documents

- All kinds of representations – added or inherent – may point to documents
- We may thus structure (cluster; map; …) documents by means of representations and their relationships
- OR the opposite:
- Structure representations by means of documents containing such keys
- This is done by means of co-occurrence

Co-occurrence applications in Scientometrics

- Data mining in databases (like IR)
- **Creation of visual maps of scientific domains**
- Demonstrate relationships between:
  - People – author co-citation; collaboration
  - Countries – collaboration
  - Institutions – collaboration; co-citation
  - Topical areas
  - Journals – co-citation – coupling
  - …..

Most important co-occurrence applications

- By bibliographic coupling: **people are connected by using the same work!!**
- By author co-citations: **the perceived connexion between people** (through bibliographic references in documents):
- *They share expertice*
  - Check consistency by ageing measures of their work (normalization?)

References vs. Citations

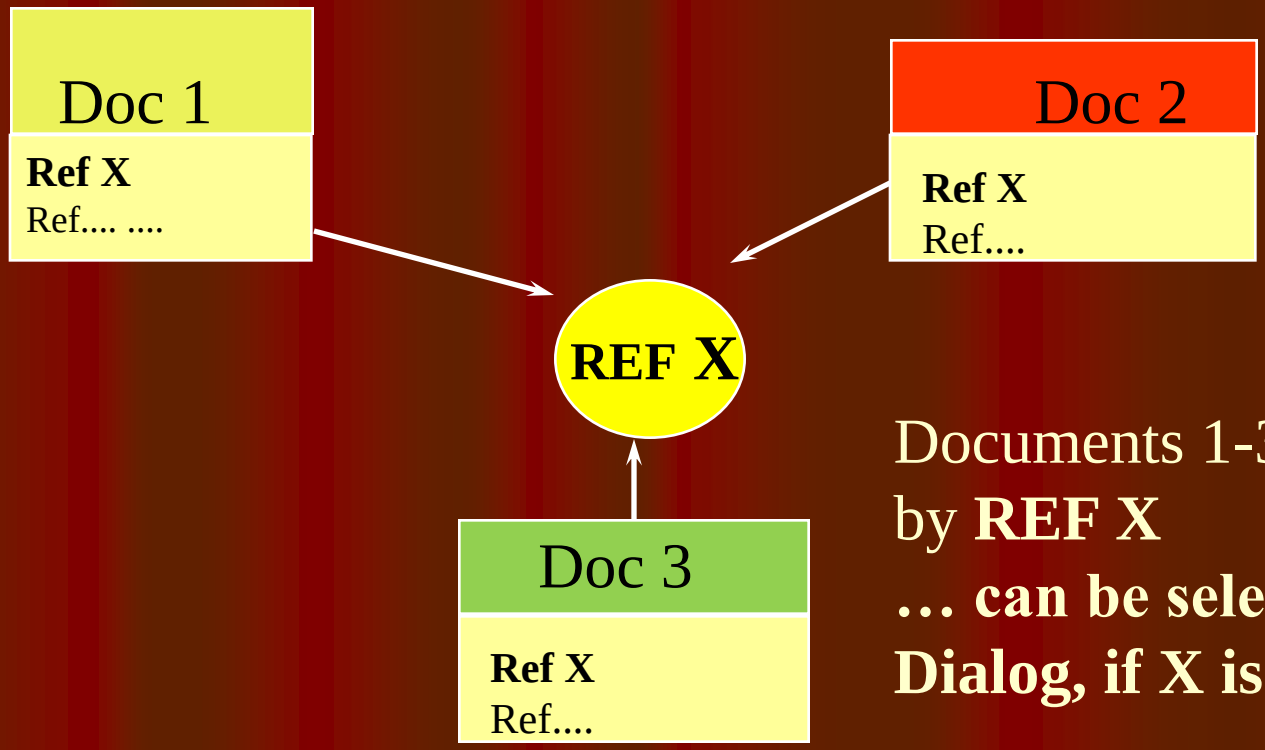
- Academic papers have list of **REFERENCES**
- When each reference is isolated and grouped together with respect to the same cited:
  - Author
  - Journal
  - Institution or country
  - Topic (term; concept; class)
- **they turn into CITATIONS!**

References as representation

- An academic reference is a 'payment' but also a kind of author-generated **additional subject/content key** to the citing doc. (this was the original idea by Eugene Garfield behind the citation indexes in 60s)
- Re-used in the IR principle of POLYREPRESENTATION (Ingwersen, 2005)

BIBLIOGRAPHIC COUPLING

- Can also be done on the Web or other social nets with link conventions!



Documents 1-3 are coupled by **REF X**

… can be selected in WoS/ Dialog, if X is known

Example of bibliographic coupling

- One article has 23 references on its reference list
- Another has 54 references on its list
- There are 4 references in common.

Strength: 4 / (23 + 54 - 4) = 0.054 (max.=1)

- **Jaccard similarity measure**

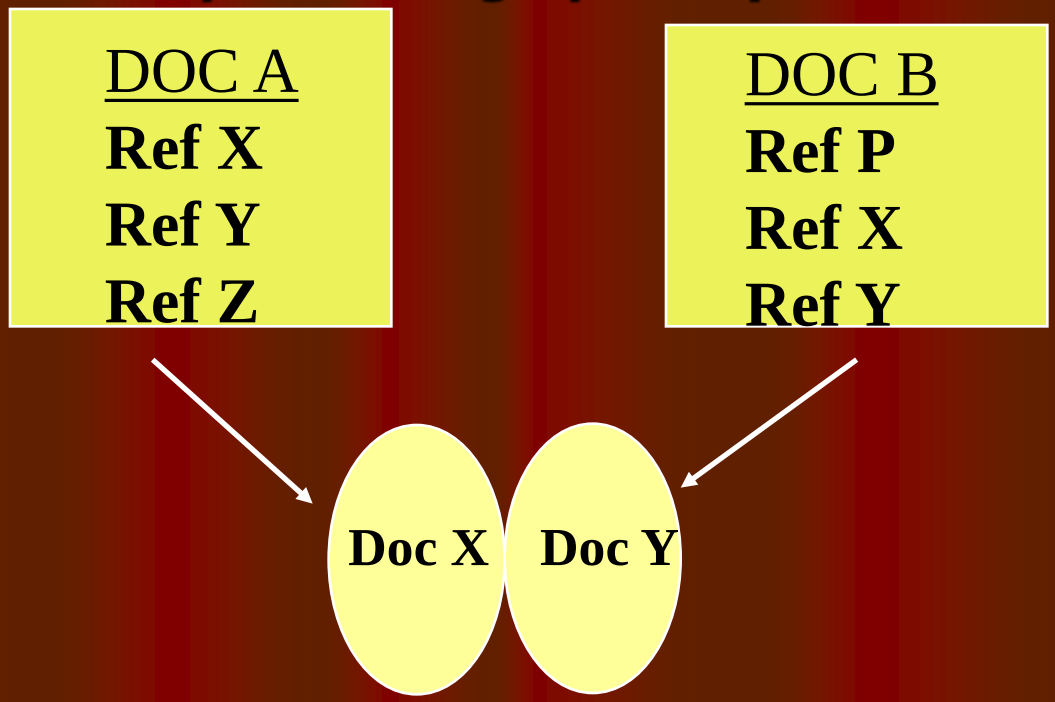
- **CA – CY – CW or cited journals could also be used**

CO-CITATION

Documents X and Y are CO-CITED Twice

on the reference lists of A and B

(Doc. A is coupled bibliographically to Doc. B by X & Y)



Co-citation illustration: KNOWN CITATIONS – known cited document

Set	Items	Description
S4	7	CA=JOSS PC(S)CY=1988(S)CW=NATURE
S8	13	CA=FABIAN AC(S)CY=1987(S)CW=NATURE
S9	6	S4 AND S8

COSINE: SIM = 6 / ($7^{1/2}$ * $13^{1/2}$) = 0,63

JACCARD: SIM = 6 / (7 + 13 - 6) = 0,43

These are BINARY calculations (document level)

Co-citation illustration: KNOWN AUTHORSHIPS – citing document level

S1	279	CA=BELKIN NJ
S2	86	CA=INGWERSEN P
S3	955	CA=SALTON G
S4	49	S1 AND S2
S5	70	S1 AND S3
S6	18	S2 AND S3

AUTHOR CO-CITATION SIMILARITY - 1990:

