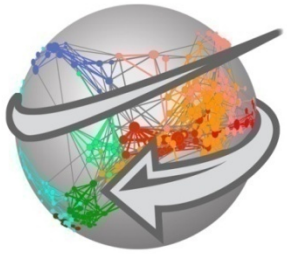
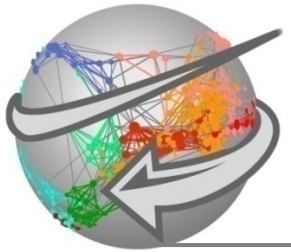


The Accuracy of Network Visualizations

Kevin W. Boyack

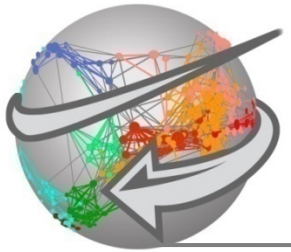
SciTech Strategies, Inc.

kboyack@mapofscience.com



Overview

- Science mapping history
  - Conceptual Mapping
  - Early Bibliometric Maps
  - Recent Bibliometric Maps
  - Summary of Improvements
- Progress in Measuring Accuracy
  - Accuracy of Relatedness Measures
  - Accuracy of Maps
- Frontier issues
  - How to increase accuracy further?



Science Mapping History

1939 –1970: Conceptual maps

- Done by hand

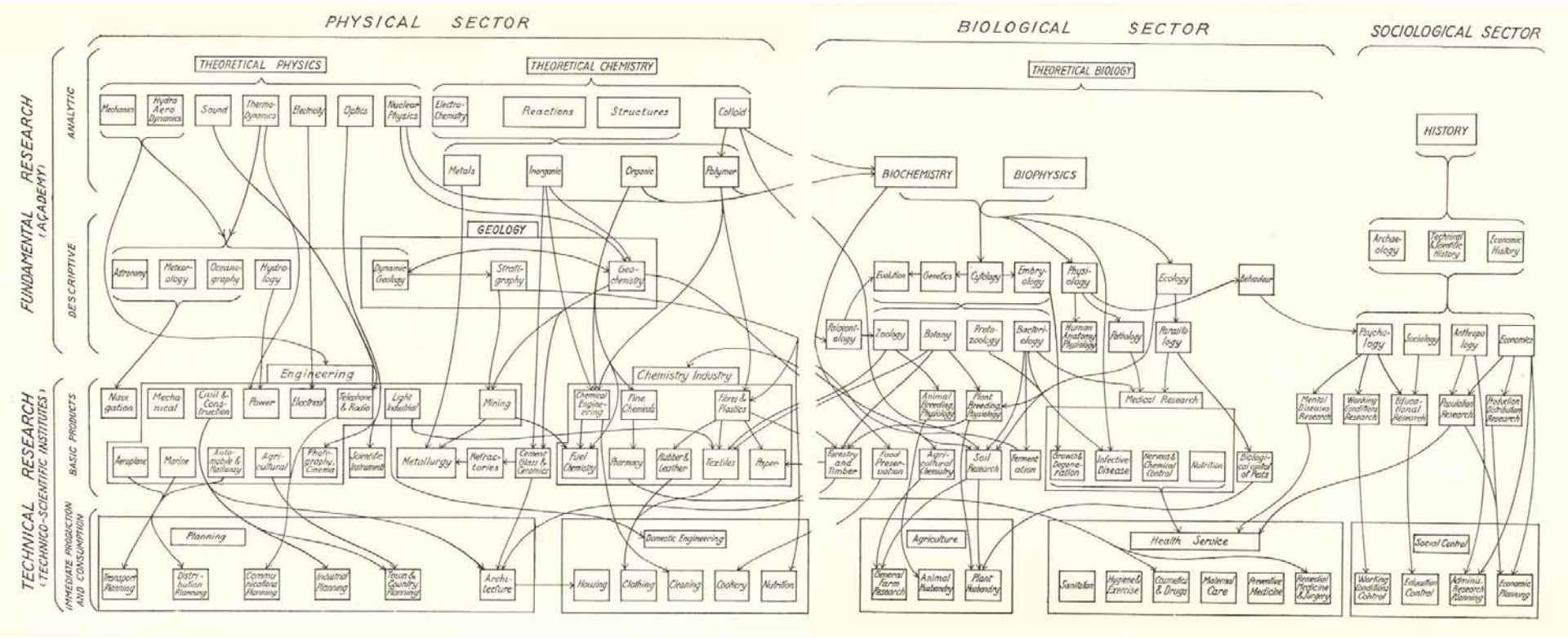
1970 – 2000: Early bibliometric maps

- Using ISI database on citing-cited behavior in articles
- Method was limited by computer hardware, software and measures

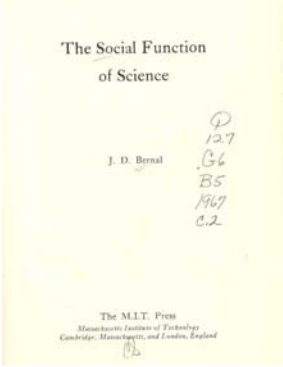
2001 – 2007: Rapid improvements in bibliometric mappings

- Computer memory
- Dimensional reduction algorithms
- Improvements in measures
- Focus on accuracy

Conceptual Map: Bernal, 1939



John D. Bernal was a world renowned physicist, a historian of science, and a sociologist of science. He is considered to have produced one of the first ‘maps’ of science.


The Social Function of Science

J. D. Bernal

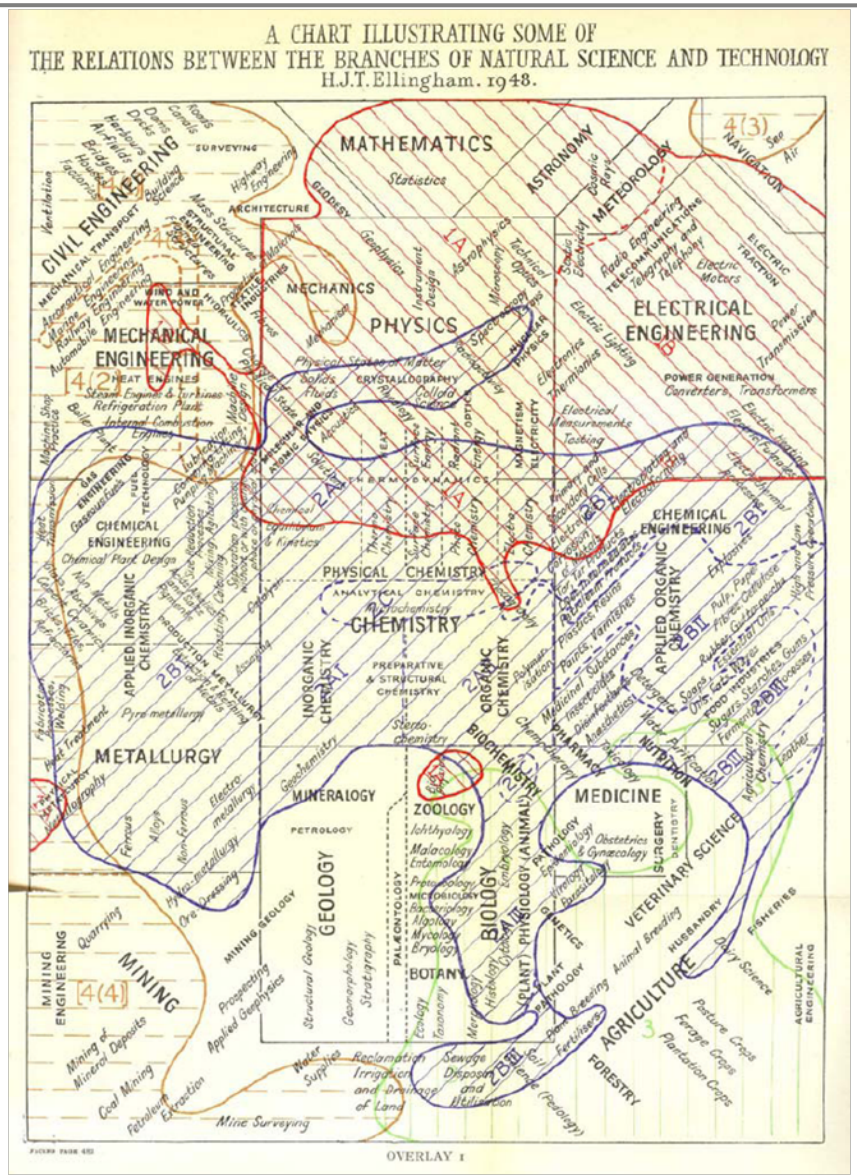
Q 127 G6 B5 1967 c.2

The M.I.T. Press

Bernal, J.D. (1939). *The Social Function of Science.* London: Routledge & Kegan Ltd.

Conceptual Map: Ellingham, 1948

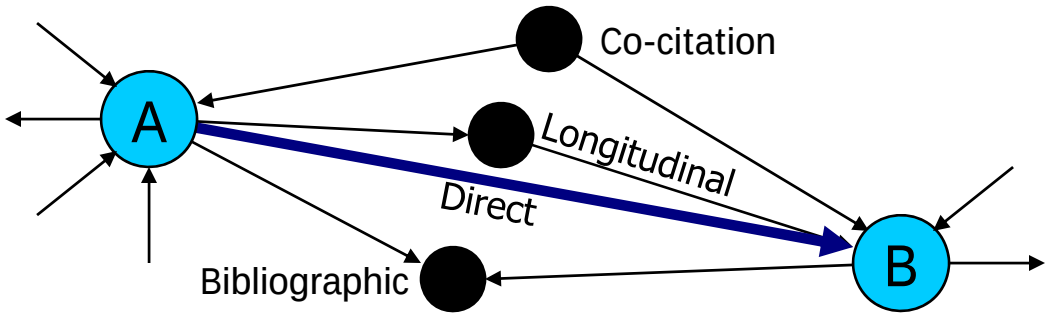
Ellingham's "Relations Between the Branches of Natural Science and Technology" with an overlay of "Abstracts or Groups of Abstracts Covering A Very Wide Field" (1948)



H. J. T. Ellingham, "Divisions of Natural Science and Technology," Royal Society Scientific Information Conference, 21 June to 2 July 1948 London: Burlington House, 1948.

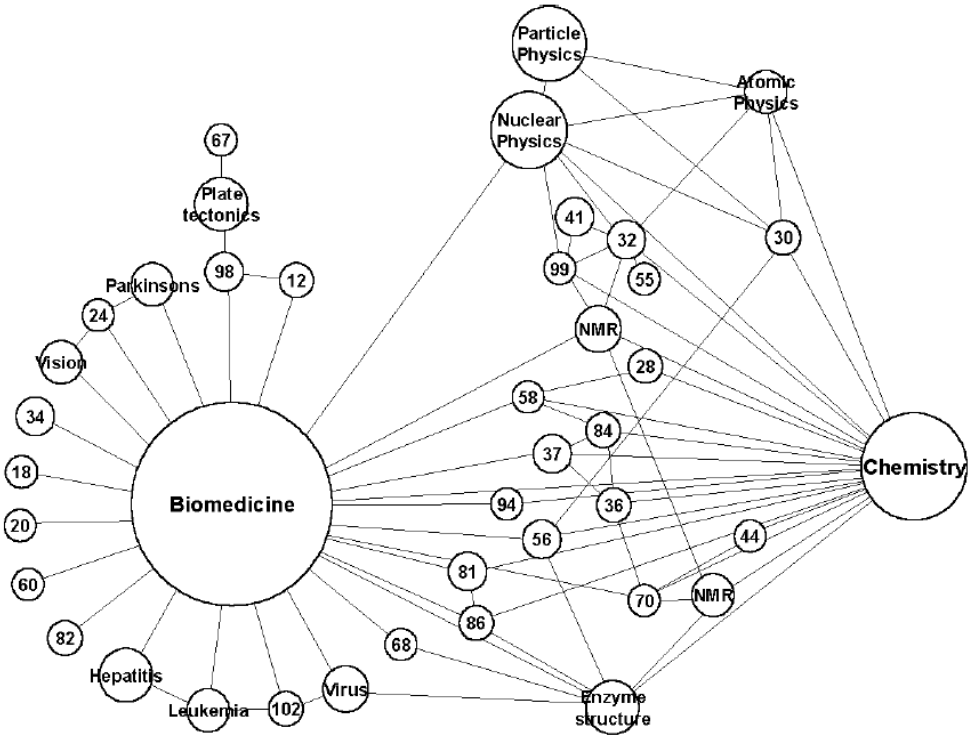
Link Types

- Direct links (real network)
  - can be binary network or weighted network
- Indirect links (implied network)
  - combinatorial nature makes this matrix very large
  - always a weighted network



