Science of Science Research and Tools
Tutorial #09 of 12

Dr. Katy Börner
Cyberinfrastructure for Network Science Center, Director
Information Visualization Laboratory, Director
School of Library and Information Science
Indiana University, Bloomington, IN
http://info.slis.indiana.edu/~katy

With special thanks to Kevin W. Boyack, Micah Linnemeier, Russell J. Duhon, Patrick Phillips, Joseph Biberstine, Chintan Tank Nianli Ma, Hanning Guo, Mark A. Price, Angela M. Zoss, and Scott Weingart

Invited by Robin M. Wagner, Ph.D., M.S.
Chief Reporting Branch, Division of Information Services
Office of Research Information Systems, Office of Extramural Research
Office of the Director, National Institutes of Health

Suite 4090, 6705 Rockledge Drive, Bethesda, MD 20892
10a-noon, July 21, 2010

12 Tutorials in 12 Days at NIH—Overview

1. Science of Science Research — **1st Week**
2. Information Visualization
3. CIShell Powered Tools: Network Workbench and Science of Science Tool

4. Temporal Analysis—Burst Detection — **2nd Week**
5. Geospatial Analysis and Mapping
6. Topical Analysis & Mapping

7. Tree Analysis and Visualization — **3rd Week**
8. Network Analysis
9. **Large Network Analysis**

10. **Using the Scholarly Database at IU** — **4th Week**
11. **VIVO National Researcher Networking**
12. **Future Developments**

12 Tutorials in 12 Days at NIH—Overview

[#09] Large Network Analysis and Visualization

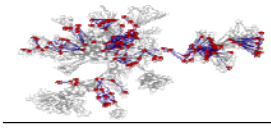
- General Overview
- Designing Effective Network Visualizations
- Sci2-Reading and Modeling Networks
- Sci2-Analysing Large Networks
- Sci2-Visualizing Large Networks and Distributions
- Outlook
- Exercise: Identify Promising Large Network Analyses of NIH Data

Recommended Reading

- NWB Team (2009) Network Workbench Tool, User Manual 1.0.0, http://nwb.slis.indiana.edu/Docs/NWBTool-Manual.pdf
- Börner, Katy, Sanyal, Soma and Vespignani, Alessandro (2007). Network Science. In Blaise Cronin (Ed.), ARIST, Information Today, Inc./American Society for Information Science and Technology, Medford, NJ, Volume 41, Chapter 12, pp. 537-607. http://ivl.slis.indiana.edu/km/pub/2007-borner-arist.pdf

[#09] Large Network Analysis and Visualization

- General Overview
- Designing Effective Network Visualizations
- Sci2-Reading and Modeling Networks
- Sci2-Analysing Large Networks
- Sci2-Visualizing Large Networks and Distributions
- Outlook
- Exercise: Identify Promising Large Network Analyses of NIH Data



Large Networks

- More than 10,000 nodes.
- Neither all nodes nor all edges can be shown at once. Sometimes, there are more nodes than pixels.

Examples of large networks

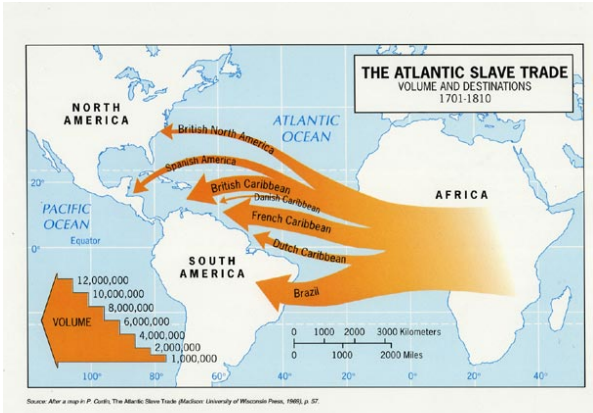
- Communication networks:

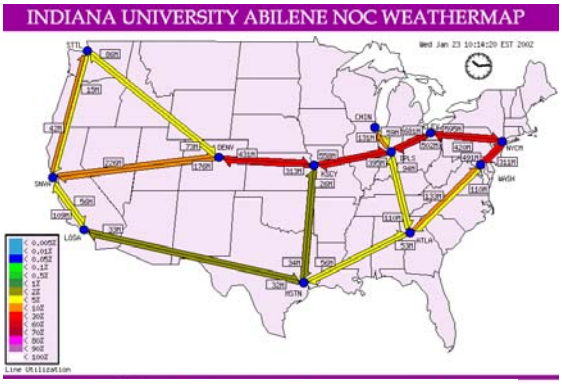
  Internet, telephone network, wireless network.
- Network applications:

  The World Wide Web, Email interactions
- Transportation network/road maps
- Relationships between objects in a data base:

  Function/module dependency graphs

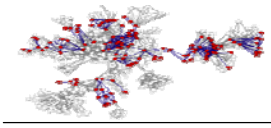
  Knowledge bases





http://loadrunner.uits.iu.edu/weathermaps/abilene/

Amsterdam RealTime project, WIRED Magazine, Issue 11.03 - March 2003

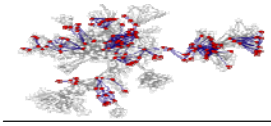


Direct Manipulation

Modify focusing parameters while continuously provide visual feedback and update display (fast computer response).

- Conditioning: filter, set background variables and display foreground parameters
- Identification: highlight, color, shape code
- Parameter control: line thickness, length, color legend, time slider, and animation control
- Navigation: Bird's Eye view, zoom, and pan
- Information requests: Mouse over or click on a node to retrieve more details or collapse/expand a subnetwork

See NIH Awards Viewer at http://scimaps.org/maps/nih/2007/

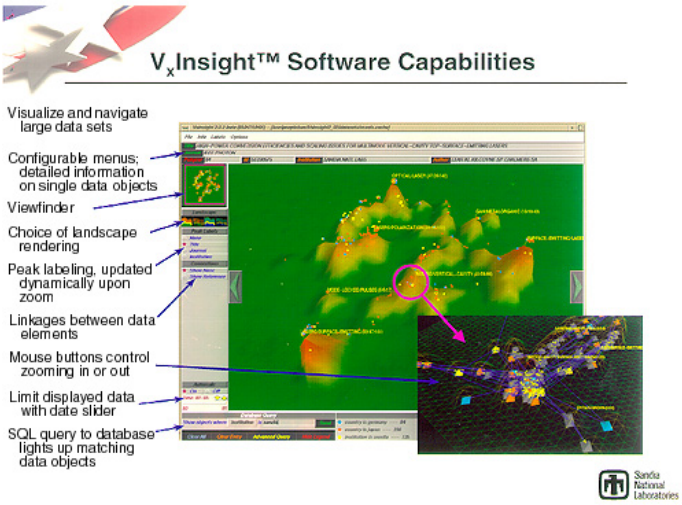


VxInsight Tool

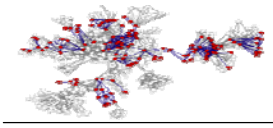
VxInsight is a general purpose knowledge visualization software package developed at Sandia National Laboratories.

It enables researchers, analysts, and decision-makers to accelerate their understanding of large databases.

Show **Insight_demo.exe**



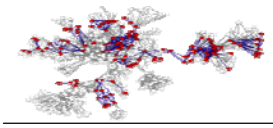
Davidson, G.S., Hendrickson, B., Johnson, D.K., Meyers, C.E., Wylie, B.N., November/December 1998. "Knowledge Mining with VxInsight: Discovery through Interaction," Volume 11, Number 3, Journal of Intelligent Information Systems, Special Issue on Integrating Artificial Intelligence and Database Technologies. pp.259-285.)