One creates an overlap matrix

	NJ Belkin 279	P Ingwersen 86	G Salton 955
NJ Belkin 279		49	70
P Ingwersen 86			18
G Salton 955			

Co-citation illustration – **Similarity matrix**: KNOWN AUTHORSHIPS – Binary, citing document level – Jaccard & Cosine

SIM(BELK/INGW): 49 / (279+86-49)= **.16**

Cosine: 49 /16.7 x 9.3 = .32

SIM(BELK/SAL): 70 / (279+955-70)=**.06**

Cosine: 70 /16.7 x 30.9 = .14

SIM(SAL/INGW): 18 / (955+86-18)=**.02**

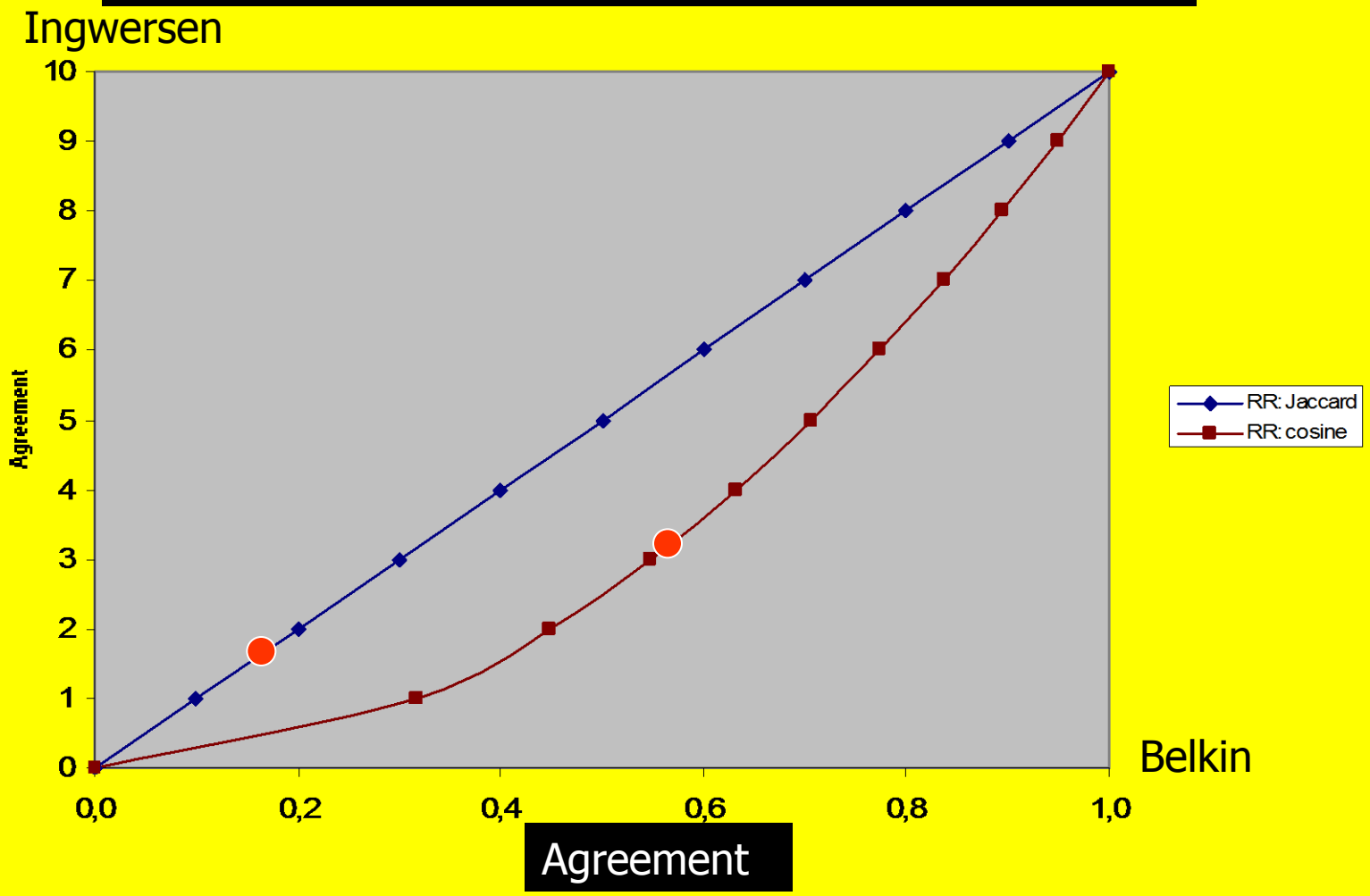
Cosine: 18 /30.9 x 9.3 = .06

One creates a similarity matrix

	NJ Belkin 279	P Ingwersen 86	G Salton 955
NJ Belkin 279		.16	.06
P Ingwersen 86			.02
G Salton 955			

At Item Level: Cosine and Jaccard OK:

Similarity in a BINARY CONTEXT



THIS WAS IN 1990 - WHAT IN 2000??

	Items	Citations	Name
S1	541	933	CA=BELKIN NJ
S2	258	382	CA=INGWERSEN P
S3	1365	2417	CA=SALTON G
S4	126	559	S1 AND S2
S5	175	680	S1 AND S3
S6	50	204	S2 AND S3

Co-citation illustration: KNOWN AUTHORSHIPS

Binary, citing document level; Jaccard & Cosine

SIM(BELKIN/INGWERSEN):

126 / (541+258-126) = **.19**

Cosine: 126 / 23.3 x 16 = **.34**

SIM(BELKIN/SALTON):

175 / (541+1365-175) = **.10**

Cosine: 175 / 23.3 x 36.9 = **.20**

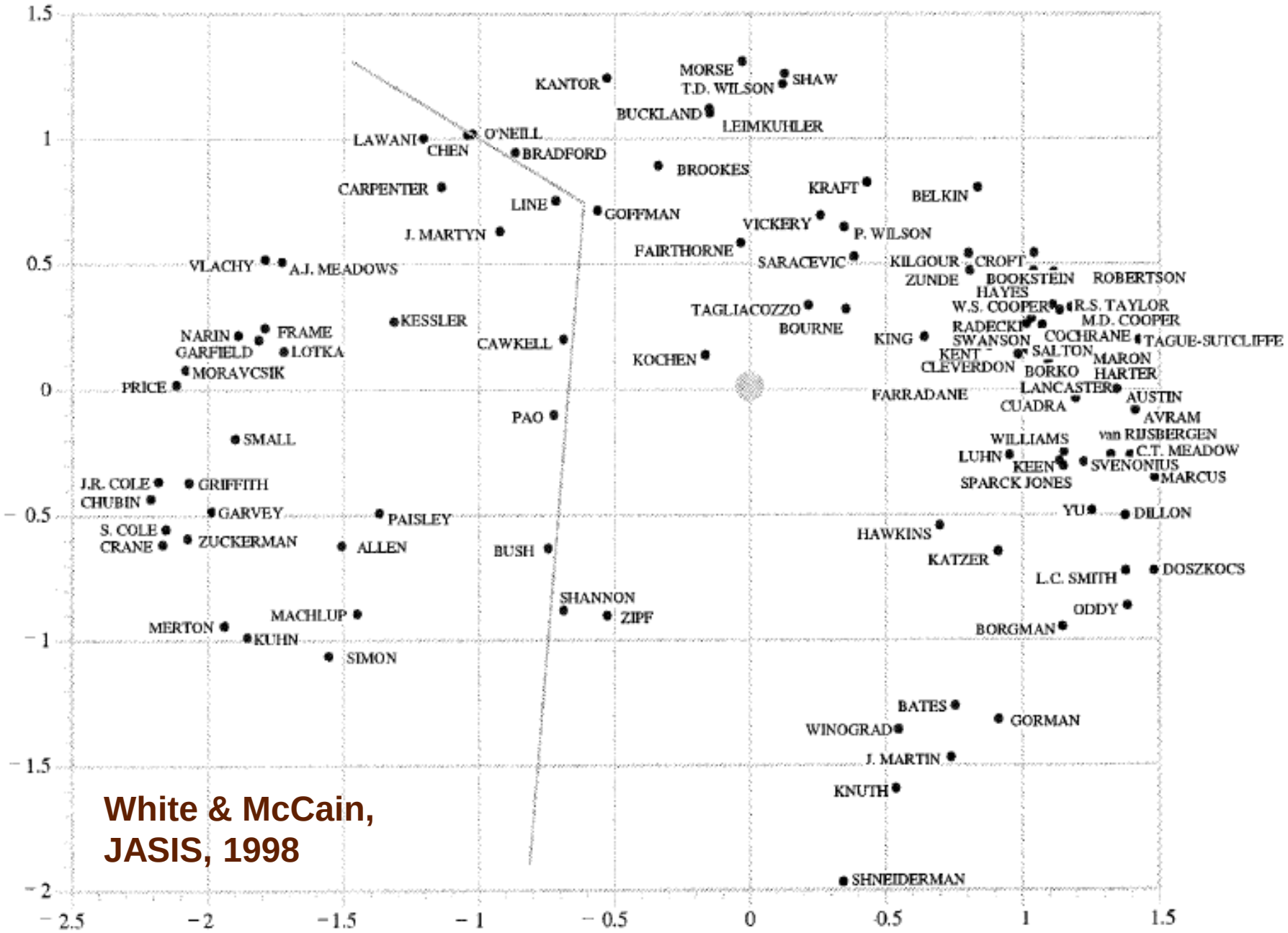
SIM(SALTON/INGWERSEN):