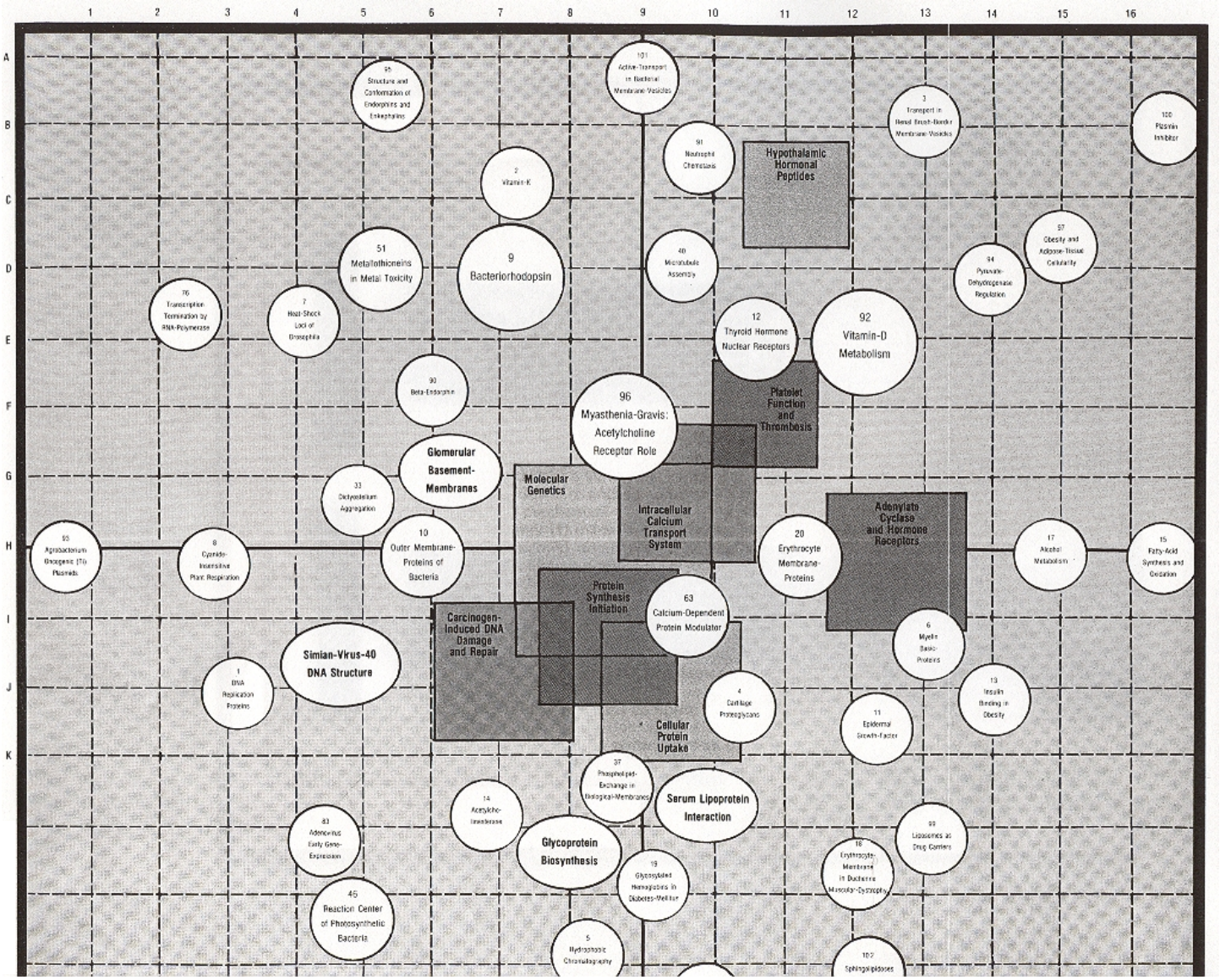
Early Bibliometric Mapping: ISI, 1974

- Paradigm map of science
  - Hierarchical clustering of scientific references
  - Node size reflects numbers of references
  - Links show relationships between clusters



Griffith, B. C., Small, H. G., Stonehill, J. A., Dey, S., Structure of scientific literatures. 2. Toward a macrostructure and microstructure for science, *Science Studies*, 4 (1974) 339–365.

Early Bibliometric Mapping: ISI, 1981



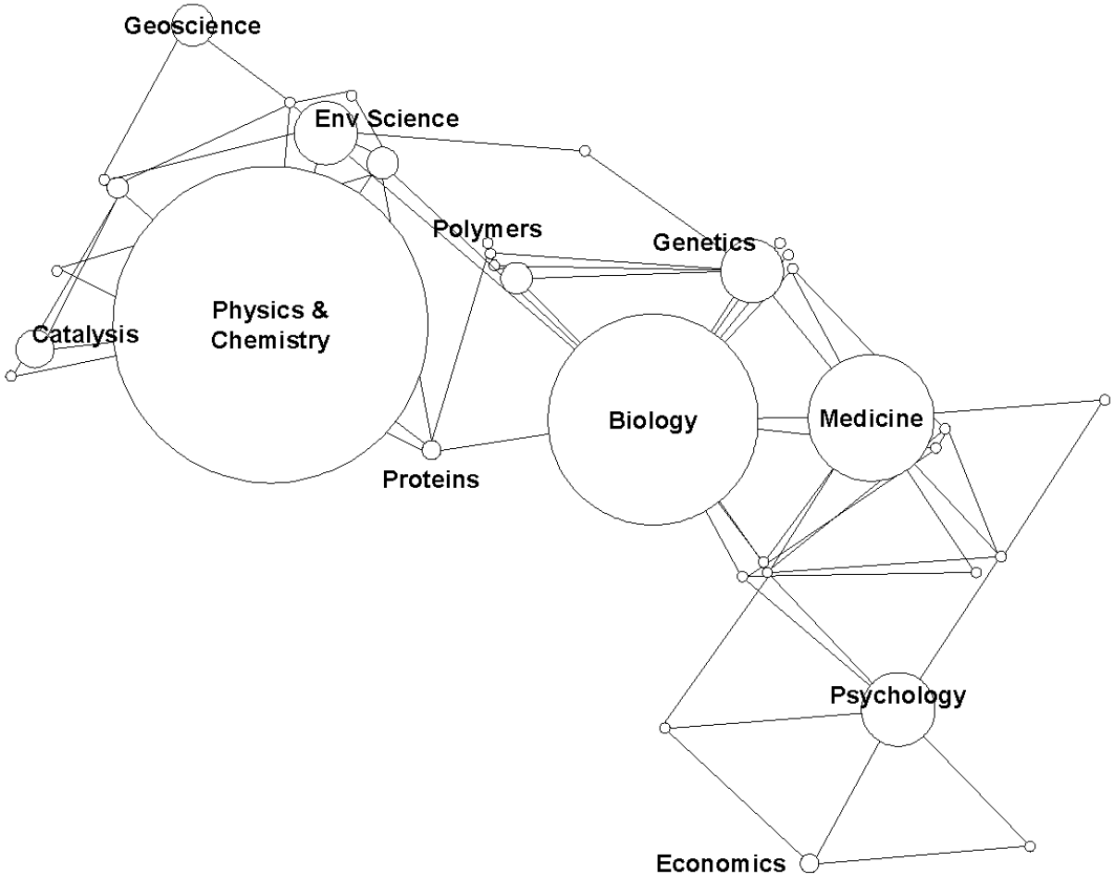
Institute for Scientific Information. (1981). *ISI atlas of science: Biochemistry and molecular biology, 1978/80*, Institute for Scientific Information, Philadelphia, PA.

Early Bibliometric Mapping: ISI, 1999

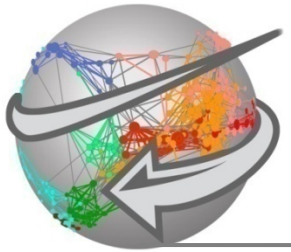
36,720 highly cited reference papers in a four-nested-level map

Triangulation process used for layout at the fourth level

High disciplinary bias due to thresholding



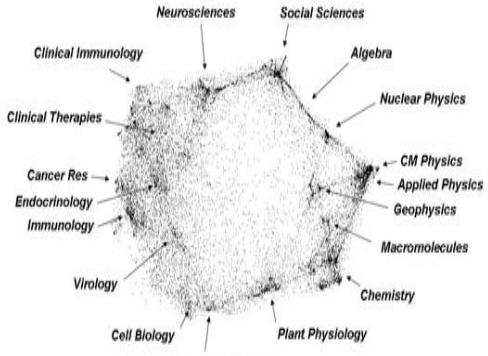
Small, H., Visualizing science by citation mapping, *Journal of the American Society for Information Science*, 50 (1999) 799–813.



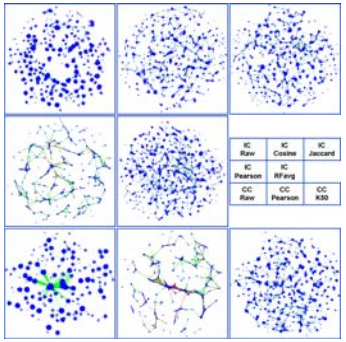
Summary: Early Bibliometric Mapping

- Technical Barriers
  - Costly (mainframes used to deal with 1MM x 1MM matrices)
  - Limitations in visualization software
    - Best method (MDS) could only handle 100 nodes
  - Limitations in clustering algorithms
    - Forced to use single link clustering results
      - Maximum cluster size was expected to be ~100 papers
      - Results were not consistent with assumptions (huge clusters appeared)
  - No methods for measuring accuracy/quality were developed
    - High costs resulted in very little experimentation in methods

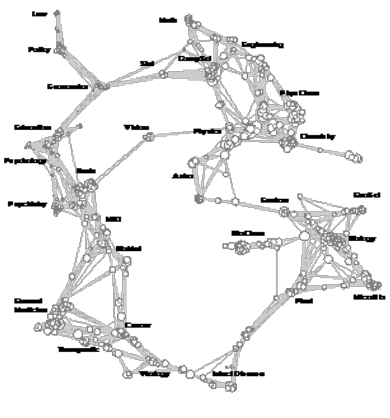
Recent Improvements



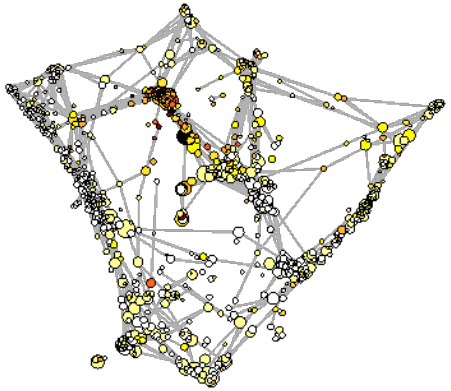
2000



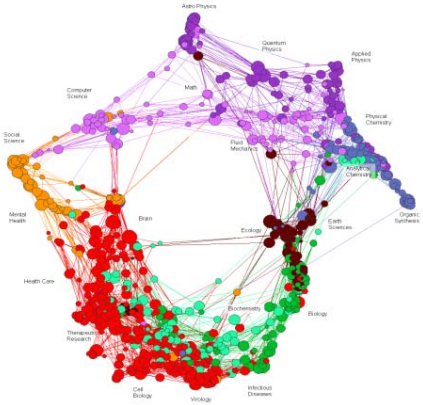
2003



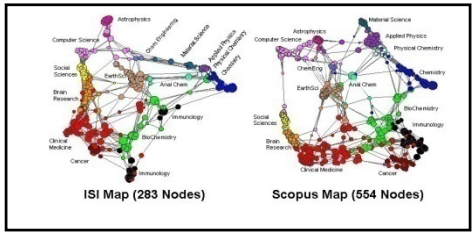
2004



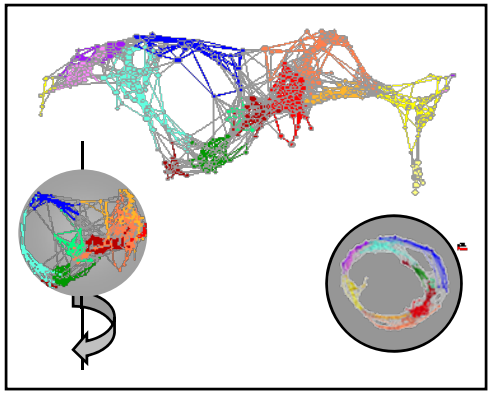
2005



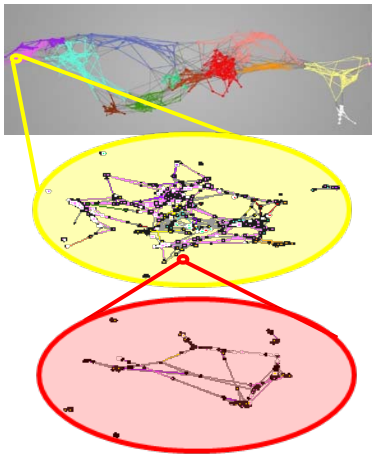
2005



2006



2007



2007

Improvements: 2000

Reference paper map linked to a journal map

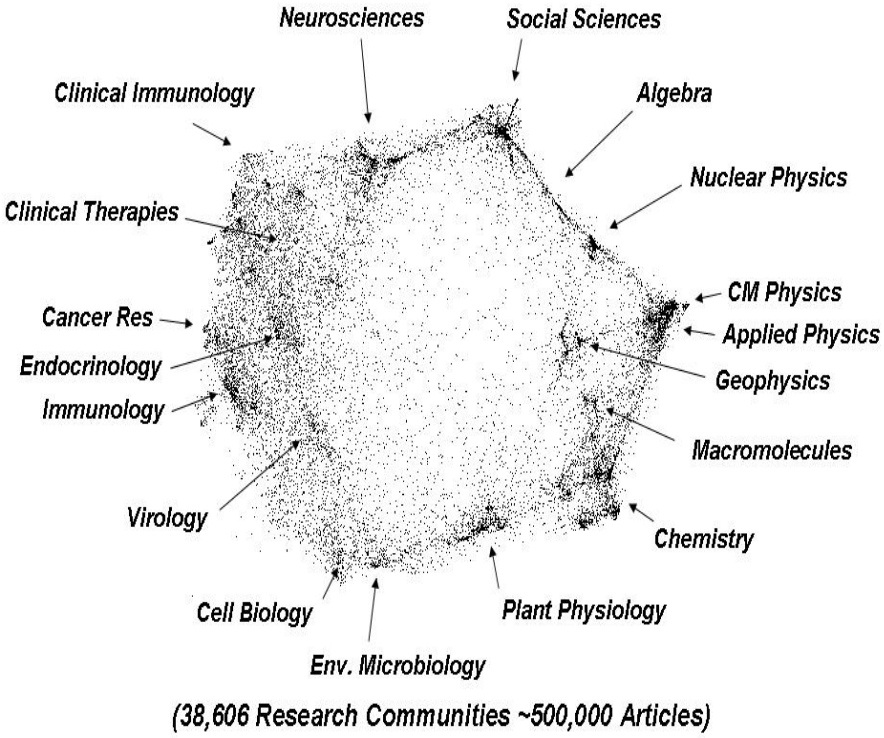
~500k papers clustered into ~38K paradigms

