Visualizing Japanese Co-authorship Data

Gavin LaRowe & Katy Börner, Indiana University, USA
Ryutaro Ichise, National Institute of Informatics, Japan

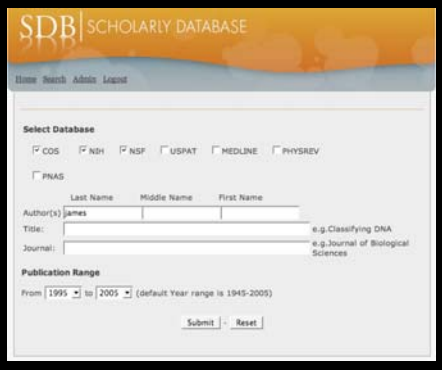
Information Visualisation Conference 2007
Zurich, Schweiz

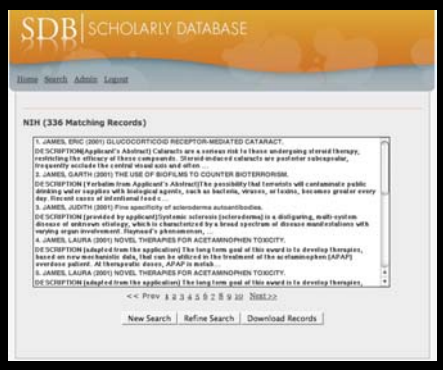
Motivation: Mapping Science

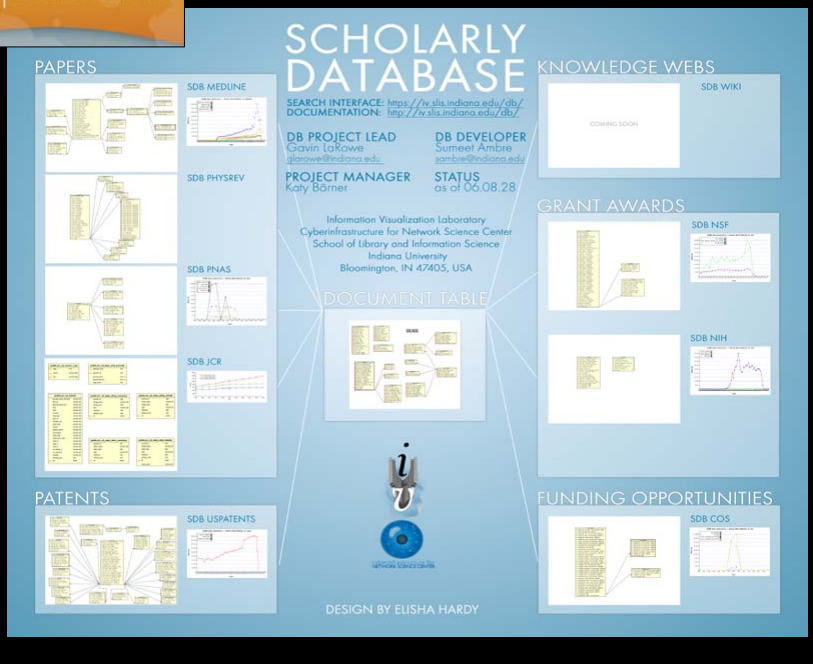
Places & Spaces: Mapping Science exhibit, see also http://scimaps.org.

Scholarly Database: Web Interface

Search across publications, patents, grants.
Download records and/or (evolving) co-author, paper-citation networks.
https://sdb.slis.indiana.edu/