50 / (1365+258-50) = **.03**

Cosine: 50 / 36.9 x 16 = **.08**

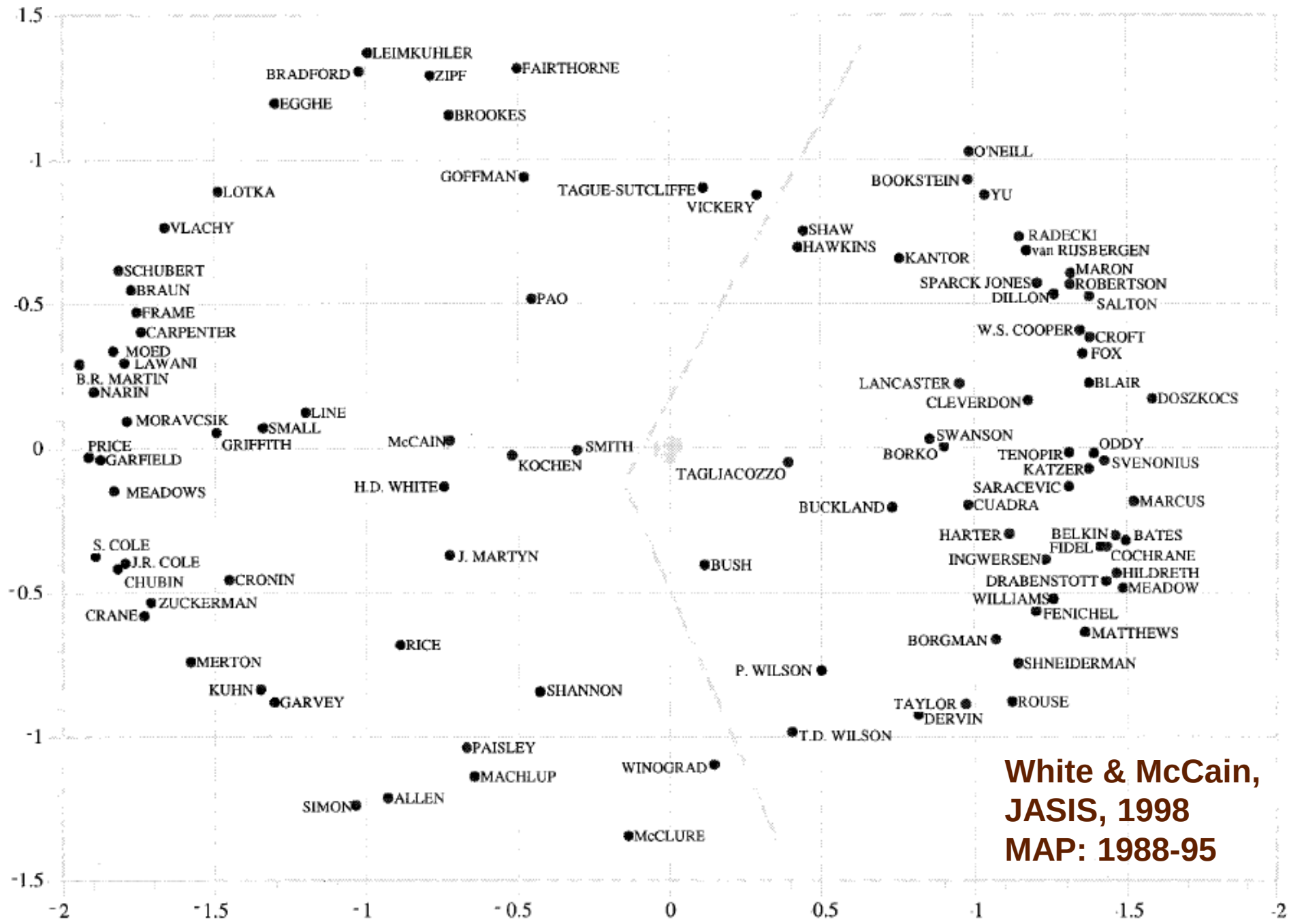
The trend 1978 - 2000

- The Belkin-Salton co-citation ratio has **increased from 0.06 to 0.10 in the 90s**
- The Belkin-Ingwersen co-citation ratio has **slightly increased, from 0.16 to 0.19**
  - **Both pairs are thus *closer* to one another seen from colleagues' views!!**
- 50 % of documents citing Ingwersen (126/258) co-cites him with Belkin after 1990. Up to 1990 this ratio was larger (49/86) = 57 %.



White & McCain, JASIS, 1998

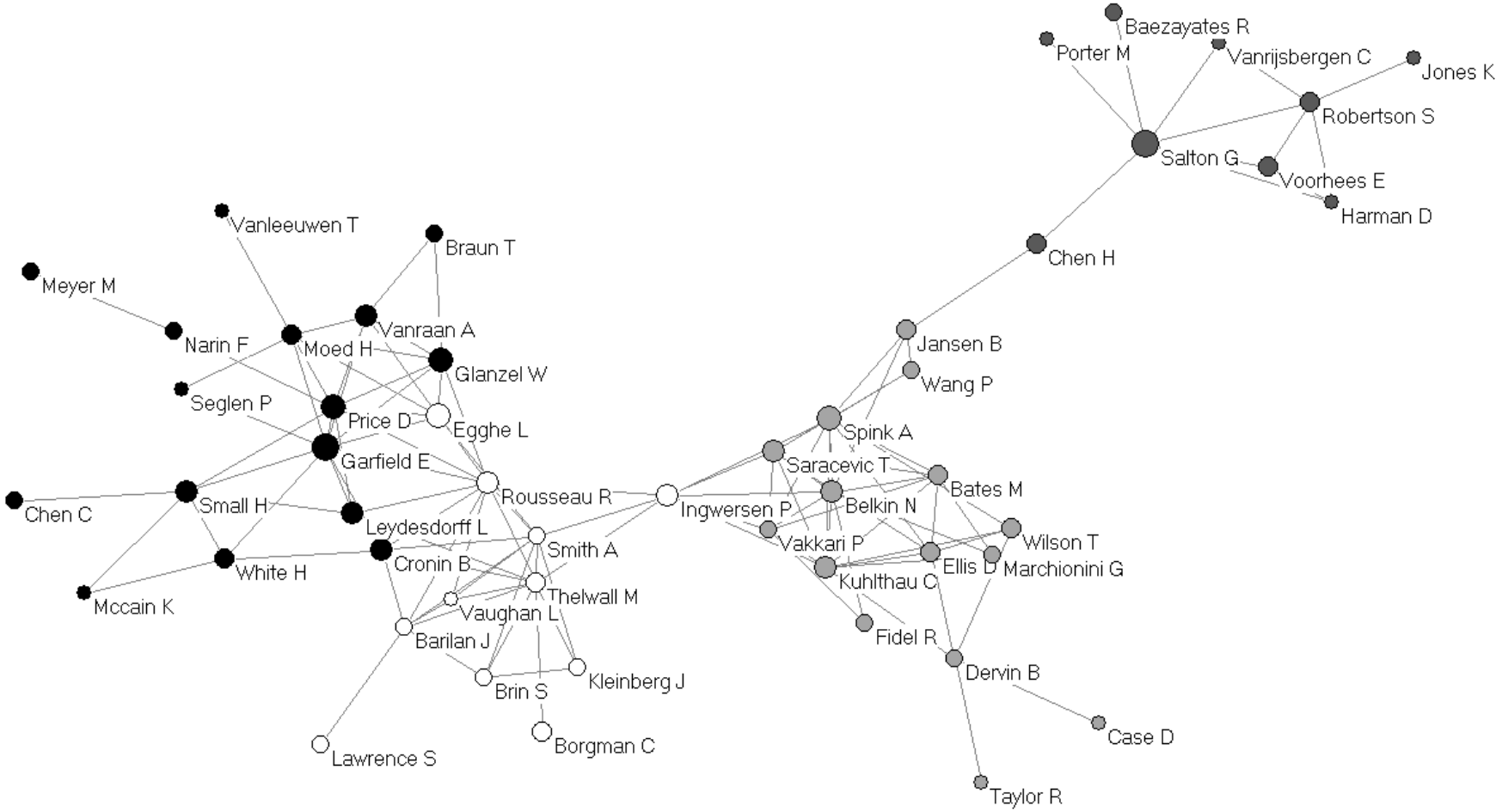
FIG. 2. Top 100 authors in information science, 1972–1979.