~5000 journals clustered into ~700 disciplines

Improvement:

Using hierarchical clustering of journals

Linking paper clusters to journal clusters



Klavans, R. (2000). Presented at the Sackler Symposium on Mapping Science, 2003, Irvine, CA

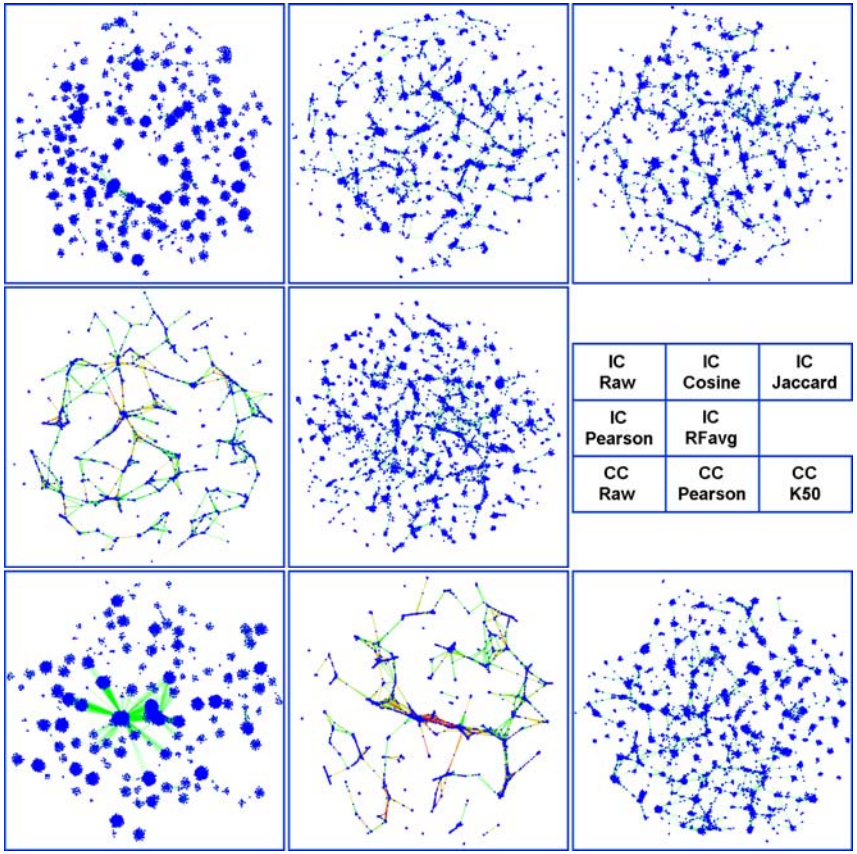
Improvements: 2003

Journal map

7000 journals (different similarity measures)

Improvement:

First approach to measure the accuracy of different maps

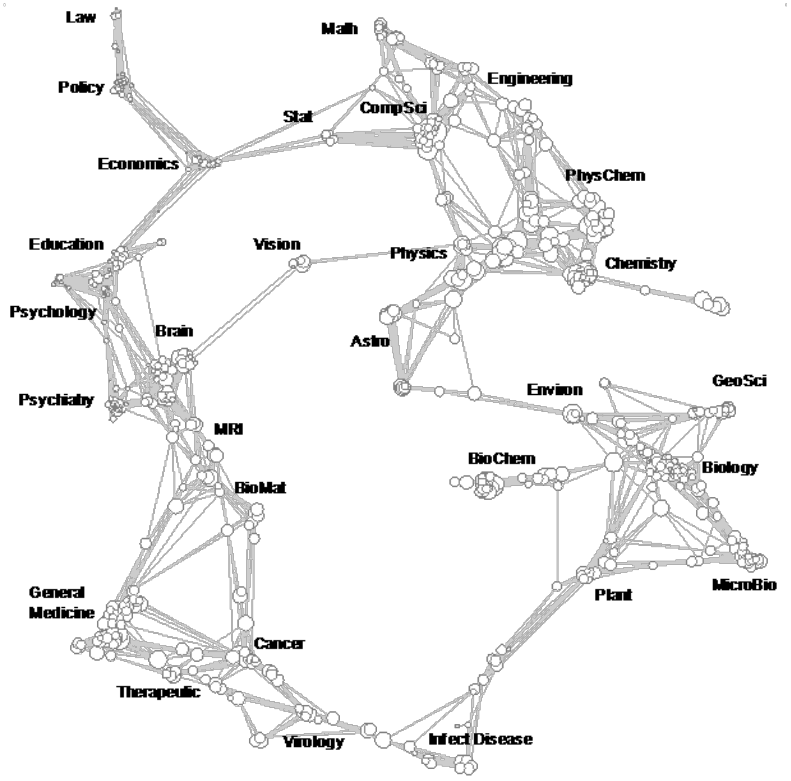


Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics* 64(3), 351-374.

Improvements: 2004

Journal map
- 7000 journals

Improvement:
- Increased accuracy by excluding multidisciplinary journals
- More intuitive macro-structure by using multiple levels of clustering



Boyack, K. W., Klavans, R., & Börner, K. (2007). Mapping the structure and evolution of chemistry research. *ISSI 2007*.

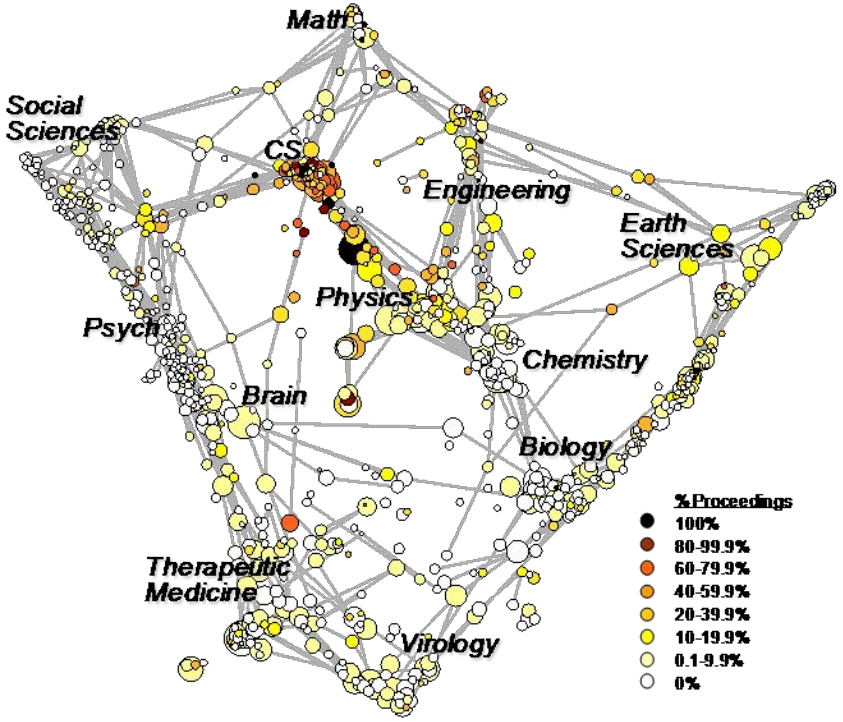
Improvements: 2005

Journal map

- 8500 journals / proceedings

Improvement:

- Added in the ISI proceedings database



Boyack, K. W. (2007). Using detailed maps of science to identify potential collaborations. *ISSI 2007*.

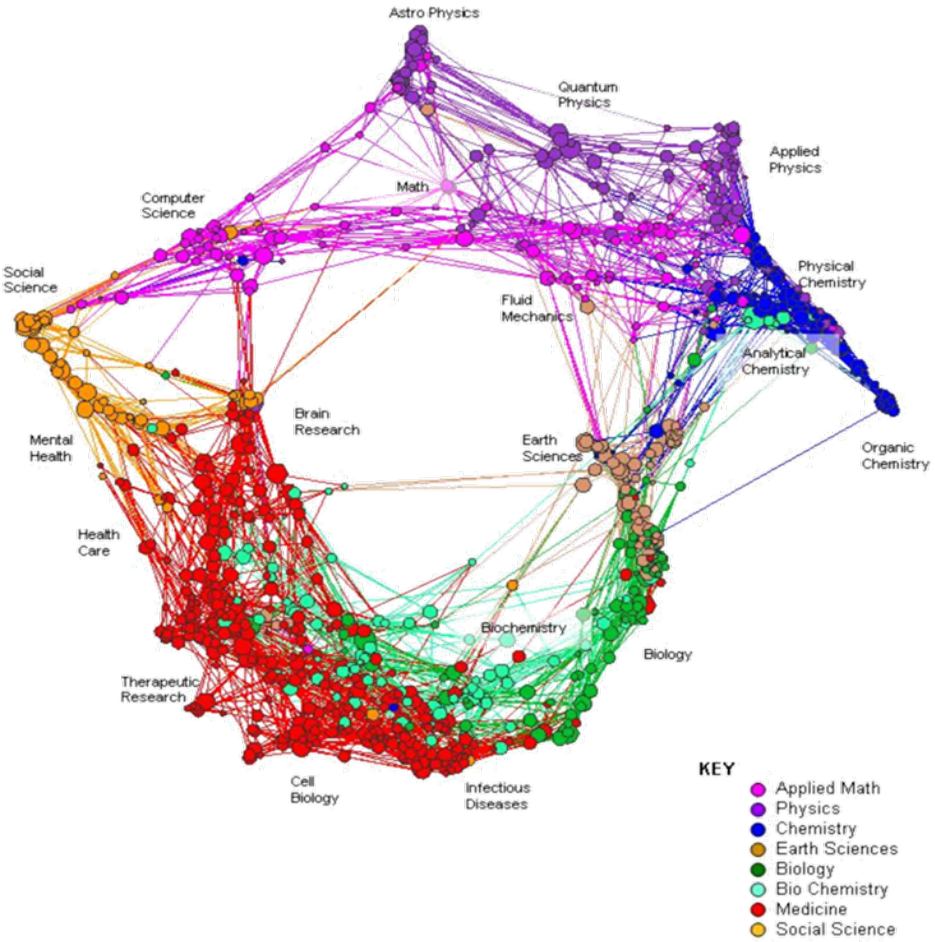
Improvements: 2005

Reference paper map

- 820,000 highly cited reference papers

Improvement:

- Replication of ‘circle of science’ using different data, measures and visualization algorithms



Klavans, R. & Boyack, K. W. (2008, forthcoming). Thought leadership: A new indicator for national and institutional comparison. *Scientometrics*.

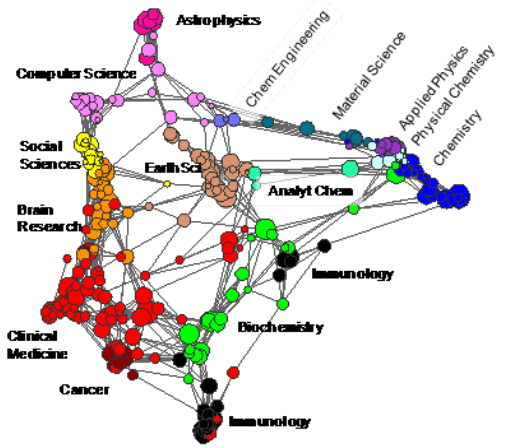
Improvements: 2006

Reference paper maps

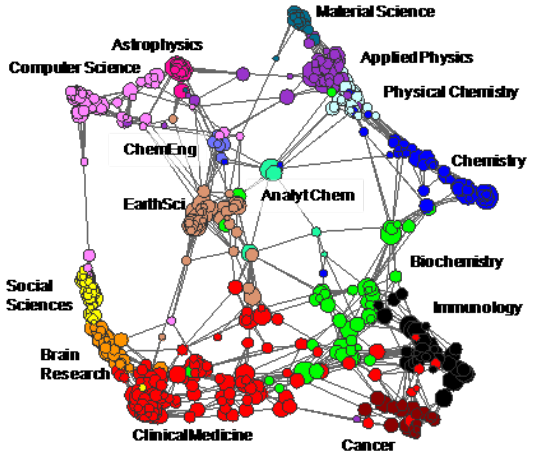
- ~800k highly cited reference papers
- two different databases and years

Improvement:

- Replication of methodology
- Comparison of databases

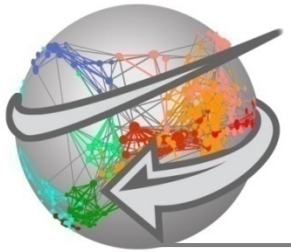


ISI Map (283 Nodes)



Scopus Map (554 Nodes)

Klavans, R. & Boyack, K. W. (2007). Is there a convergent structure of science? A comparison of maps from the ISI and Scopus databases. ISSI 2007.