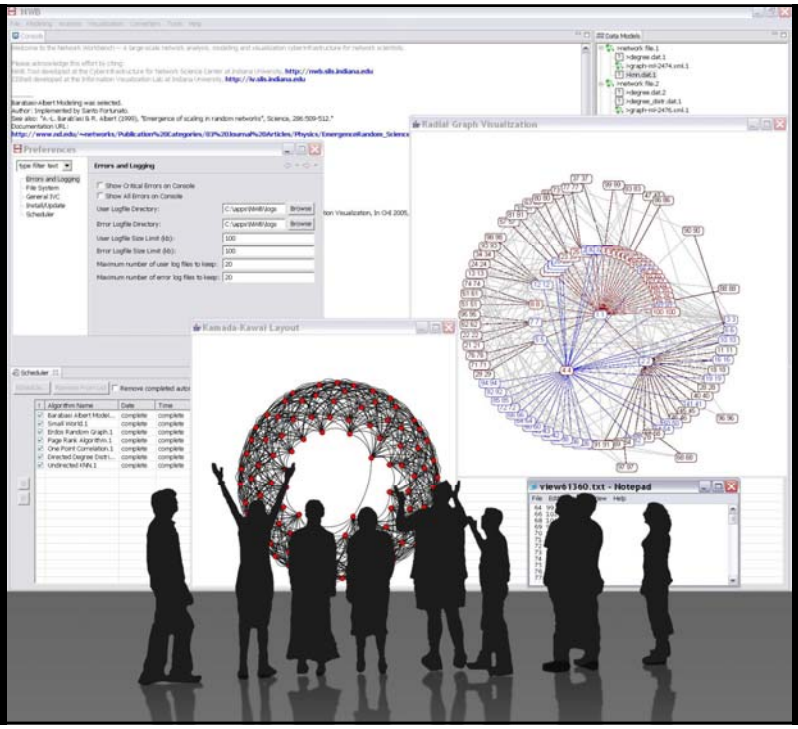


Scholarly Database: # Records & Years Covered

Datasets available via the Scholarly Database (* future feature)

Dataset	# Records	Years Covered	Updated	Restricted Access
Medline	13,149,741	1965-2005	Yes	
PhysRev	398,005	1893-2006		Yes
PNAS	16,167	1997-2002		Yes
JCR	59,078	1974, 1979, 1984, 1989 1994-2004		Yes
USPTO	3,179,930	1976-2004	Yes*	
NSF	174,835	1985-2003	Yes*	
NIH	1,043,804	1972-2002	Yes*	
Total	**18,021,560**	**1893-2006**	**4**	**3**

Aim for comprehensive geospatial and topic coverage.



Network Workbench (NWB)

Investigators: Katy Börner, Albert-Laszlo Barabasi, Santiago Schnell, Alessandro Vespignani & Stanley Wasserman, Eric Wernert

Software Team: Lead: Weixia (Bonnie) Huang

Developers: Bruce Herr, Ben Markines, Santo Fortunato, Cesar Hidalgo, Ramya Sabbineni, Vivek S. Thakre, & Russell Duhon

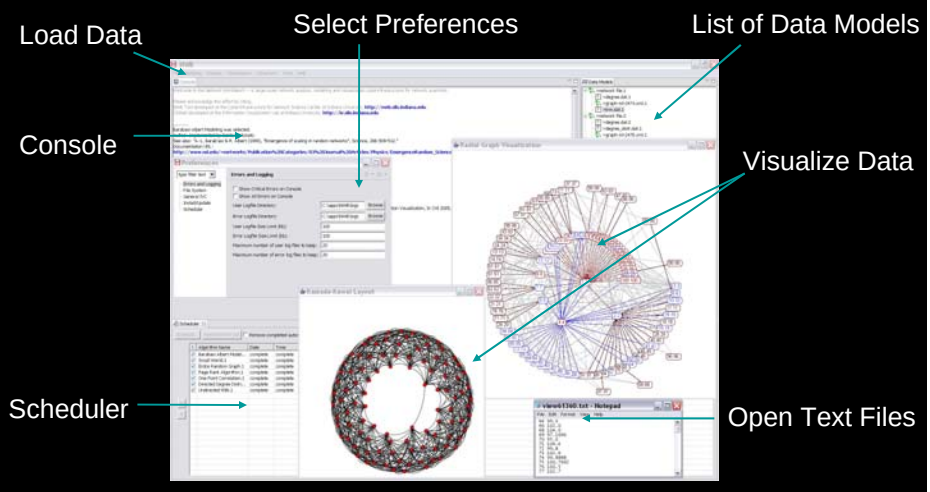
Goal: Develop a large-scale network analysis, modeling and visualization toolkit for biomedical, social science and physics research.