White & McCain, JASIS, 1998
MAP: 1988-95

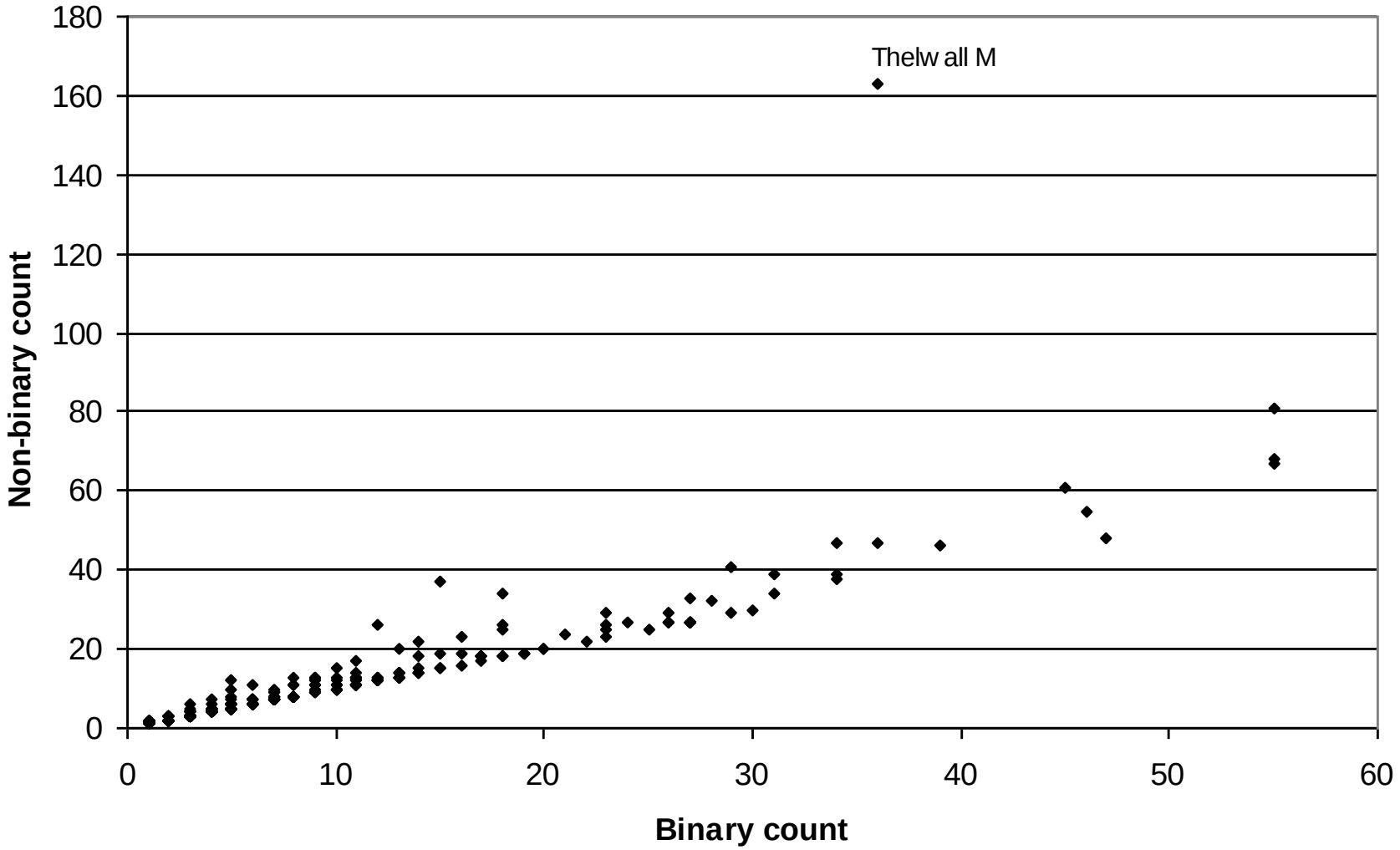
Underlying data

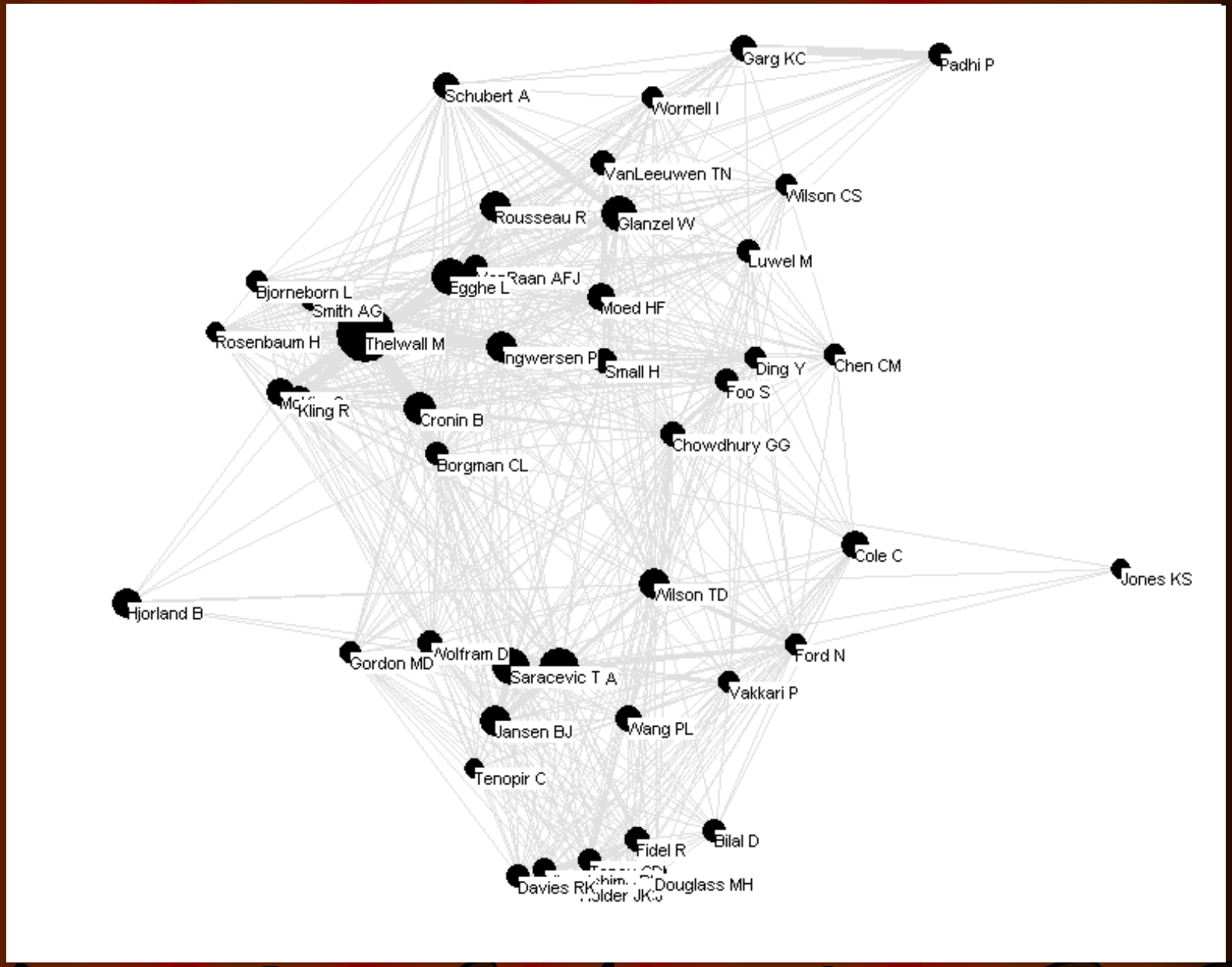
- 12 LIS/IR journals
- Top-100 most cited authors
- Author-co-citation matrices
- Similarity: Pearson correlation Coeff.
- NW: Scientometrics/bibliometrics
- SW & S: Sociology of Science – Philosophy
- N: Meta-LIS; NE: Hard core IR;
- SE: IR interaction; centre: IR & Bibliometrics