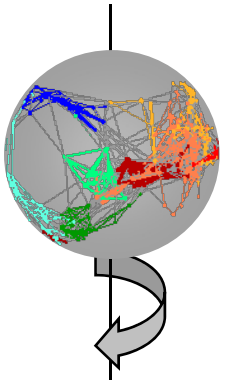
Improvements: 2007

Journal Map

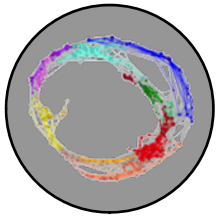
16,000 Journals into 554 disciplines

Improvements:

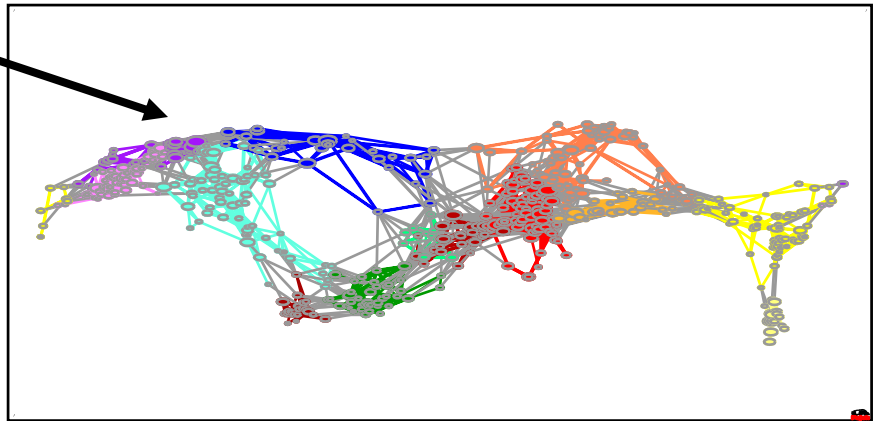
greater accuracy by using the surface of a sphere



Radial projection replicates ‘circle of science’



Mercator projection views the entire surface



Klavans, R. (2007, unpublished)
http://www.mapofscience.com/

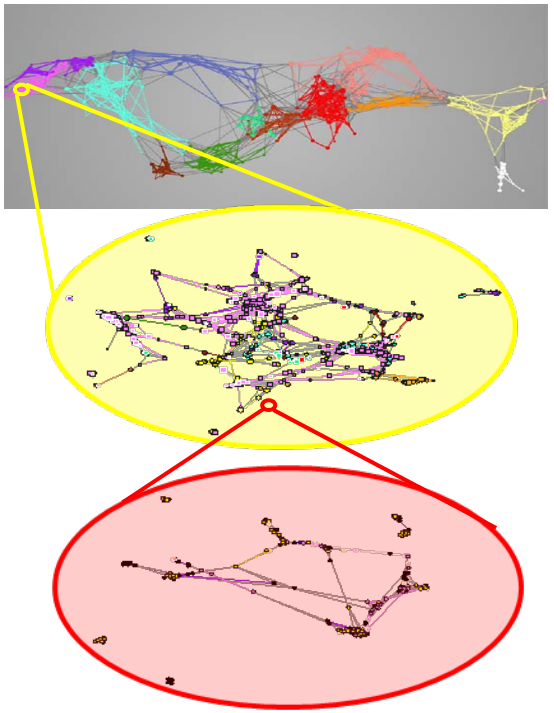
Improvements: 2007

Journal map linked to paradigm map linked to paper map
- 16,000 journals into 554 disciplines
- 2,160,000 highly cited references into 40,000 paradigms
- 4,870,000 Papers (2002-2006) assigned to 40,000 paradigms

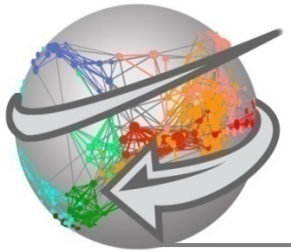
Improvements:
- Details for each of the disciplines

  Paradigm networks
  - for every 1 paradigm assigned to a discipline there are approximately 5 related paradigms

  Paper level maps show performance
  - nodes are highly cited references
  - edges are assigned current papers

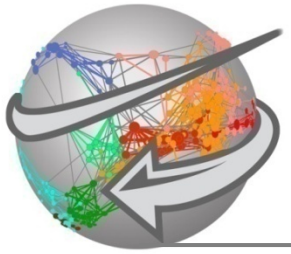


Klavans, R. (2007, unpublished)
http://www.mapofscience.com/



Summary: Improvements

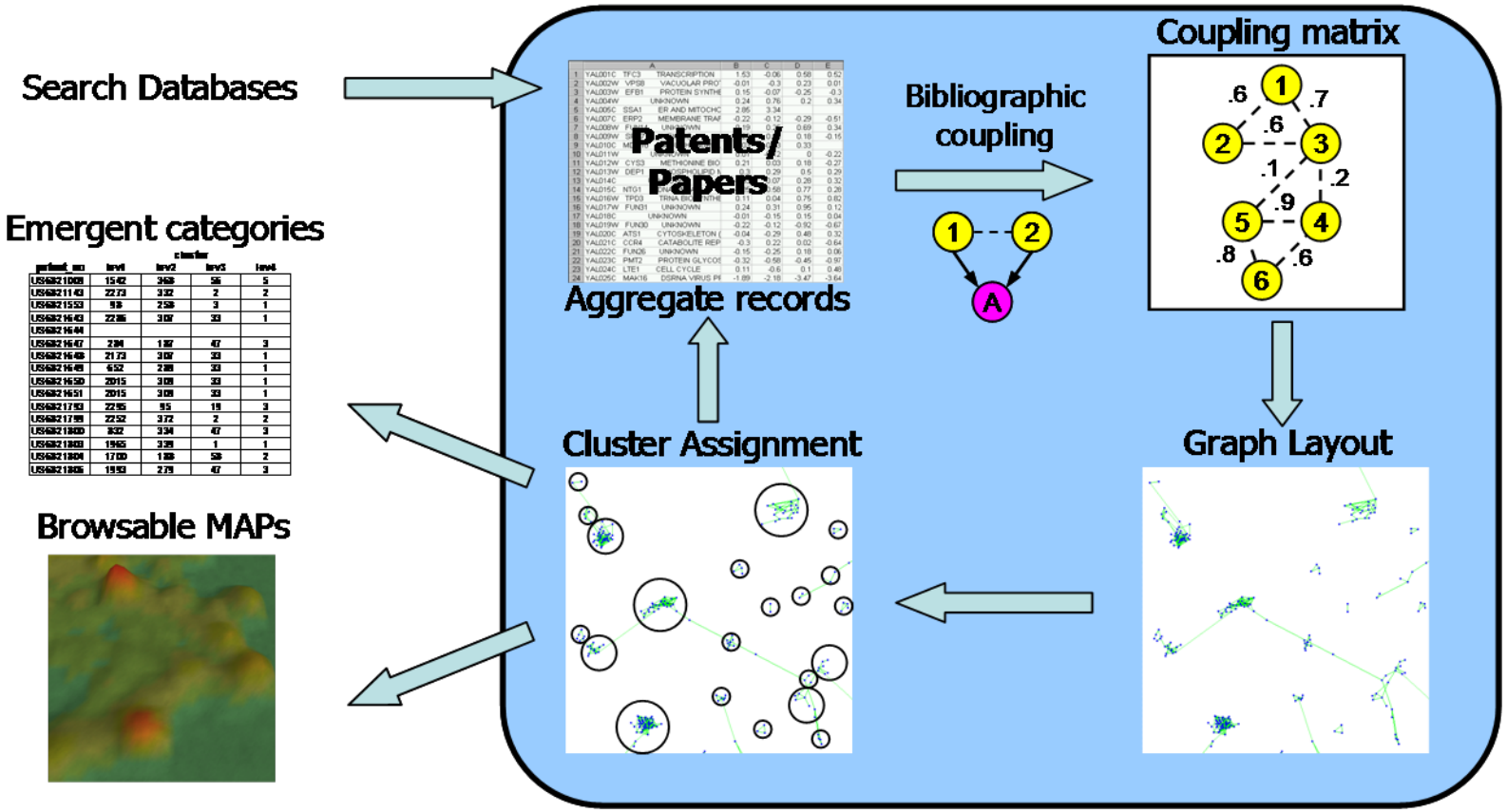
- Protocols for measuring accuracy
  - Experiment: New measures that generate more accurate maps
  - Experiment: New algorithms that generate more accurate maps
  - Experiment: Tradeoffs between accuracy and other factors
  - Continue experimenting until
    - Either one reaches convergence
    - Or alternatives become clearly differentiated
- New measures of strength
  - Network strength vs. Isolated strength
  - Publication leadership vs. Thought leadership

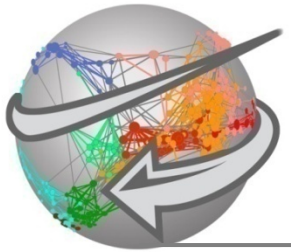


Progress in Measuring Accuracy

- Journal Maps
  - Local accuracy
  - Regional accuracy
  - Text & citations
  - Lessons learned
- Paper-Level Maps
  - Local accuracy
  - Disciplinary bias
  - Lessons learned

Hierarchical Mapping Process



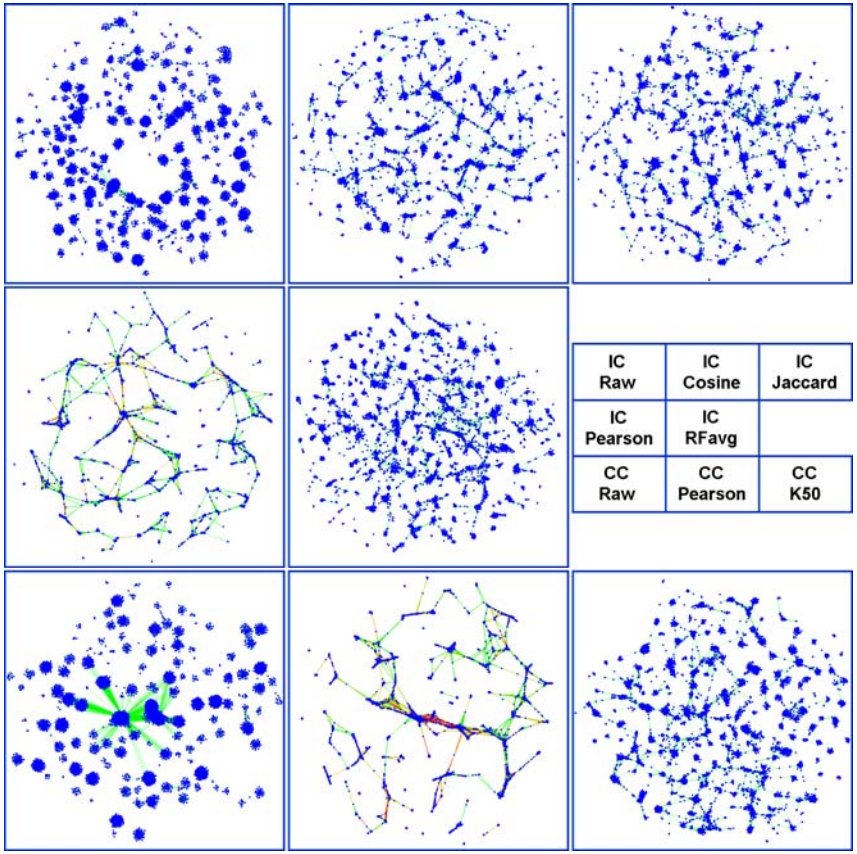


Methodological Limitation

- VxOrd is most applicable where local conditions dominate over far-flung or global conditions
  - A journal network is a multidimensional system
  - However, each journal only cares about its local environment (ask a journal editor how many journals she considers as competitors)
- Pruning of the network (rather than thresholding) is appropriate
  - Pruning (removing all but the top 10 edges per node) retains nodes, does not preserve degree, does preserve macro-structure, provides visual separability
  - Thresholding (removing edges under a certain weight) removes nodes, preserves relative degree, leads to over-aggregation

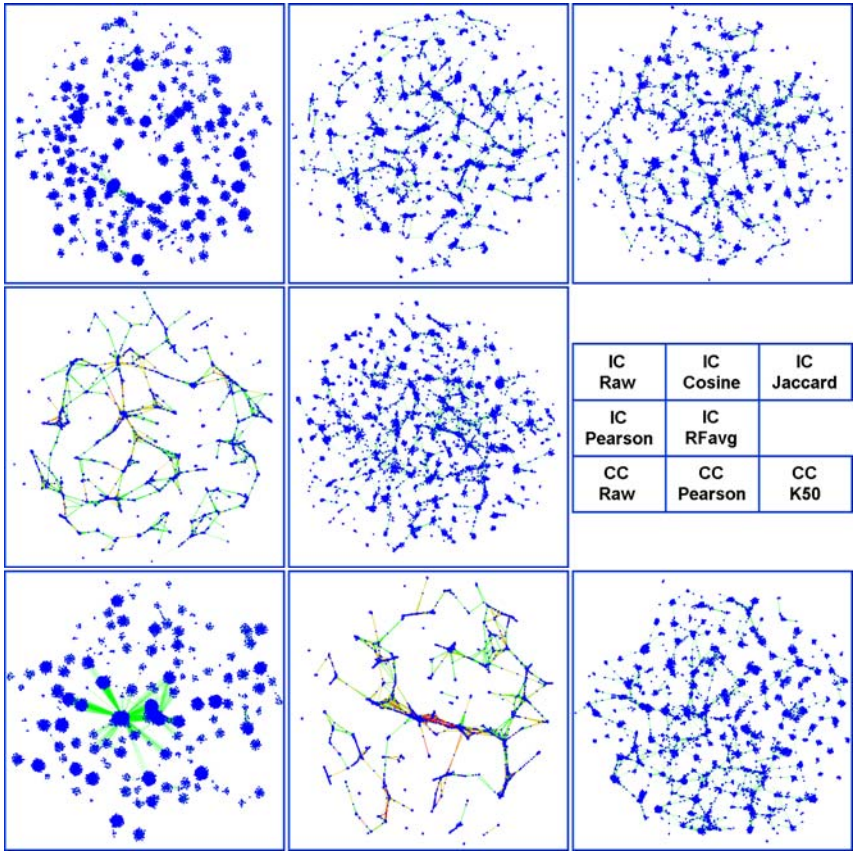
Example: Journal Maps

- Same Data
  - ISI SCIE & SSCI file year 2000
- 8 different similarity metrics
  - 5 Inter-citation (raw counts, cosine, Jaccard, Pearson, RF Average)
  - 3 Co-citation (raw counts, Pearson, modified cosine)



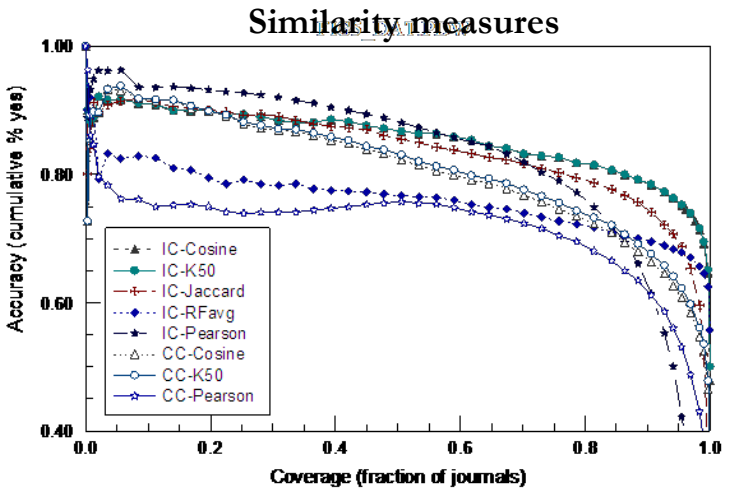
Accuracy of Journal Maps

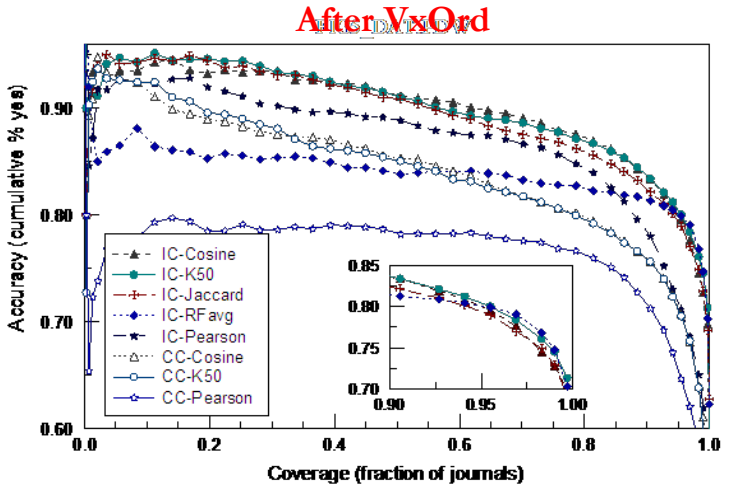
- Local Accuracy
  - The journal that is closest to you is assigned to the same disciplinary category using a gold standard (the ISI disciplinary classification system).
- Regional Accuracy
  - Clusters of journals correspond to the gold standard (the ISI disciplinary clasification system).