Amount: $1,120,926 NSF IIS-0513650 award.

Duration: Sept. 2005 - Aug. 2008

Website: http://nwb.slis.indiana.edu

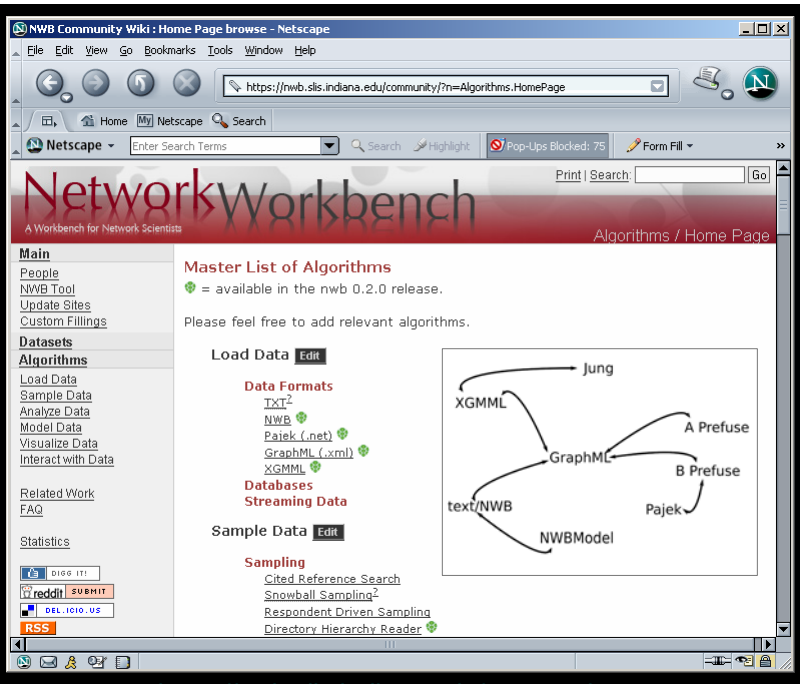
NWB Tool: Interface Elements



NWB Tool 0.2.0: List of Algorithms

Category	*Algorithm*	*Language*
Preprocessing	Directory Hierarchy Reader	JAVA
	Erdös-Rényi Random	FORTRAN
	Barabási-Albert Scale-Free	FORTRAN
Modeling	Watts-Strogatz Small World	FORTRAN
	Chord	JAVA
	CAN	JAVA
	Hypergrid	JAVA
	PRU	JAVA
	Tree Map	JAVA
	Tree Viz	JAVA
	Radial Tree / Graph	JAVA
Visualization	Kamada-Kawai	JAVA
	Force Directed	JAVA
	Spring	JAVA
	Fruchterman-Reingold	JAVA
	Circular	JAVA
	Parallel Coordinates (demo)	JAVA
Tool	XMGrace	

Analysis Algorithm	*Language*
Attack Tolerance	JAVA
Error Tolerance	JAVA
Betweenness Centrality	JAVA
Site Betweenness	FORTRAN
Average Shortest Path	FORTRAN
Connected Components	FORTRAN
Diameter	FORTRAN
Page Rank	FORTRAN
Shortest Path Distribution	FORTRAN
Watts-Strogatz Clustering Coefficient	FORTRAN
Watts-Strogatz Clustering Coefficient Versus Degree	FORTRAN
Directed k-Nearest Neighbor	FORTRAN
Undirected k-Nearest Neighbor	FORTRAN
Indegree Distribution	FORTRAN
Outdegree Distribution	FORTRAN
Node Indegree	FORTRAN
Node Outdegree	FORTRAN
One-point Degree Correlations	FORTRAN
Undirected Degree Distribution	FORTRAN
Node Degree	FORTRAN
k Random-Walk Search	JAVA
Random Breadth First Search	JAVA



https://nwb.slis.indiana.edu/community

Visualizing Japanese Co-authorship Data

Gavin LaRowe & Katy Börner, Indiana University, USA
Ryutaro Ichise, National Institute of Informatics, Japan