2003-2007 – 63 authors; 2542 LIS articles (=White/McCain, 1998). Persson, 2008

Most cited authors



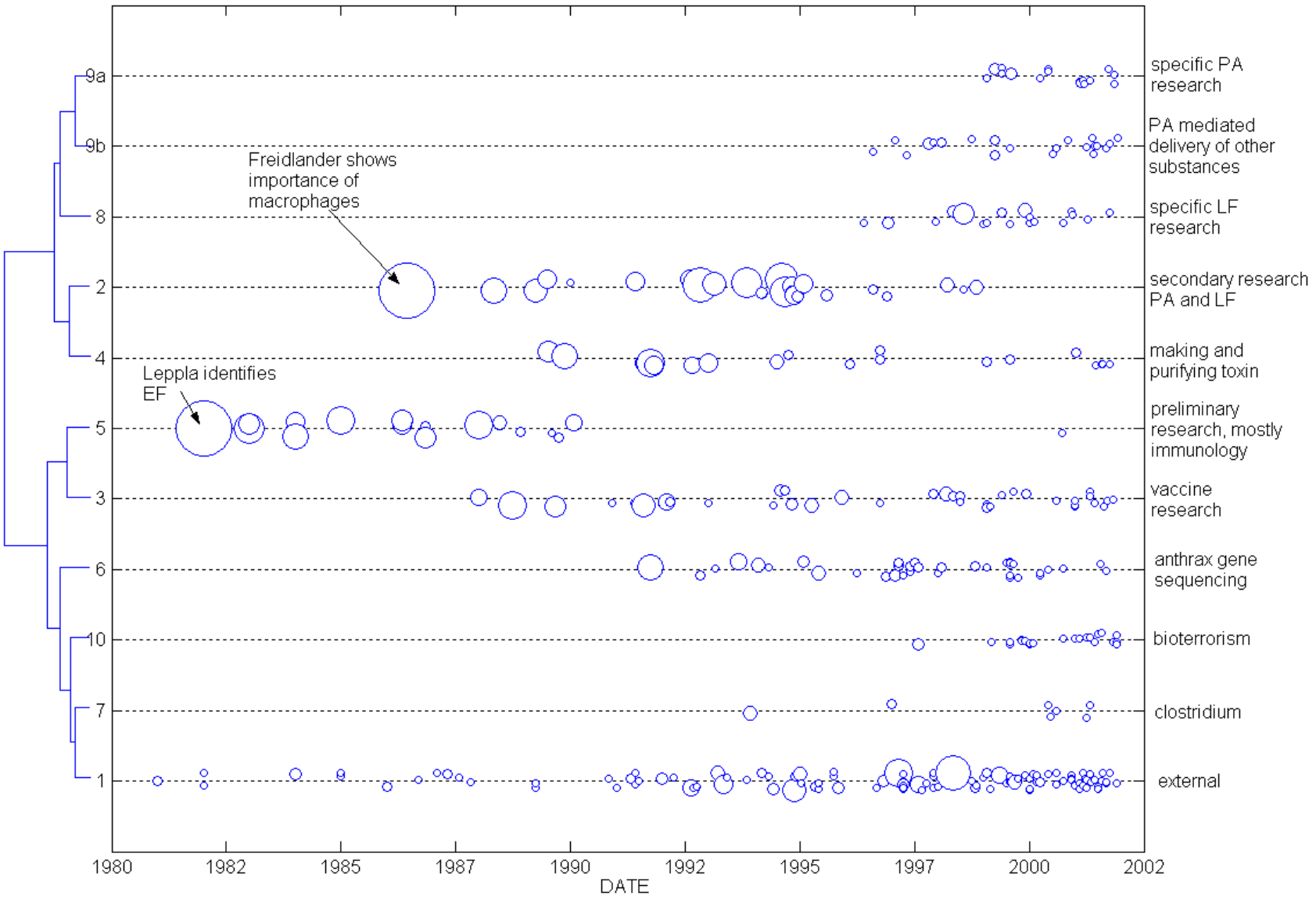


Non-binary integer count

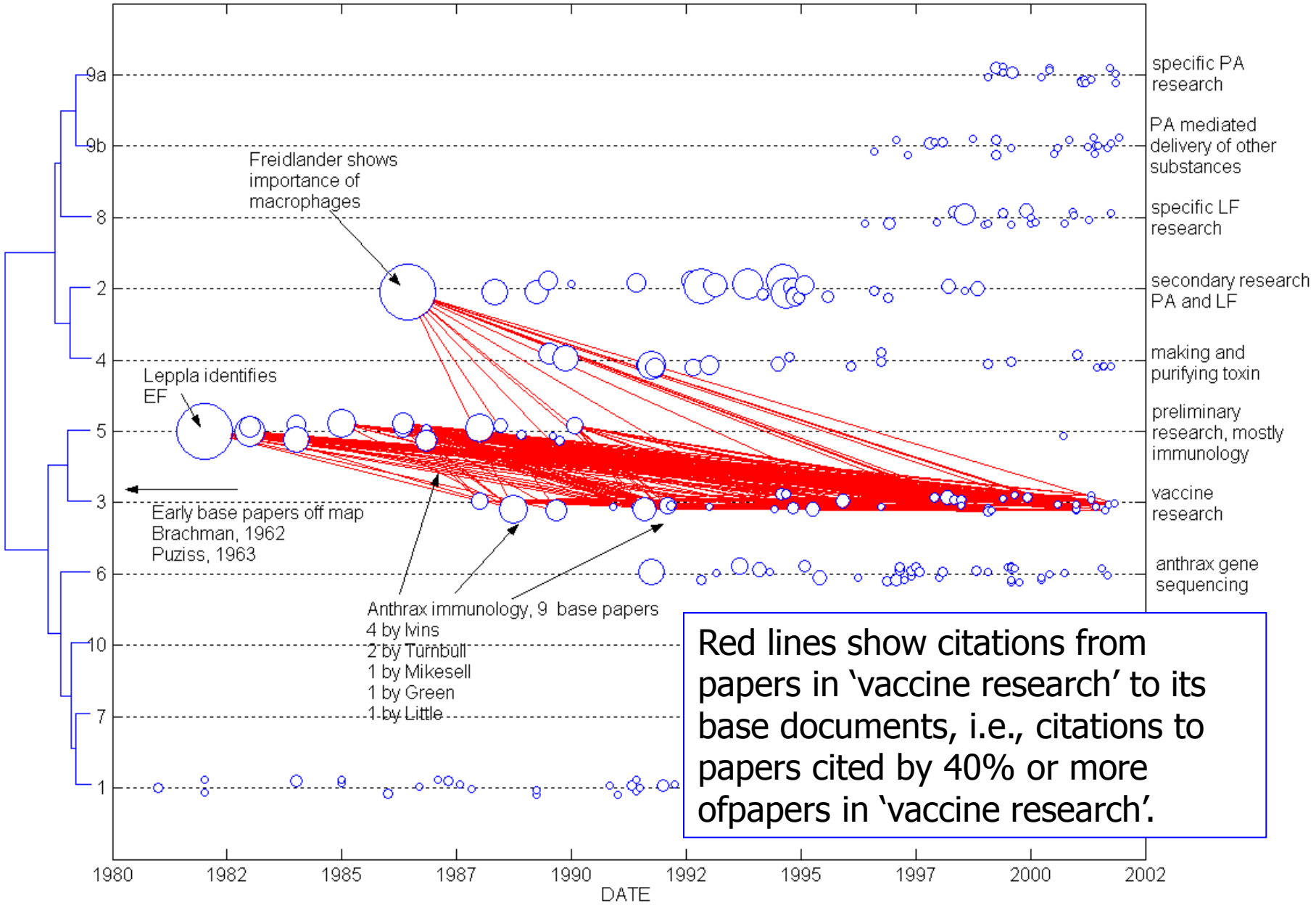
Online visualization

Where Information Retrieval and Scientometrics meet!

Anthrax research