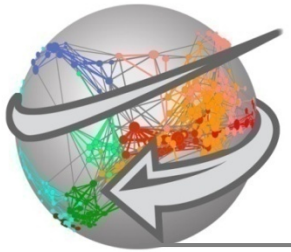
Journal Maps: Local Accuracy

- ISI category structure used as basis for comparison
- Similarity pairs ordered by value, pairs assigned a 1/0 binary score (same category?)
- Accuracy vs. coverage curves were generated for each similarity measure
- "After VxOrd" based on distances in layout rather than similarity values
- Results after layout using VxOrd were more accurate than the raw measures





Klavans, R., & Boyack, K. W. (2006). Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology* 57(2), 251-263.

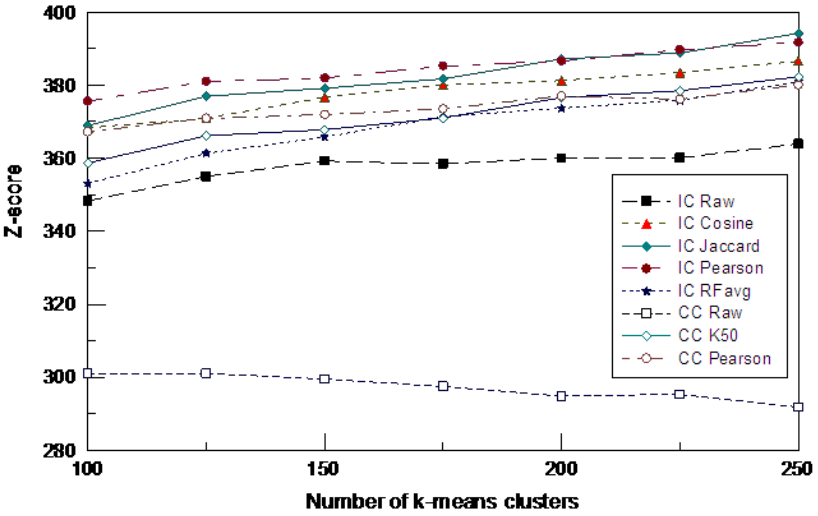


Computing Mutual Information

- Use method of Gibbons and Roth (Genome Research v. 12, pp. 1574-1581, 2002)
- K-means clustering (MATLAB) for each graph layout
  - 3 different k-means runs at 100, 125, 150, 175, 200, 225, 250 clusters
- Quality metric (mutual information) calculated as
  - MI(X,Y) = H(X) + H(Y) – H(X,Y)
  - where $H = -\sum P_i \log_2 P_i$
  - P_i are the probabilities of each [cluster, category] combination
  - X (known ISI category assignments), Y (k-means cluster assignments)
- Z-score (indicates distance from randomness, Z=0=random)
  - $Z = (MI_{real} – MI_{random})/ S_{random}$
  - MI_{random} and S_{random} vary with number of clusters, calculated from 5000 random solutions

Journal Maps: Regional Accuracy

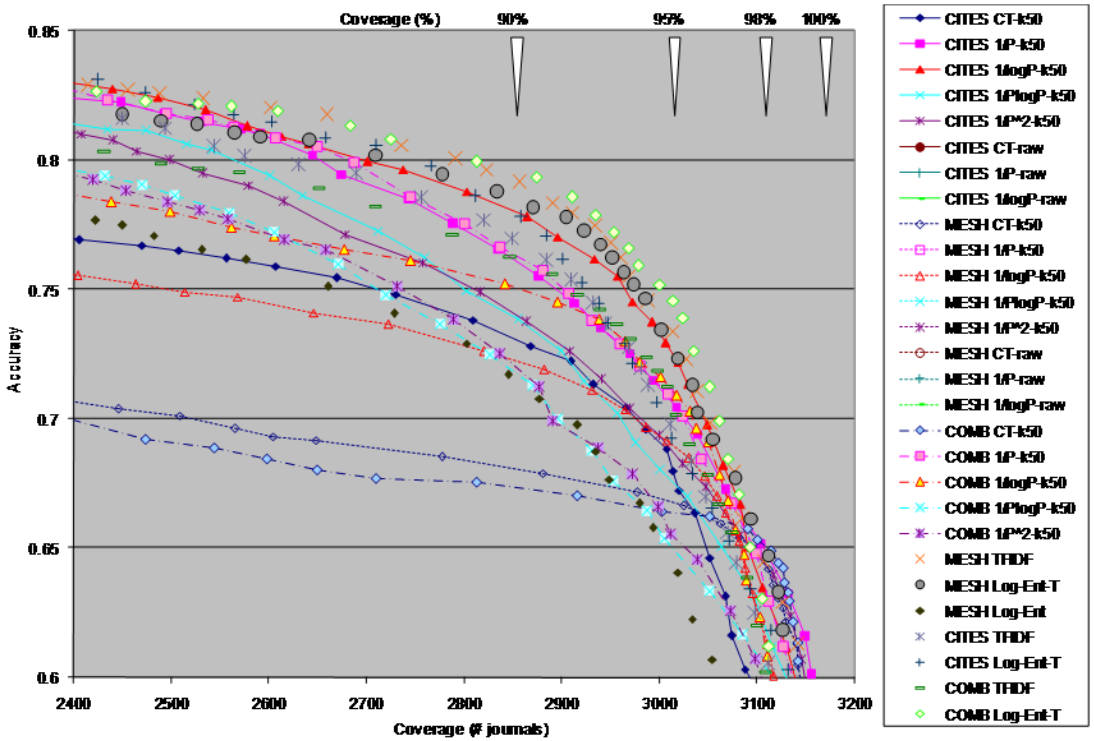
- ISI category structure used as basis for comparison
- Layouts clustered (k-means) with different numbers of clusters
- Resulting cluster memberships were compared to actual structure using entropy/ mutual information method
- Increasing Z-score indicates increasing distance from a random solution
- Most similarity measures are within several percent of each other



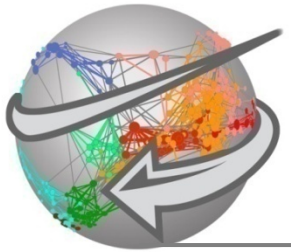
Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics* 64(3), 351-374.

Journal Maps: Citations vs. Text

- Can keywords or controlled vocabularies compete with citations?
- Data: ~3300 Medline journals covered by ISI
  - Citation data from ISI
  - MeSH terms from Scopus
- Computed accuracy/ coverage curves for similarity measures using citations, MeSH, and a combination
- Some MeSH and combined measures do as well as citation measures



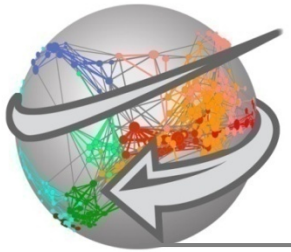
Boyack & Klavans (2007, unpublished).



Journal Maps: Citations vs. Text

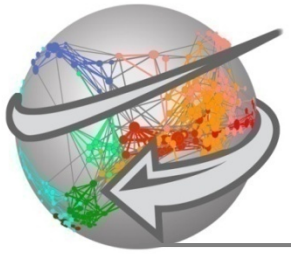
- Better text analysis/extraction is needed
  - Our results with MeSH are encouraging, BUT …
    - MeSH is a standardized vocabulary
    - assigned by a select group of human indexers
    - in one field (medicine) of science
  - Would a non-standard vocabulary do as well?

- Accurate (automatic) labeling is another issue



Journal Maps: Lessons Learned

- Clear tradeoff between coverage and accuracy.
- Pearson correlation is more accurate for low levels of coverage and is not scalable to hundreds of thousands of items.
- K50 (normalized cosine) is the most accurate overall measure of journal:journal relatedness for higher levels of coverage.
- Data reduction (e.g. VxOrd) actually increases accuracy!
- Combining text & citations can increase accuracy. Further work needed.
- Best practice: Use distance measure from the 2 dimensional (x,y) map generated by VxOrd using K50 as the input journal:journal measure of relatedness. Coverage ~90% still maintains high enough levels of accuracy.

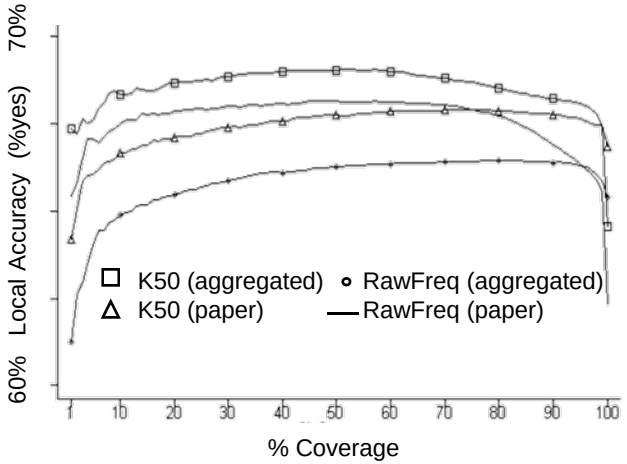


Accuracy of Paper Maps

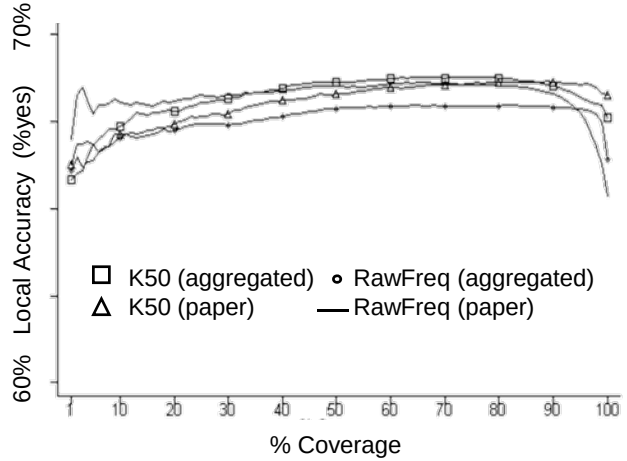
- Local accuracy
  - The paper that is closest to you is assigned to the same disciplinary category using a gold standard (the ISI disciplinary classification system).
- Disciplinary bias
  - What threshold is necessary to generate acceptable levels of disciplinary bias?

Paper Maps: Local Accuracy

- Two maps (current and reference), two measures (raw and modified cosine-K50), two aggregation levels (edge-cutting levels)
- Paper pairs ordered by distance on the map, pairs assigned a 1/0 binary score (same ISI category?)
- Accuracy vs. coverage curves were generated for all 2x2x2 cases
- K50 measures have high accuracy at high coverage



Local Accuracy (Current Papers)

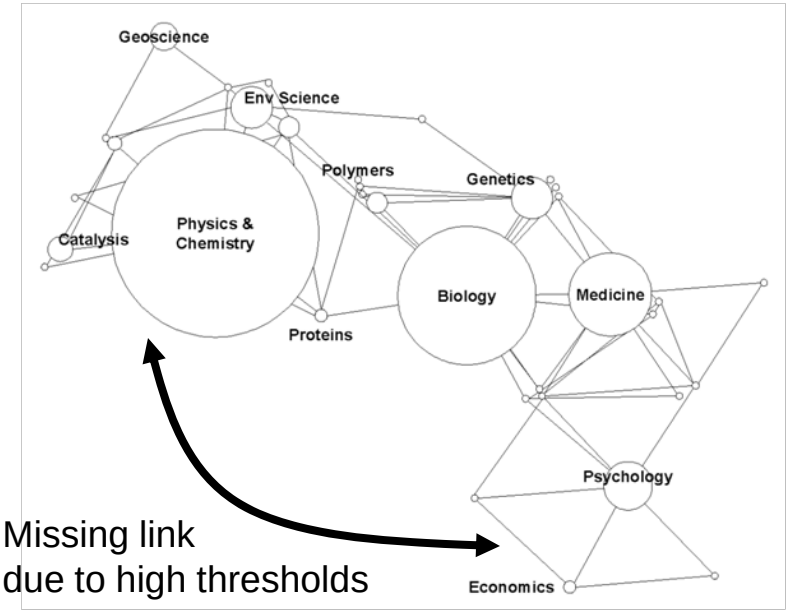


Local Accuracy (Reference Papers)

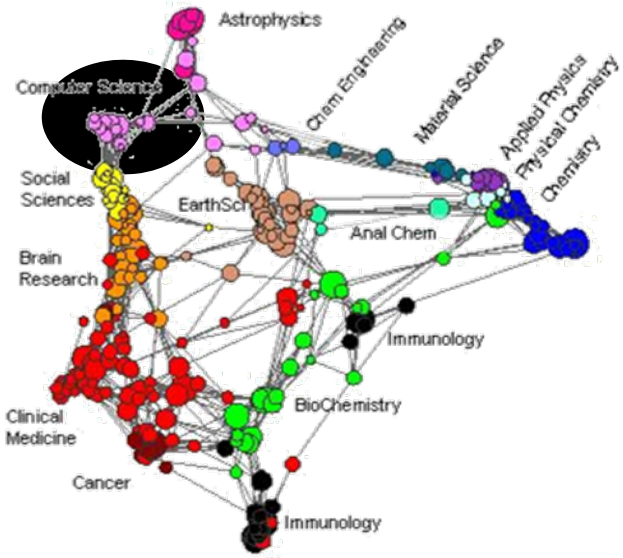
Klavans, R. & Boyack K. W. (2006). Quantitative evaluation of large maps of science. *Scientometrics* 68(3), 475-499.