Information Visualisation Conference 2007
Zurich, Schweiz

Introduction

This paper reports a bilbiometric analysis of an evolving co-author network composed of 5,009 articles from Transactions D. Information Systems journal of the Institute of Electronics Information and Communication Engineers (IEICE) for the years 1993 to 2005.

Networks from this data set were subsequently generated, producing metrics used for further analysis. We were particularly interested in whether the characteristics of these networks were similar or different than those of often-cited networks found in popular literature regarding co-authorship networks for other scientific disciplines.

Prior Research

Most of the prior research regarding co-authorship networks in Japanese literature was performed during the mid-1990s by public policy analysts focusing on academic collaboration.

Recent studies by Professor Ichise and others have looked at co-authorship networks in the context of data mining and information visualization.

Other studies in Japan have used co-authorship networks as a mechanism to study the effect conferences play in initiating and sustaining collaborations between researchers.

Method

Data

- Provider: National Institute of Informatics, Tokyo, Japan
- Years: 1993 - 2005
- Institute of Electronics Information and Communication Engineers - Japanese analogue to IEEE

- Four main journals:
  - A. Fundamentals
  - B. Communications
  - C. Electronics
  - D. Information Systems
- 12,337 articles
- 5,009 unique authors

Method

Data Processing

- Transformation: converted initial data from EUC_JP to UTF-8
- For each year, unique authors extracted using Japanese surnames. Custom scripts used to lean/identify/disambiguate names.
- Data status: < 3% transcription errors. Identifiable errors were cleaned manually.
- Data parsed into individual lexemes and proper names
- Data placed into relational database
- Functions in database used to build network tables in Pajek format
- R used to generate time-series metrics

IEICE Co-authorship Networks

Metrics

Network	*Size/AU*	*AR*	*AC*	*DC*	*CC*	*BC*	*DE*	*DI*	$\langle k \rangle$	l	C
1993	861	402	2.9	0.4599	0.00116	0.03019	0.00053	2	1.208	0.46	0.562
1994	758	377	2.87	0.5646	0.00132	0.00660	0.00053	2	1.045	0.57	0.6777
1995	733	327	2.77	0.3329	0.00136	0.00682	0.00075	2	1.076	0.33	0.6034
1996	900	406	2.94	0.3778	0.00111	0.00889	0.00045	2	1.086	0.37	0.5878
1997	1127	491	3.12	0.3762	0.00089	0.00887	0.00042	2	1.086	0.37	0.6912
1998	1125	466	3.24	0.3733	0.00089	0.00356	0.00033	2	1.037	0.37	0.6462
1999	995	415	2.95	0.2573	0.00100	0.00302	0.00033	2	1.045	0.26	0.6263
2000	1144	452	3.68	0.4336	0.00087	0.04808	0.00026	3	1.329	0.43	0.5745
2001	1194	454	3.66	0.3317	0.00084	0.00838	0.00038	2	1.092	0.33	0.6807
2002	681	257	3.14	0.2702	0.00147	0.05286	0.00028	3	1.514	0.27	0.6316
2003	722	343	3.16	0.3047	0.00139	0.00970	0.00040	2	1.113	0.3	0.6667
2004	924	276	3.63	0.4159	0.00108	0.00541	0.00042	2	1.05	0.42	0.6687
2005	1173	343	3.75	0.3444	0.00085	0.00597	0.00029	2	1.065	0.34	0.6866
Total	12,337	5,009	3.00	1.2900	0.00013	5.33230	0.00017	15	1.29	4.3	0.505
SPIRES	56,627								173	4.0	0.726
NCSTRL	11,994								3.59	9.7	0.496
Math	70,975								3.90	9.5	0.590
Neurosci.	209,293								11.5	6.0	0.760