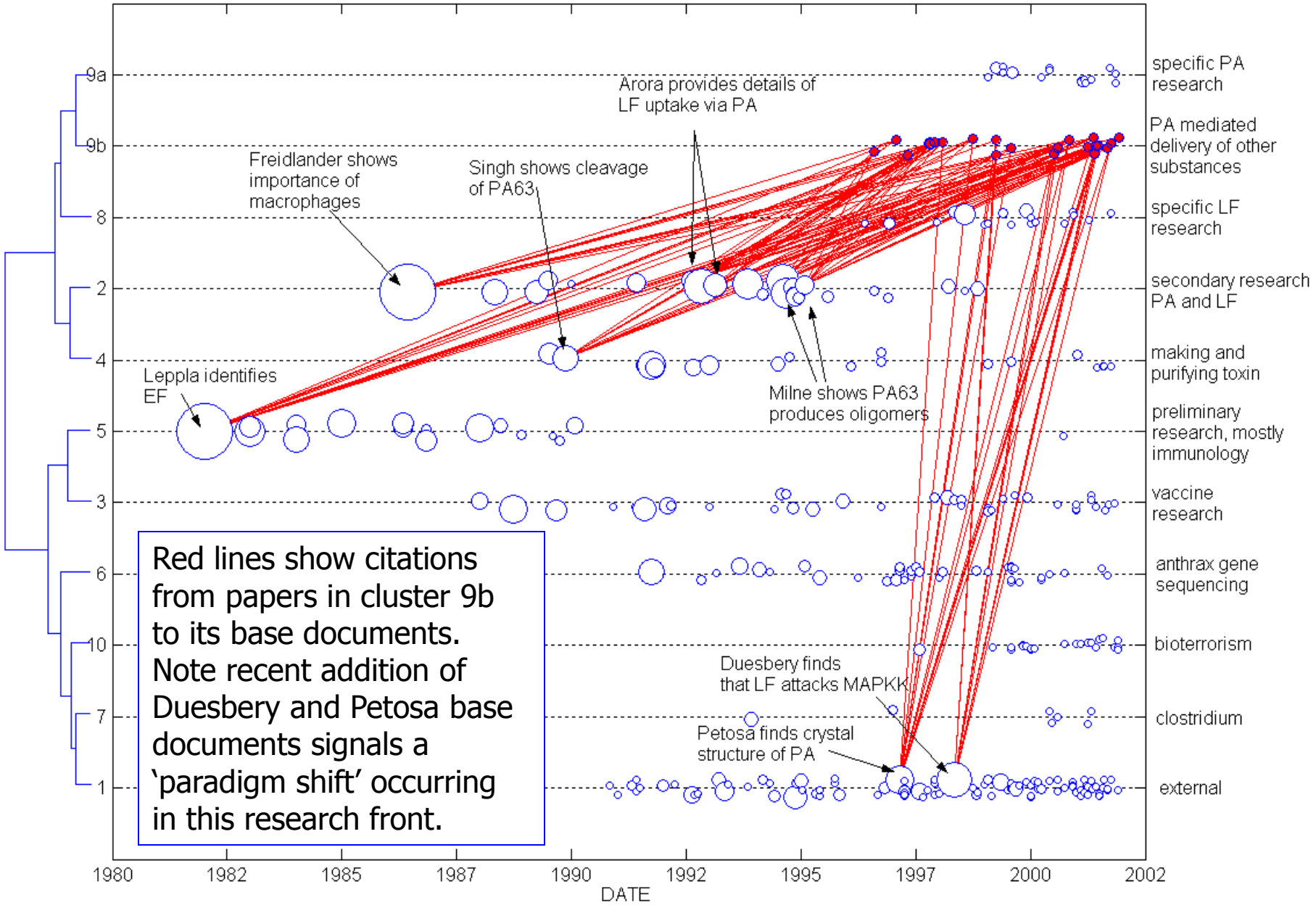


Anthrax research: base documents for vaccine research



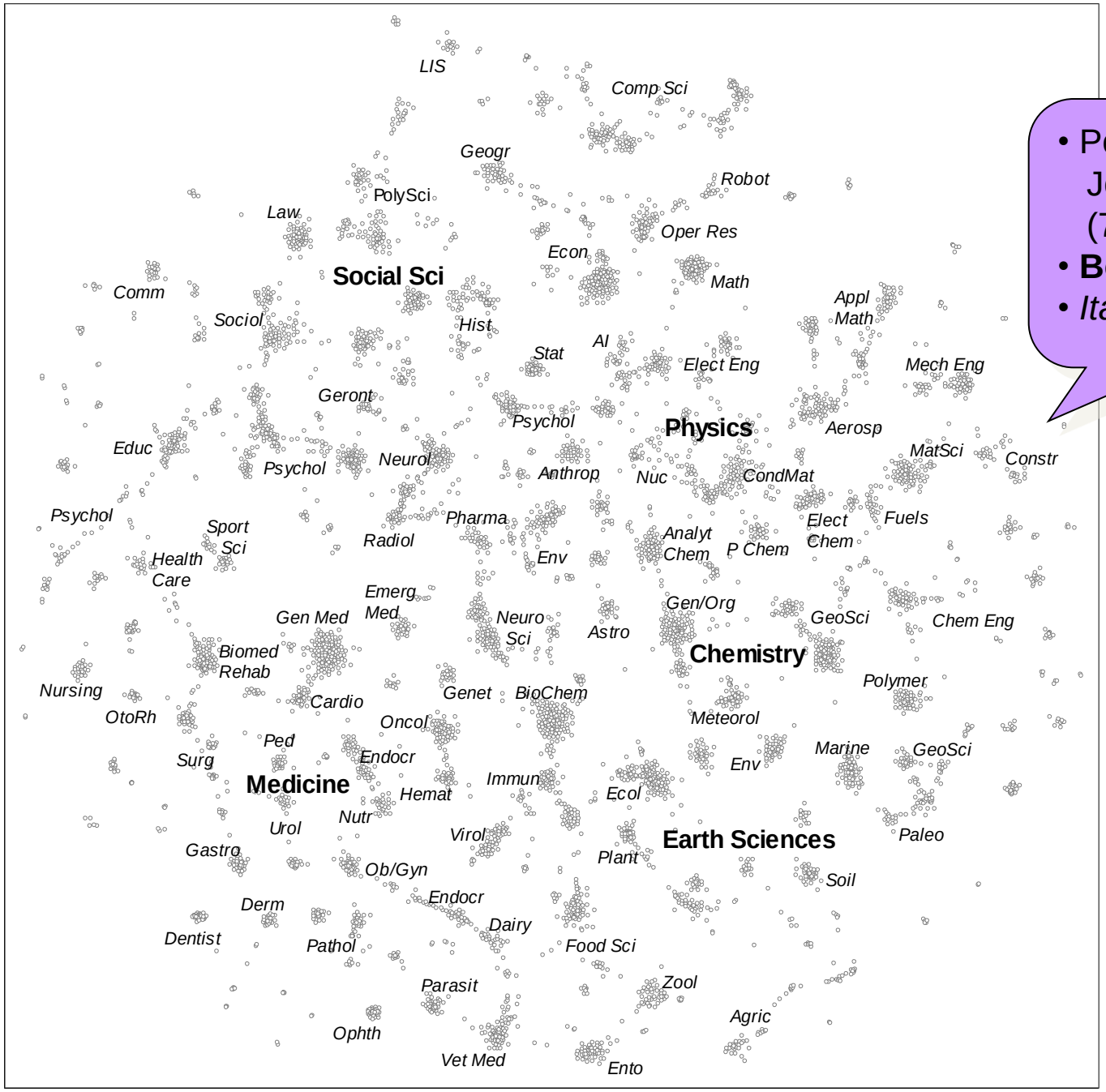
Red lines show citations from papers in 'vaccine research' to its base documents, i.e., citations to papers cited by 40% or more ofpapers in 'vaccine research'.