Paper Maps: Disciplinary Bias

Small [1999] presented a more linear map of science, but also reported high disciplinary bias. Disciplinary bias is due to using a high citation threshold for selecting reference papers.



Klavans & Boyack [2007], using the same ISI database with much lower thresholds, suggest that there is a strong link between economics and physics – going through the disciplines that are under-represented by Small [1999].

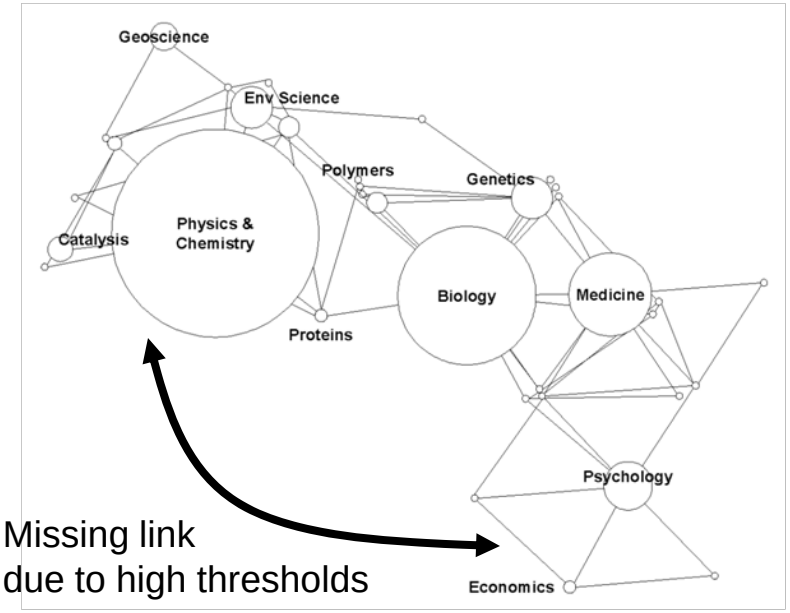


Small, H., Visualizing science by citation mapping, *Journal of the American Society for Information Science*, 50 (1999) 799–813.

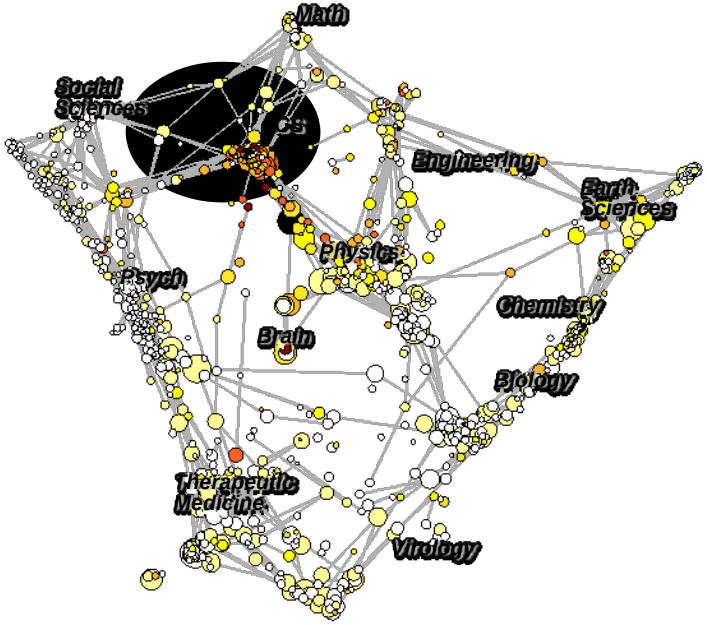
Klavans, R. & Boyack, K. W. (2007). Is there a convergent structure of science? A comparison of maps from the ISI and Scopus databases. ISSI 2007.

Paper Maps: Disciplinary Bias

Small [1999] presented a more linear map of science, but also reported high disciplinary bias. Disciplinary bias is due to using a high citation threshold for selecting reference papers.



Boyack [2007], also using a lower thresholds and including the proceedings database, suggests that the missing link between economics and physics goes through the disciplines that were under-represented by Small [1999].

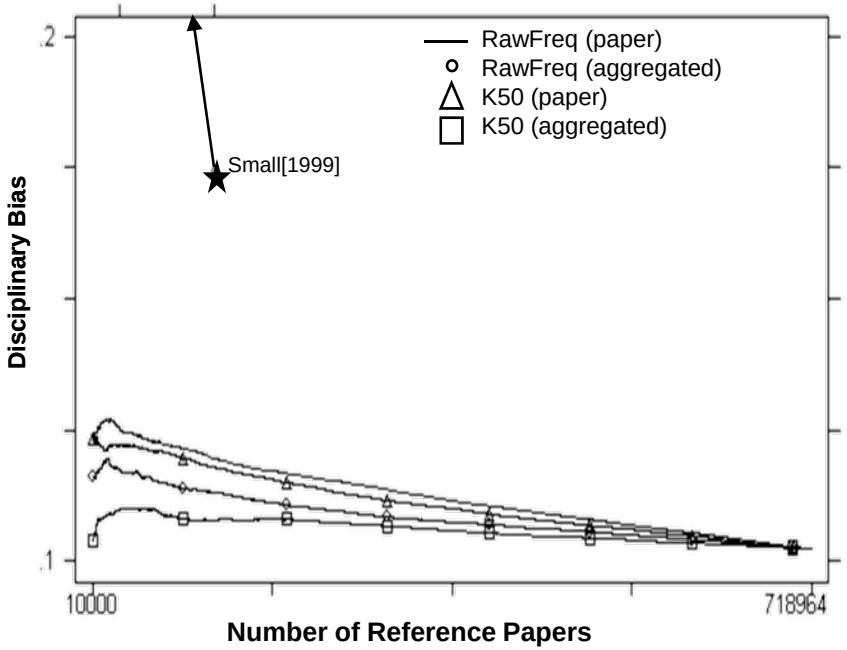


Small, H., Visualizing science by citation mapping, *Journal of the American Society for Information Science*, 50 (1999) 799–813.

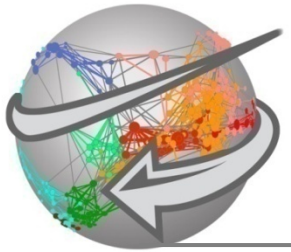
Boyack, K. W. (2007). Using detailed maps of science to identify potential collaborations. *ISSI 2007*.

Paper Maps: Disciplinary Bias

- Disciplinary bias reported by Small is very high.
- Disciplinary bias, using the new methodologies developed by SciTech Strategies is very low.
- K50 (modified cosine) measure of paper-paper relatedness generates the lowest levels of disciplinary bias.

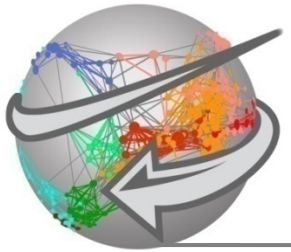


Klavans, R. & Boyack K. W. (2006). Quantitative evaluation of large maps of science. *Scientometrics* 68(3), 475-499.



Paper Maps: Lessons Learned

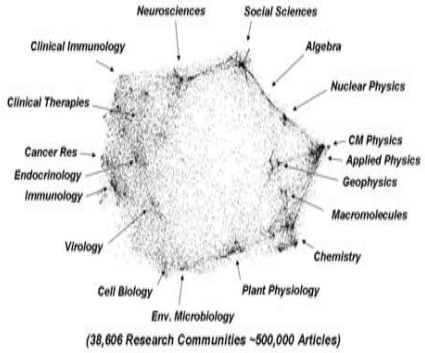
- Clear tradeoff between disciplinary bias and accuracy.
- Choosing lower thresholds creates more accurate maps.
- Choosing lower thresholds requires the use of algorithms that can handle square matrices that have millions (not thousands) or rows.
- K50 (modified cosine) remains the most accurate, scalable measure of relatedness.
- Best practice: Use the lowest reasonable thresholds; K50 as the measure of paper-paper relatedness to generate x,y coordinates using VxOrd, and then use the distance measure from the coordinates map.



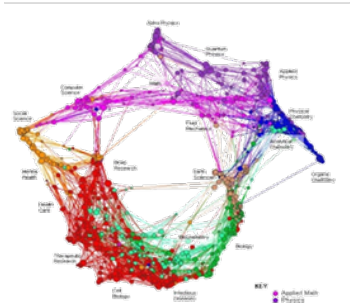
Frontier Issues

- Is there convergence?
- How can network science help?

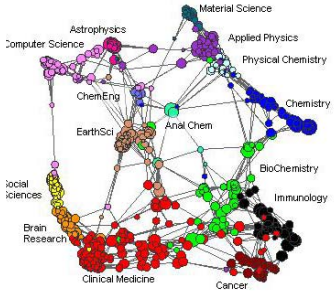
Is there Convergence Yet?