Table 1: Metrics per year and totals: No. nodes (*Size*), no. articles (AR), avg. no. of collaborations per author per year (AC), *degree centrality (DC), *closeness centrality (CC), *betweenness centrality (BT), density (DE), diameter (DI), average degree ($\langle k \rangle$), average path length(l), and clustering co-efficient (C); * = mean value indicated. Other metrics taken from p. 8 of Albert and Barabási's paper, *Statistical Mechanics of Complex Networks*, in Reviews of Modern Physics, 74, 47 (2002).

Analysis Results

- We computed centrality measures such as degree, closeness, betweenness as well as distributions for centrality data for each year and plotted using a q-q plot to identify significant changes. Clustering coefficient and average path length were also generated for each year.

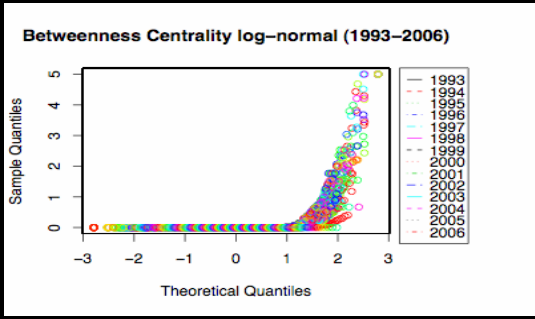
- Degree distribution does not deviate from other popular co-authorship networks; fat-tail distribution.
- Changes in coauthorship pattern or paradigm almost always reflected in clustering coefficient and average path length.
- No significant increases in average no. of co-authors, etc.

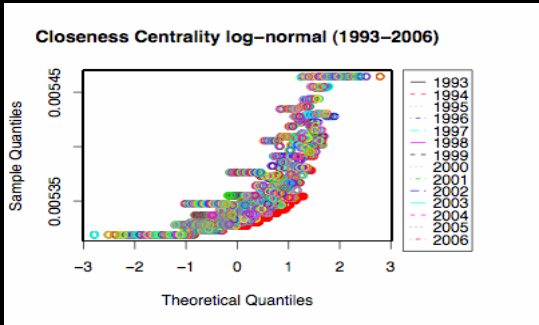
Analysis Results

Q-q plots for betweenness and closeness centrality computed for years 1993-2005.

No significant deviation for any one year.

Quantile distributions could also have been used.

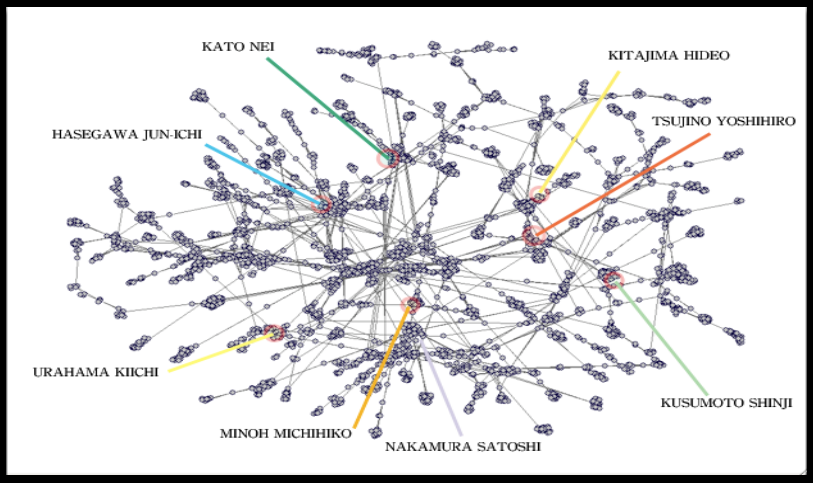




Largest Connected Component

Transactions D. (1993-2005): 3,961 nodes showing top eight collaborators.