Anthrax research: base documents



Red lines show citations from papers in cluster 9b to its base documents. Note recent addition of Duesbery and Petosa base documents signals a 'paradigm shift' occurring in this research front.

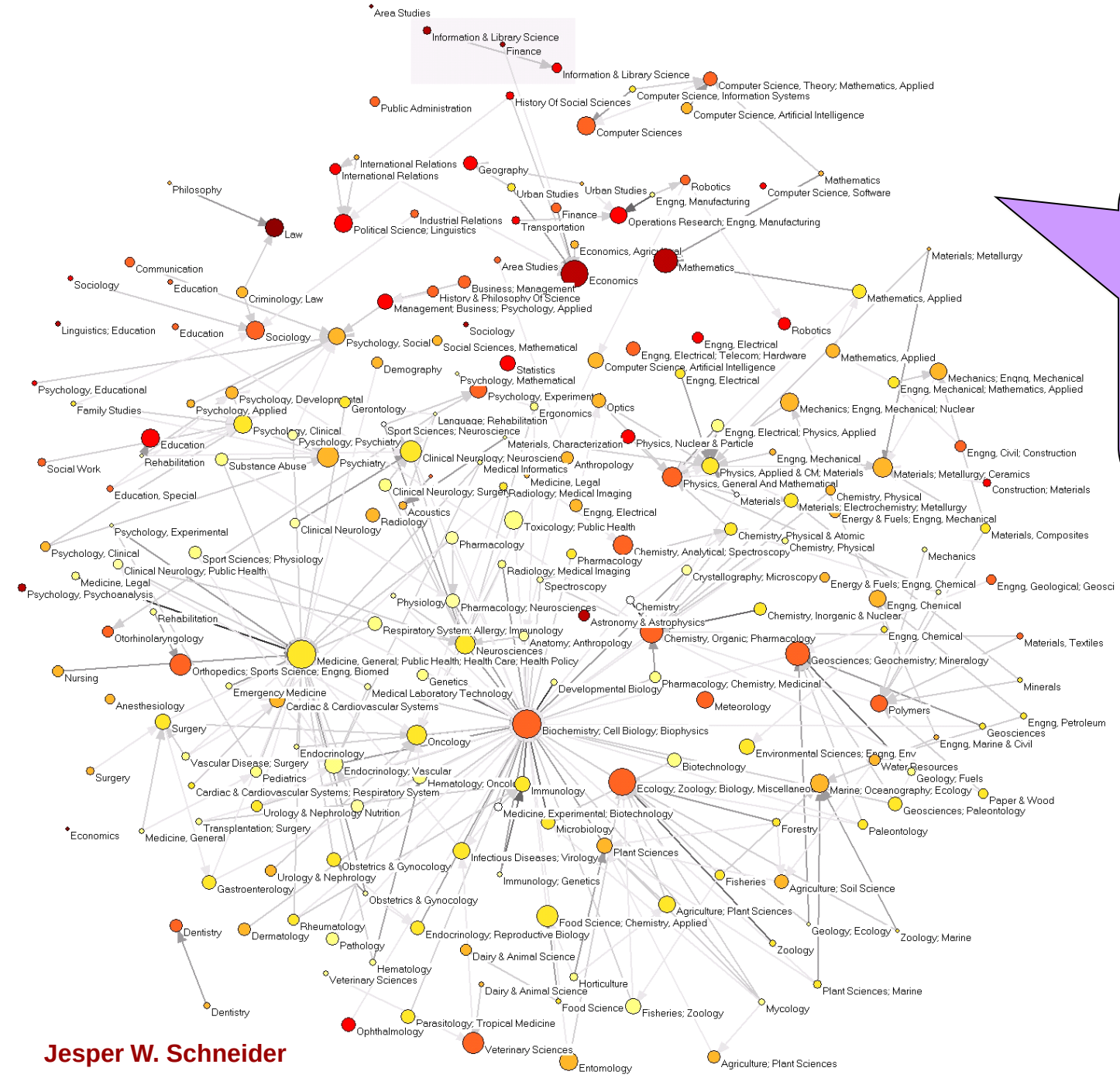
Academic fields



- Points = Journals (7121)
- **Bold** = Sciences
- *Italics* = Discipl.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. Scientometrics, 64(3), 351-374.