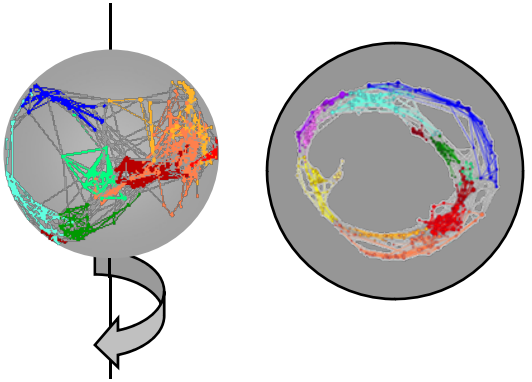
2000 ISI Disciplinary Map



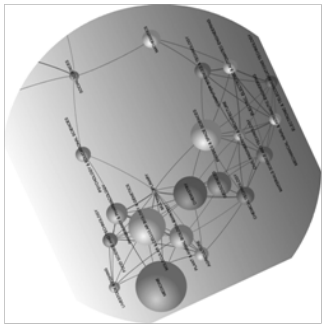
2003 ISI Paradigm Map



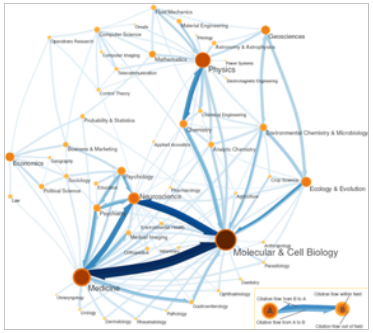
2004 Scopus Paradigm Map



2001- 2005 UCSD Disciplinary Map

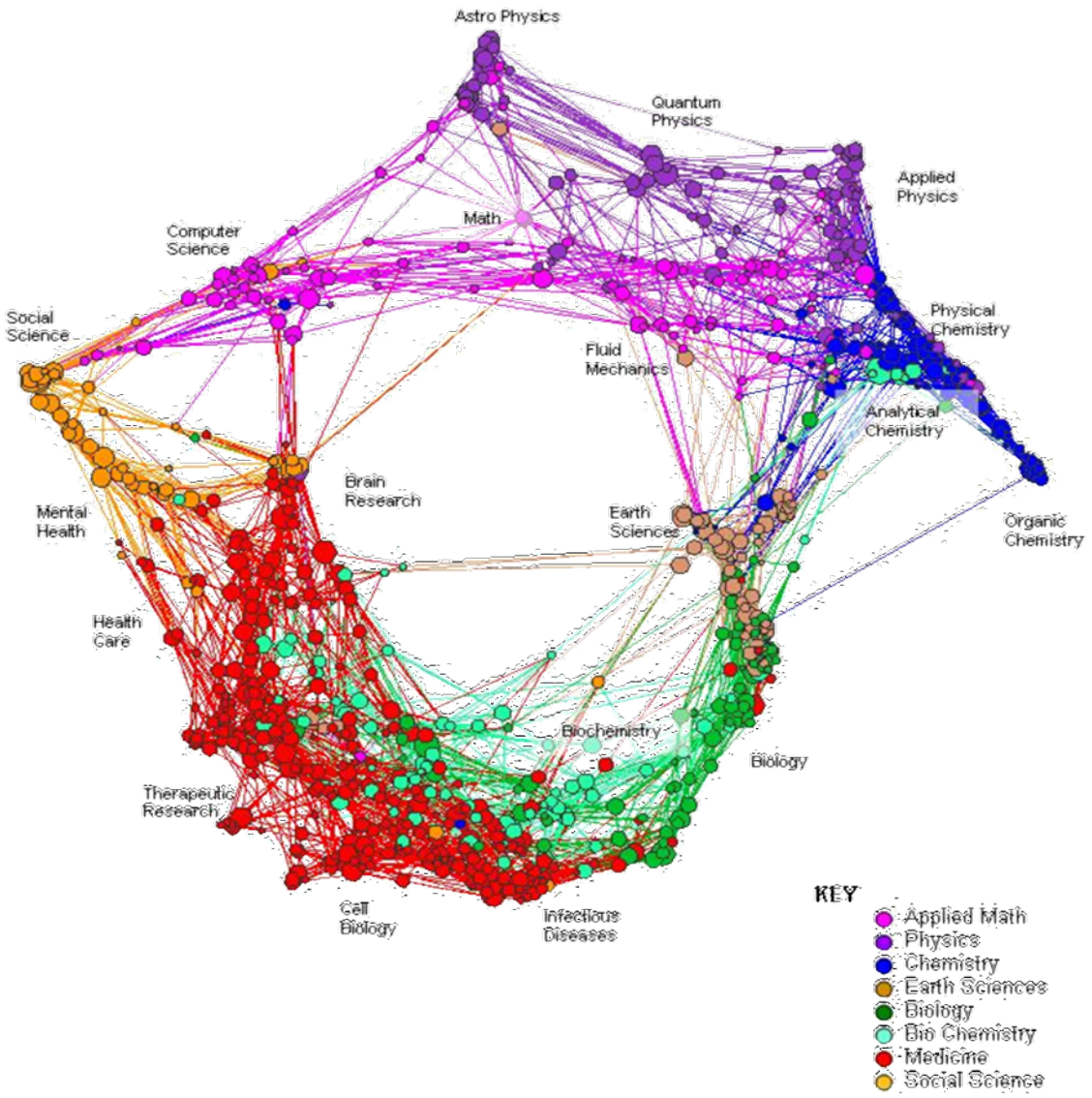


Scimago: 2000 ISI Co-category Map

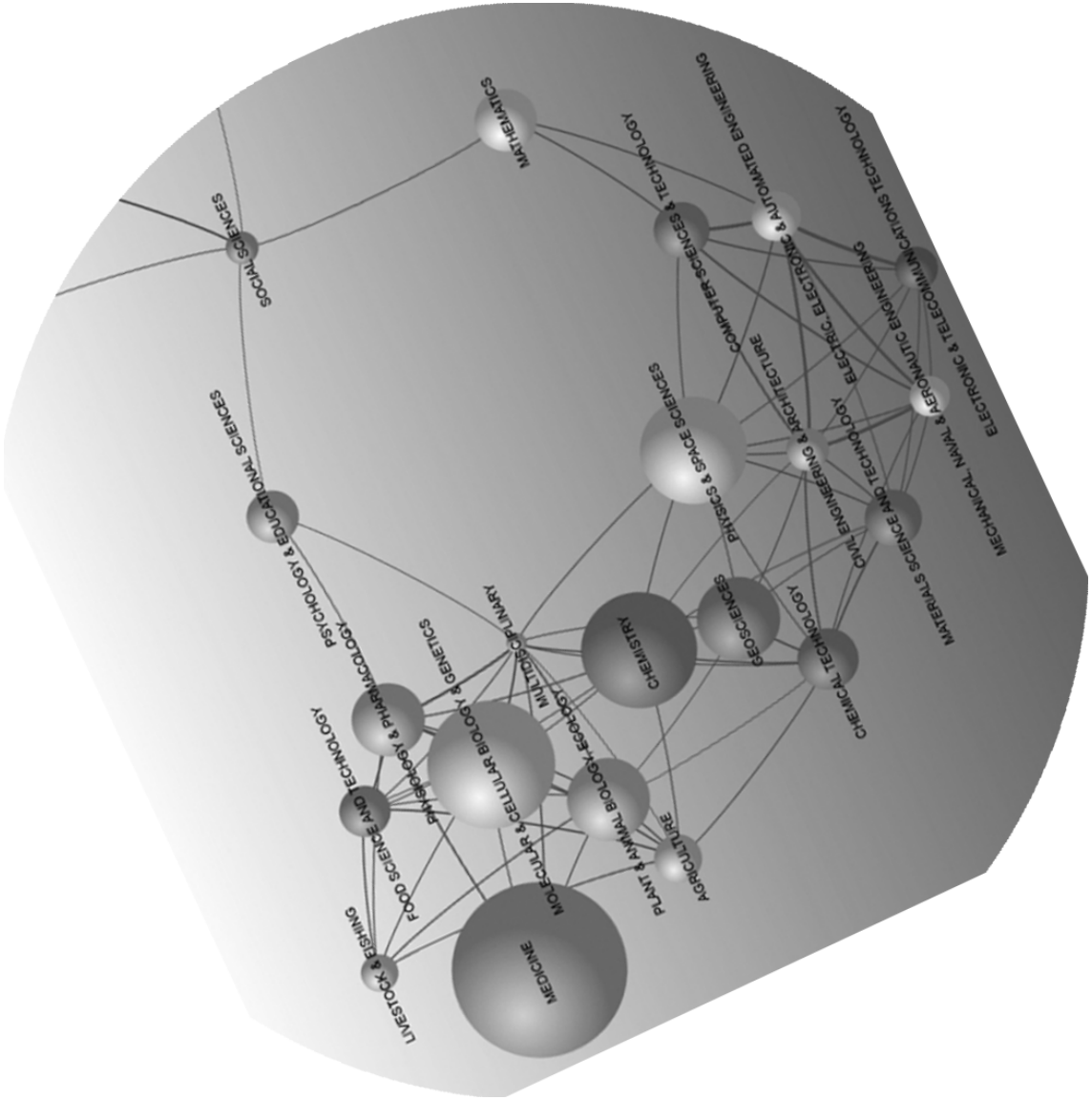


U. Wash: 2004 ISI Disciplinary Map

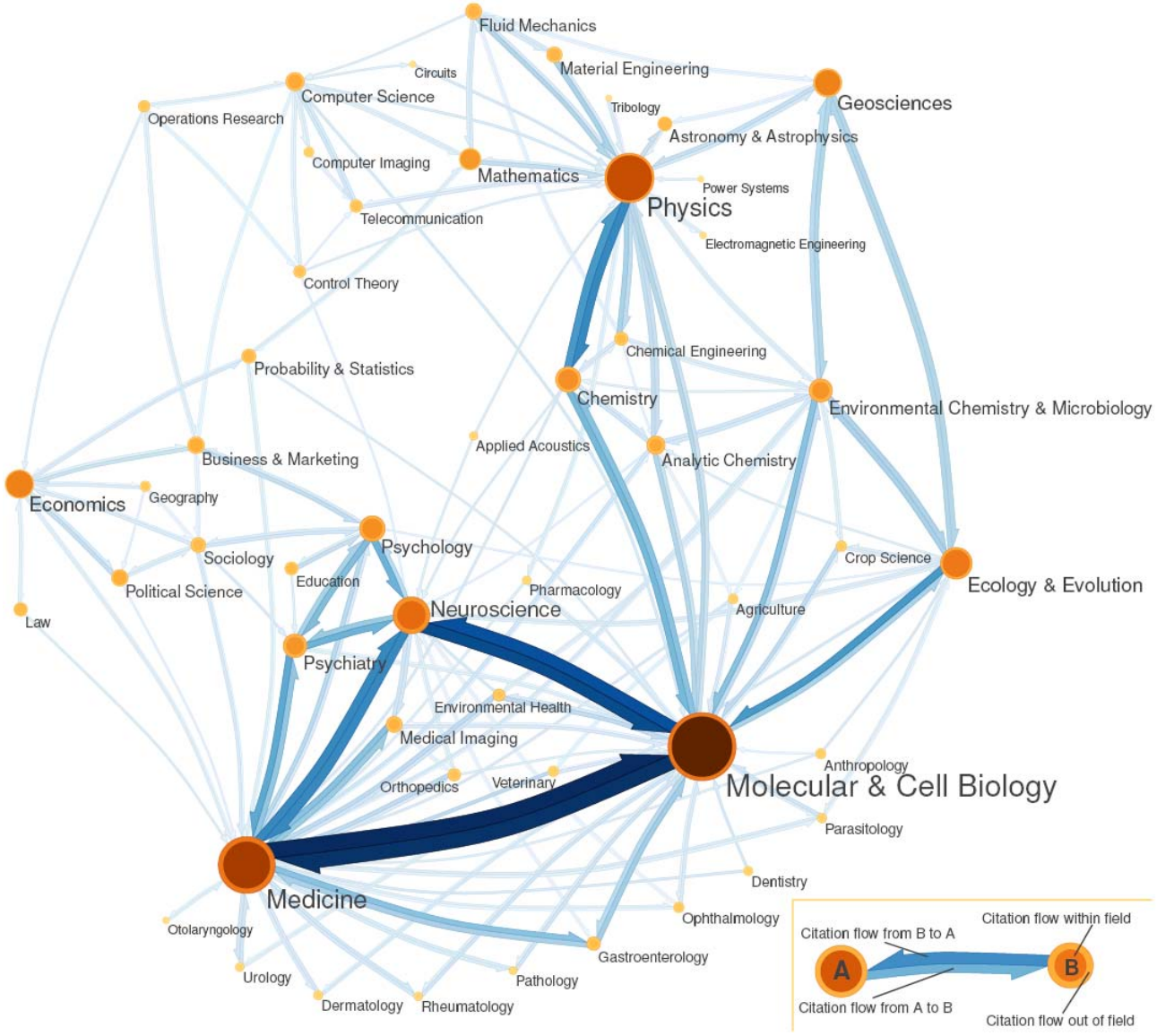
Is there Convergence Yet?



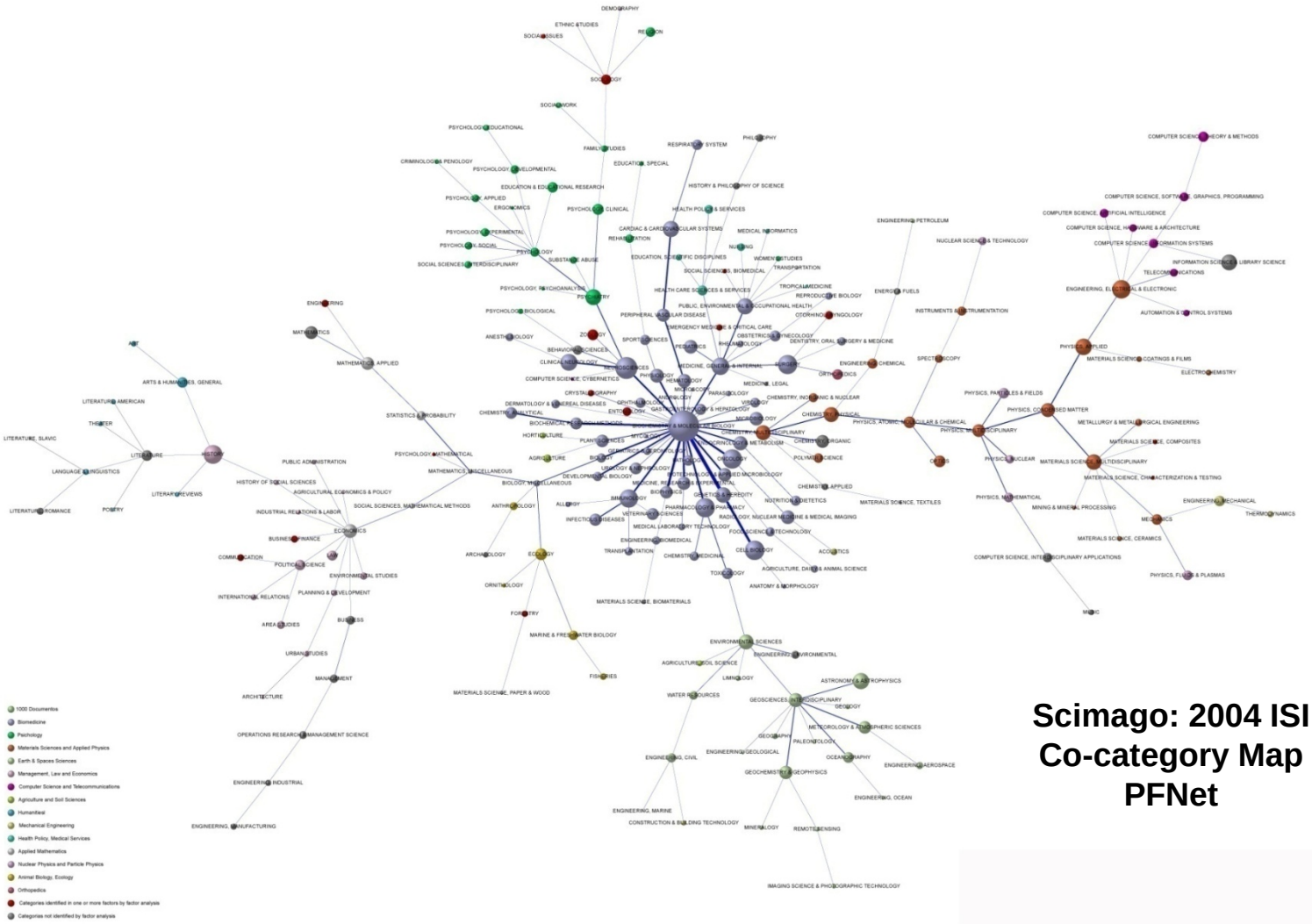
Is there Convergence Yet?



Is there Convergence Yet?