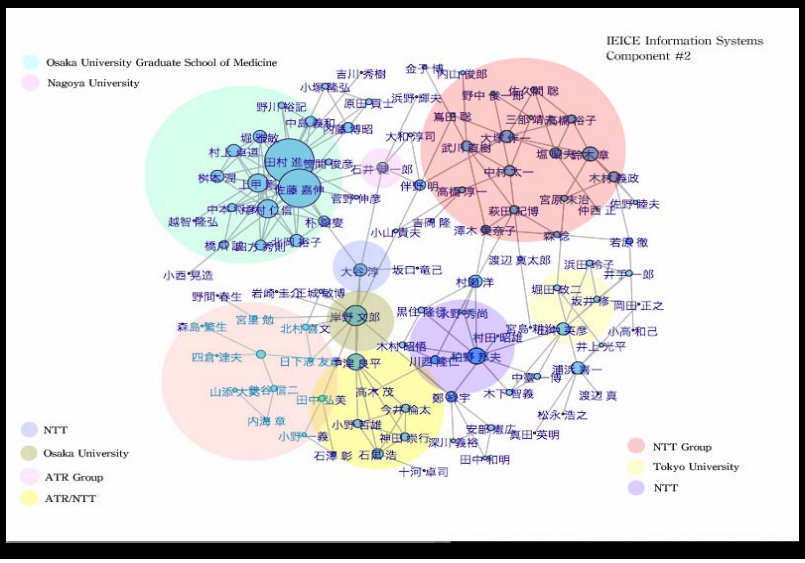
12,337 articles 5,009 authors



Largest Component #2: IEICE Transactions D. (1993-2005)

*Ellipses indicate general affiliation.

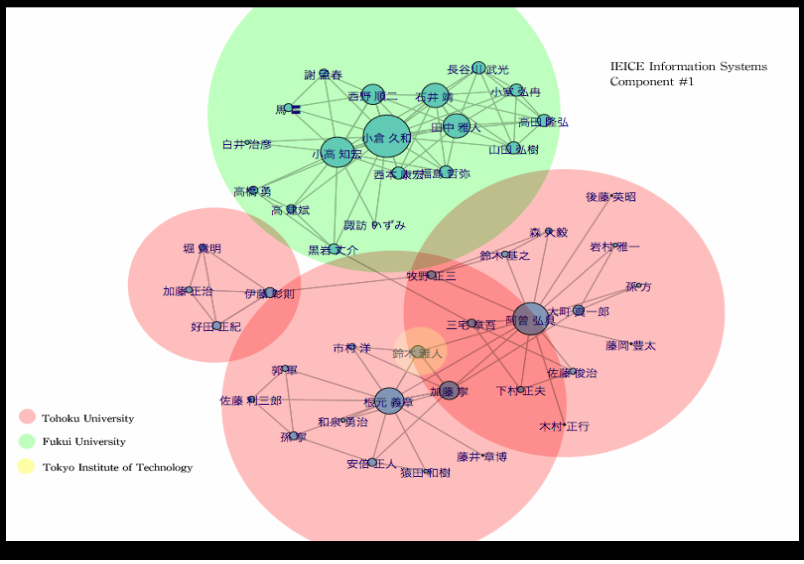
12,337 articles 5,009 authors



Largest Component #1: IEICE Transactions D. (1993-2005)

*Ellipses indicate general affiliation.

12,337 articles 5,009 authors



Conclusions

- IEICE Transactions D. network is very similar to SPIRES and other co-authorship data.
- Average path length and clustering coefficient similar, again pointing out the significance of the degree distribution in regard to other metrics.
- $P(k) \alpha k = 2.216$ (power-law network)
- Scale-free behavior (small-world network)

Acknowledgements

We’d like to thank the National Institute of Informatics, Tokyo, Japan for funding this work by a MOU grant and for providing the data used in this study.