Jesper W. Schneider

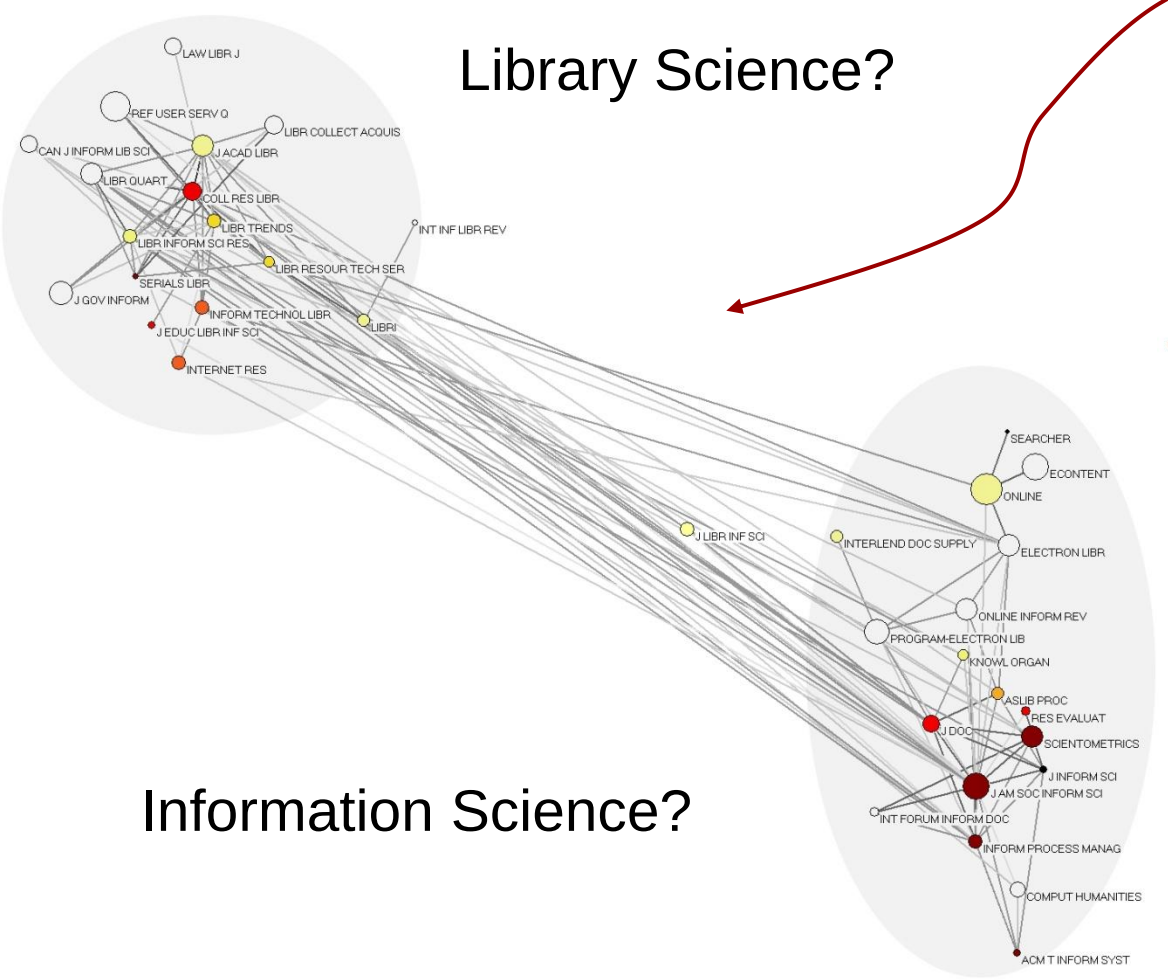


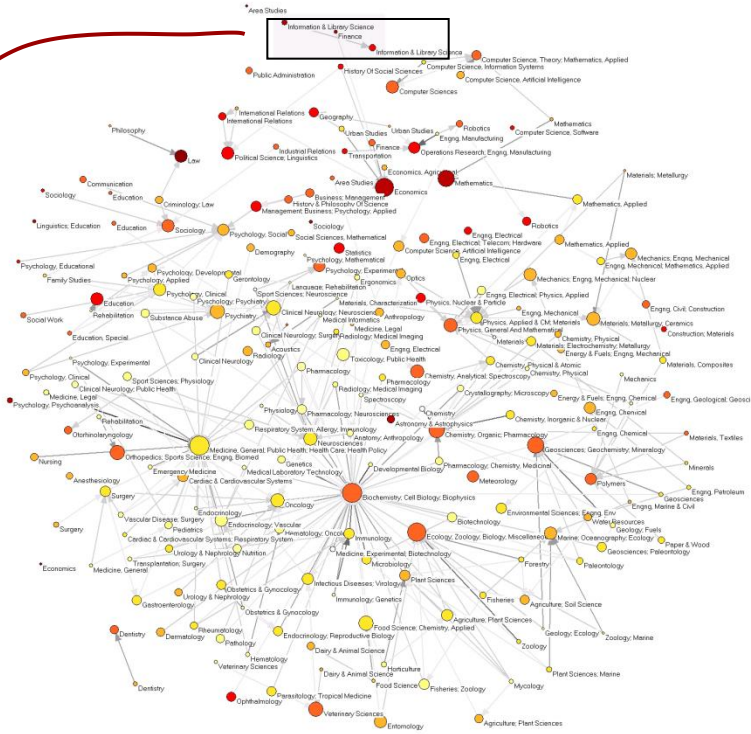
- 212 disciplines
- Circles = clusters
- Size = Journals #
- Color ="independence"
- Lines = strongest relations (k_{cit} > 7.5%)

Jesper W. Schneider

Zoom on LIS

Library Science?





Information Science?

Jesper W. Schneider

Citations vs. Inlinks on Web

- Academic citations follow tacit conventions:
- References are given in ‘normative ways’ (Merton) or from ‘rhetoric perspectives’ (Latour)
- That is the reason that we can use them in impact analyses (quality) and visualization of sciences; they presuppose:
- Academic publications represent research
- References / Citations imply recognition/impact
- Publications remain in space (reproducibility)

Links

- Many different reasons for linking
- Links can be empty (10 %)
- Web publications are modified / deleted / reborn / fused ….
- Link impact corresponds mostly to visibility on the web (volume of web publications)
- Some very weak correlation with RAE

References & Citations vs. relevance

- References have been shown to improve IR (e.g. in polyrepresentation: Larsen et al., JASIST, 2009) as additional keys to the contents
- Citations are not necessarily good markers of relevance, because impact and relevance might not always be overlapping phenomena

(In)Links and Citations

- Links and citation impact are not simply corresponding to one another
- Original Kleinberg or PageRank algorithms (1998) are based on the false assumptions that 1) links are similar to citations and 2) providing a kind of 'relevance' perspective to searching.

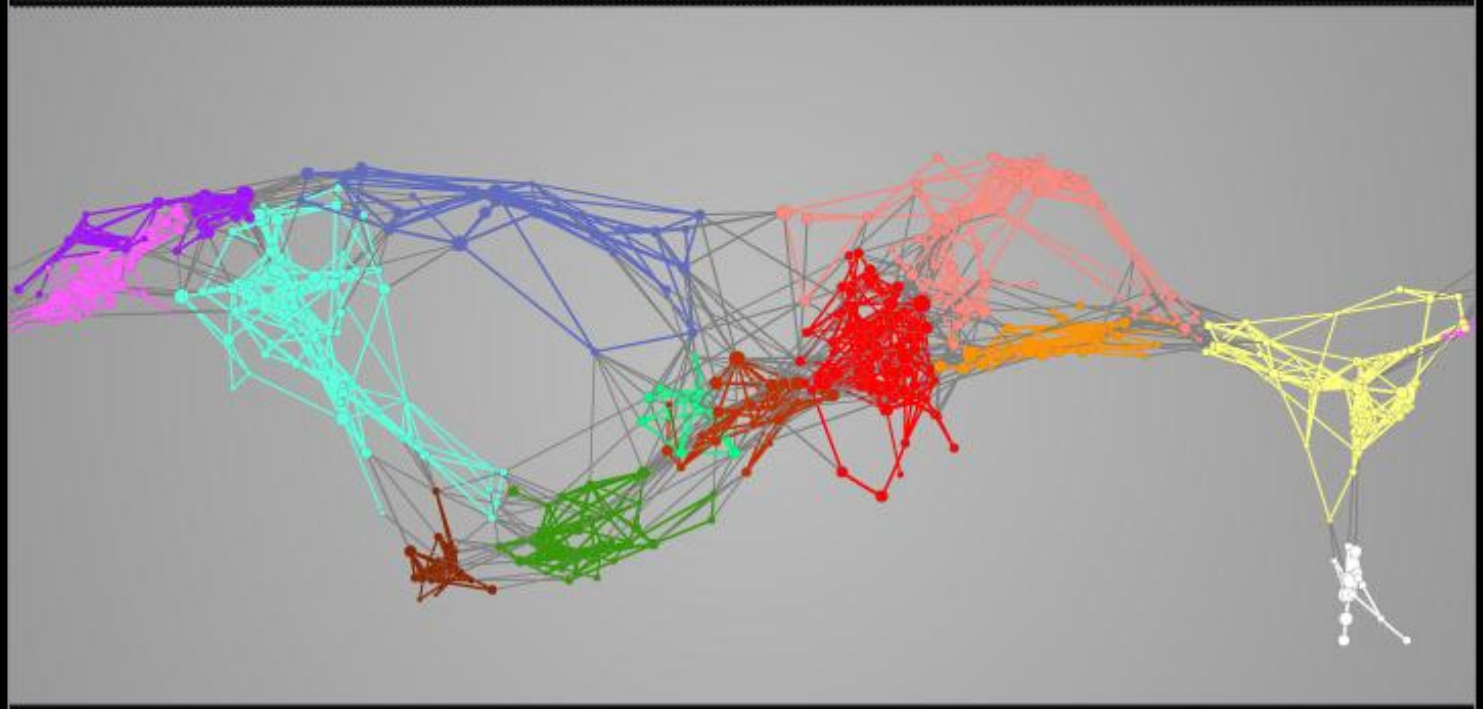
(In)Links and Citations

- Rather, HIT and PageRank provide kinds of (graph theoretical) impact analyses where the strength of citation chains (relationships) are cumulated to result in a final score.
- Can be applied to citation networks too.

MAPS OF SCIENCE

Math & Physics
Chemistry
Computer Science & EE
Other Engineering
Biotechnology
Earth Sciences
Biology
Infectious Diseases
Medical Specialties
Brain Research
Health Professionals
Social Sciences
Humanities
All

Scientific Disciplines



Concluding remarks

Scientometrics and IR (& visualization) have a lot to gain from collaboration

as done by Bookstein; Swanson; Tague Suttcliffe; Salton & Ingwersen in the good old times!

Interesting stuff

- Kate Börner
- Boyack & Klavens
- Chen CM (Cite space)
- Loet Leydesdorf
- Jesper W. Schneider
- Chowdhury GG (real time IR visualization)
- Ying Ding (real time IR visualization)