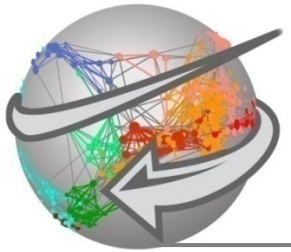


Maybe Not



Scimago: 2004 ISI Co-category Map PFNet

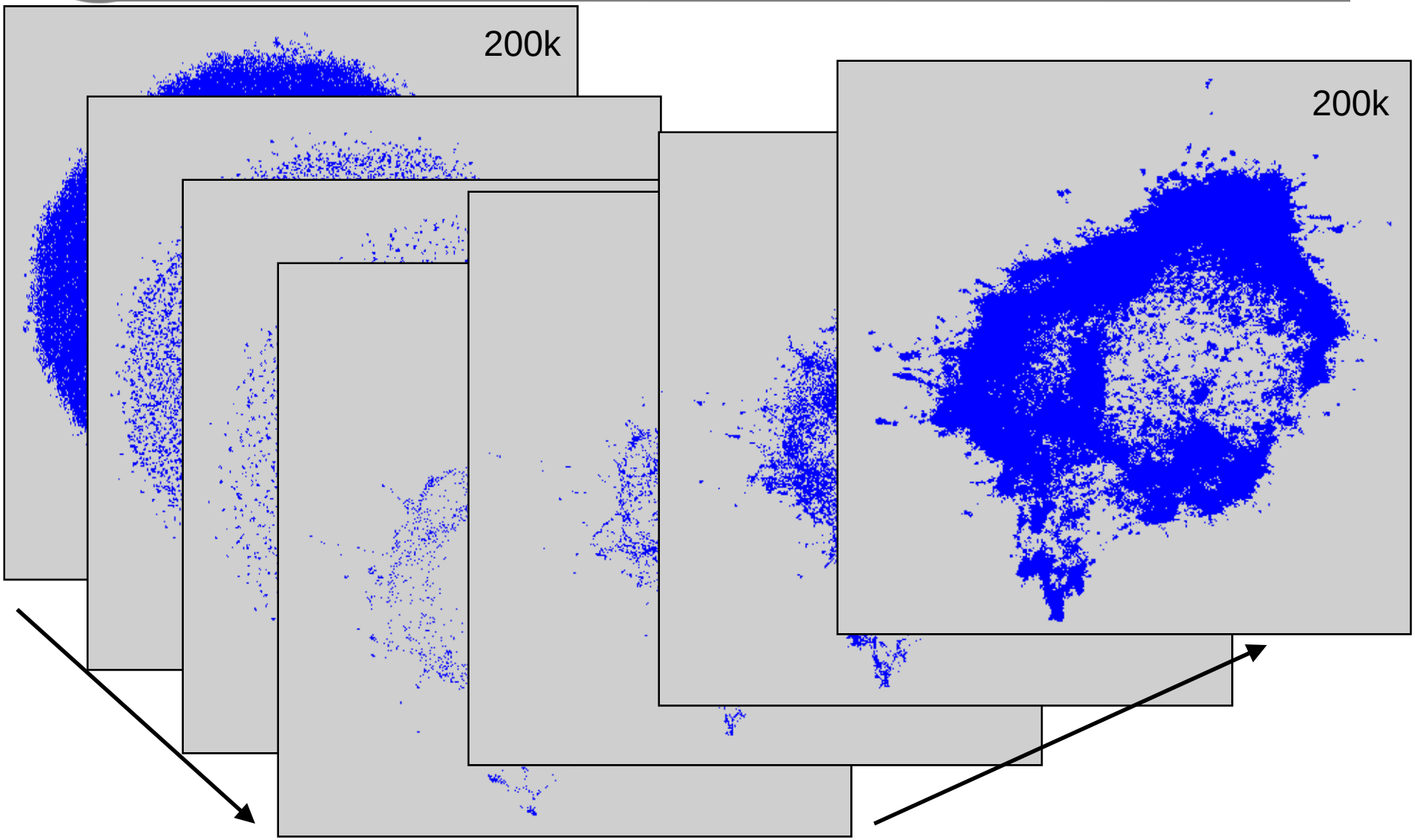
Vargas-Quesada, B. & Moya-Anegón, F. (2007). *Visualizing the Structure of Science*. Springer.



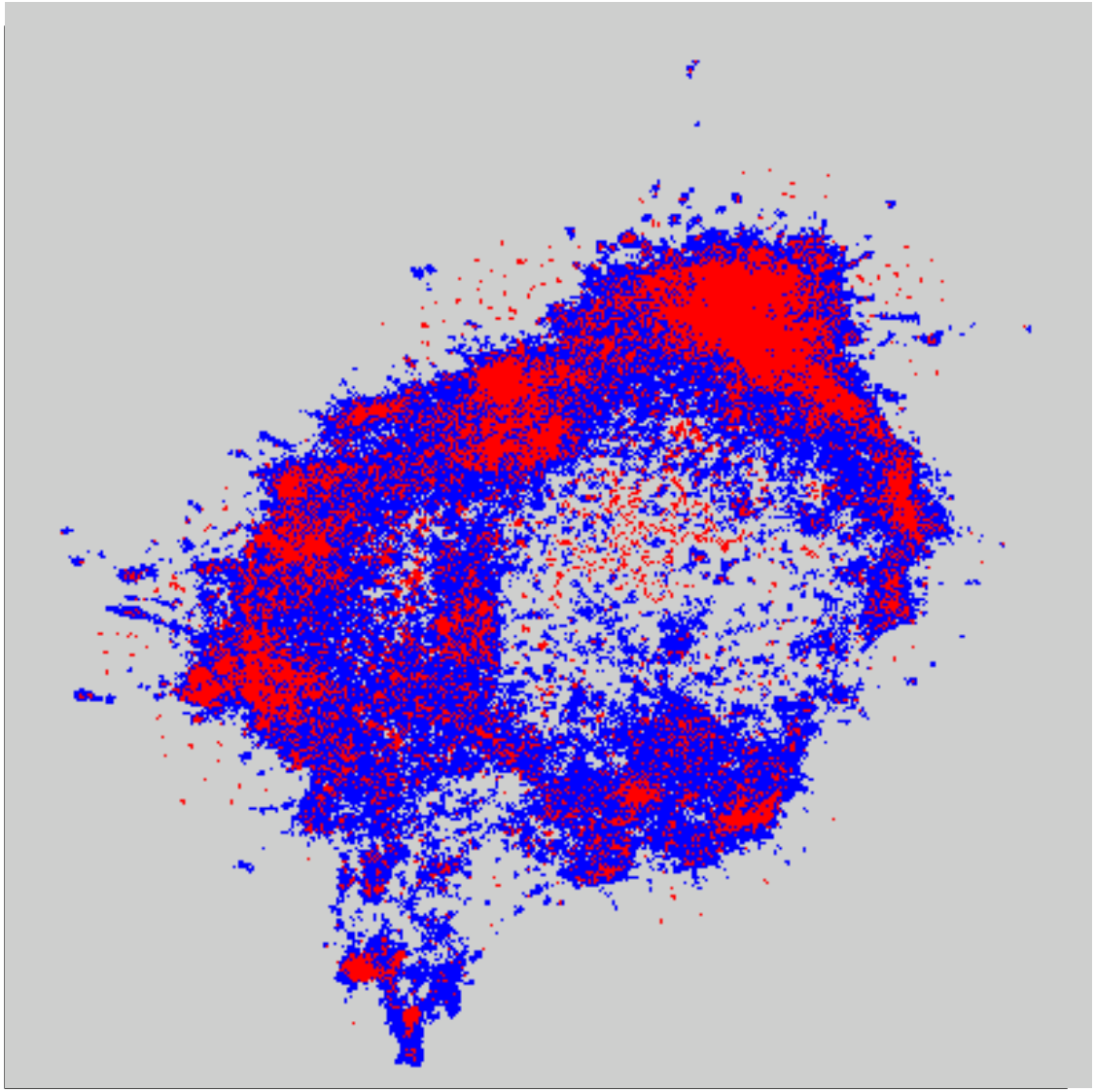
How can Network Science Help?

- Adopt methods for measuring accuracy
  - Yes… that means citing the information science literature !
- Improved dimension reduction algorithms
  - Including methods to test/compare accuracy of these algorithms (e.g. VAST)
- Improvements in force-directed algorithms
  - Scalability and speed
- Solution of the multi-point problem
  - Allow actors to have ‘n’ positions
- Euclidean space? or perhaps Riemann space?
  - Compare accuracy of solutions using different topological assumptions

Smart Coarsening



Real-time Clustering



These maps were constructed by sorting more than 16,000 journals into disciplines. Disciplines, represented as circles, are sets of journals that cite a common literature. Links (the lines between disciplines) are pairs of disciplines that share a common literature. Links are treated as rubber bands (attempting to bring two disciplines close to each other). The lack of a link is similar to a repulsive force (pushing two disciplines apart). A mathematical model, using these attractive-repulsive forces, generates the visualizations shown below directly from the data.

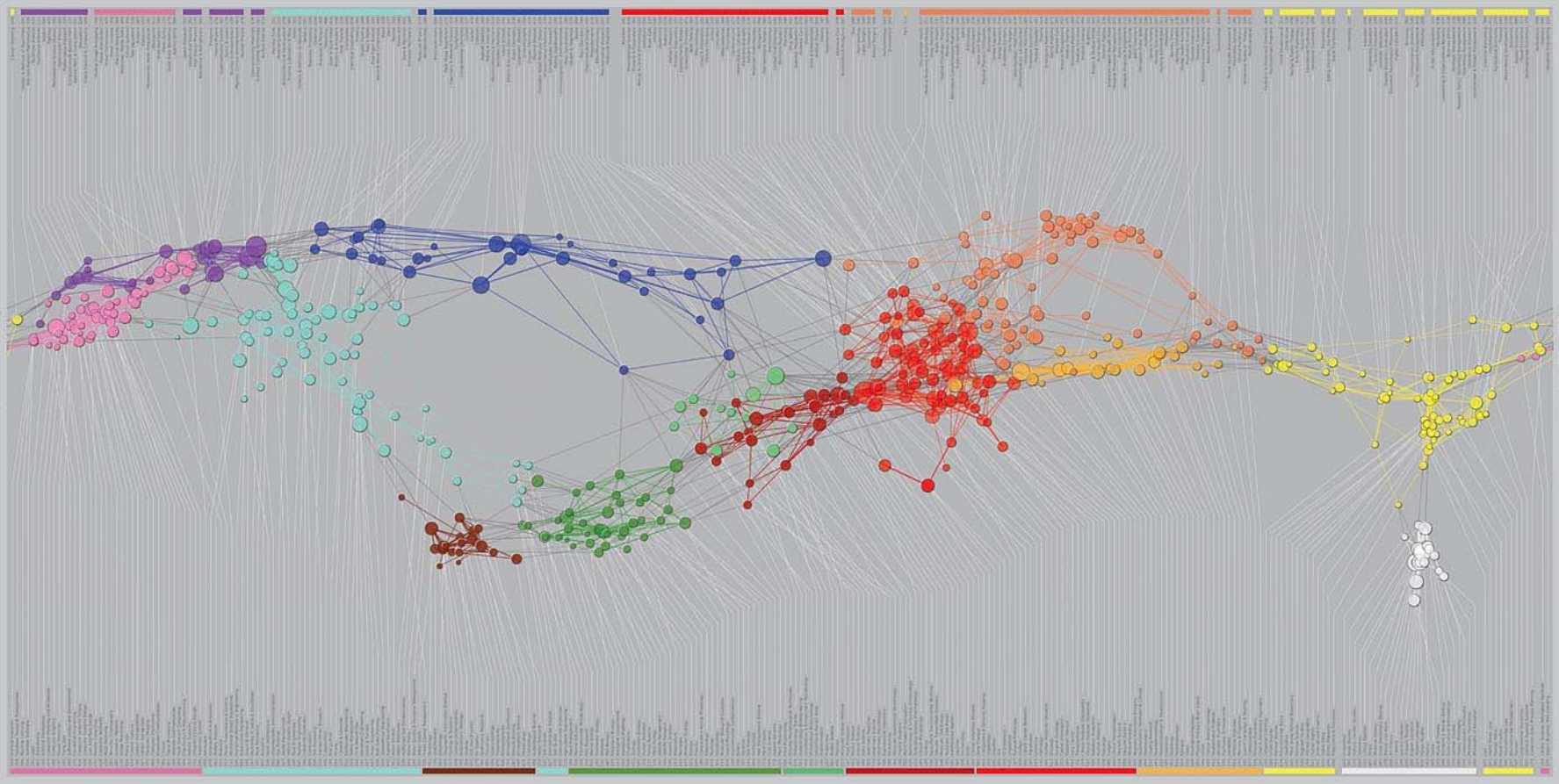
The maps on this poster represent a tradeoff between accuracy (shown on right) and coverage (shown below).

Maps of Science

Visualizing disciplines on the surface of a sphere result in the most accurate map of science. Five rotations are shown below.

A linescape projection (shown below) allows us to see the entire surface of the sphere.

A visualization of 7.2 million scholarly documents appearing in over 16,000 journals, proceedings or symposia between Jan, 2001 and Dec, 2005.



color key
- Math & Physics
- Electrical Engineering & Computer Science
- Chemical, Mechanical & Civil Engineering

color key
- Chemistry
- Biology
- Earth Sciences
- Biotechnology

color key
- Infectious Diseases
- Health Professionals
- Medical Specialties
- Brain Research

color key
- Social Sciences
- Humanities

Source: http://www.mapofscience.com/ Color images: © Regents of the University of California. The underlying data came from two sources: Thomson ISI and Scopus. Mapping methodology and descriptive text by Dick Klavans - President, SciTech Strategies, Inc. Graphics & typography by Ethan Mollick, Mike Pitsis, and Kevin Boyack. Special acknowledgements to Katy Borner, Kevin Boyack, Art Ellis, W. Brad Paley, Sam Small and Henry Small.

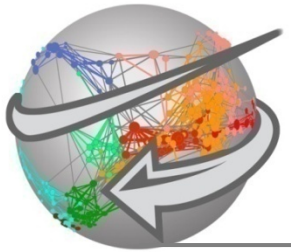
Places & Spaces: Mapping Science

a science exhibit that introduces people to maps of sciences, their makers and users.

Exhibit Curators:

Dr. Katy Börner &

Deborah MacPherson



Final Thoughts

- We are adamant about the accuracy of network visualization.
- We hope others will join us in this effort
  - The “community identification” work in physics and network science is part of this, but needs to go further in validation
- Thanks for listening !!