Other Tools

Tool	Year	Domain	Description	Open Source	Operating System	References
S&T Dynam. Toolbox	1985	Scientom.	Tools from Loet Leydesdorff for organization analysis, and visualization of scholarly data.	No	Windows	(Leydesdorff , 2008)
In Flow	1987	SocSci	Social network analysis software for organizations with support for what-if analysis.	No	Windows	(Krebs, 2008)
Pajek	1996	SocSci	A network analysis and visualization program with many analysis algorithms, particularly for social network analysis.	No	Windows	(Batagelj & Mrvar, 1998)
BibExcel	2000	Scientom	Transforms bibliographic data into forms usable in Excel, Pajek, NetDraw, and other programs.	No	Windows	(Persson, 2008)
Boost Graph Library	2000	CS	Extremely efficient and flexible C++ library for extremely large networks.	Yes	All Major	(Siek et al., 2002)
UCINet	2000	SocSci	Social network analysis software particularly useful for exploratory analysis.	No	Windows	(Borgatti et al., 2002)
Visone	2001	SocSci	Social network analysis tool for research and teaching, with a focus on innovative and advanced visual methods.	No	All Major	(Brandes & Wagner, 2008)
Cytoscape	2002	Bio	Network visualization and analysis tool focusing on biological networks, with particularly nice visualizations.	Yes	All Major	(Cytoscape-Consortium, 2008)

See http://ivl.slis.indiana.edu/km/pub/2010-borner-et-al-nwb.pdf for references.



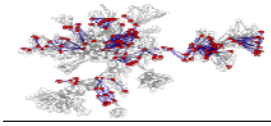
Other Tools cont.

Tool	Year	Domain	Description	Open Source	Operating System	References
GeoVISTA	2002	Geo	GIS software that can be used to lay out networks on geospatial substrates.	Yes	All Major	(Takatsuka & Gahegan, 2002)
iGraph	2003	CS	A library for classic and cutting edge network analysis usable with many programming languages.	Yes	All Major	(Csárdi & Nepusz, 2006)
Tulip	2003	CS	Graph visualization software for networks over 1,000, 000 elements.	Yes	All Major	(Auber, 2003)
CiteSpace	2004	Scientom	A tool to analyze and visualize scientific literature, particularly co-citation structures.	Yes	All Major	(Chen, 2006)
GraphViz	2004	Networks	Flexible graph visualization software.	Yes	All Major	(AT&T-Research-Group, 2008)
Hittite	2004	Scientom	Analysis and visualization tool for data from the Web of Science.	No	Windows	(Garfield, 2008)
R	2004	Statistics	A statistical computing language with many libraries for sophisticated network analyses.	Yes	All Major	(Ihaka & Gentleman, 1996)
Prefuse	2005	Visualiz.	A general visualization framework with many capabilities to support network visualization and analysis.	Yes	All Major	(Heer et al., 2005)
NWB Tool	2006	Bio, IS, SocSci, Scientom	Network analysis & visualization tool conducive to new algorithms supportive of many data formats.	Yes	All Major	(Huang, 2007)

See http://ivl.slis.indiana.edu/km/pub/2010-borner-et-al-nwb.pdf for references.

[#09] Large Network Analysis and Visualization

- General Overview
- Designing Effective Network Visualizations
- Sci2-Reading and Modeling Networks
- Sci2-Analysing Large Networks
- Sci2-Visualizing Large Networks and Distributions
- Outlook
- Exercise: Identify Promising Large Network Analyses of NIH Data



QVR Dataset Provided by Robert F. Moore, Deepshikha Roychowdhury, Emilee Pressman, and Matthew Eblen

All NIH projects that received funding in 1998-2009 (Oct 1, 1997-Sept 30, 2009) and their associated publications (max 100 per project so that SAS can handle the data. Note that some projects had 5000+ publications! We do miss much data here.)

168,764 grant records collapsed by base project.

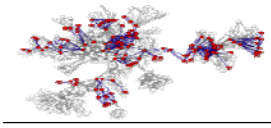
119,230 grants have a linked publications (pubid).

There are 157,376 unique publications.

pubid	proj	proj1	ADMII	first_year	ACTIVITY_	last_year
9485464;9096302	C06CA058690	CCA058690	CA	1994	C06	1995
20527532;8858722;20427856;2(	C06CA059267	CCA059267	CA	1992	C06	1995
16913728;16362150	C06RR011192	CRR011192	RR	1996	C06	1998
16698792;16534782;17518562;1	C06RR012088	CRR012088	RR	1996	C06	1996
9714740	C06RR012176	CRR012176	RR	1996	C06	1996
19248166;18071382;18838156;1	C06RR012463	CRR012463	RR	1997	C06	1997
15345738;11994348;12586855;1	C06RR012488	CRR012488	RR	1997	C06	1997

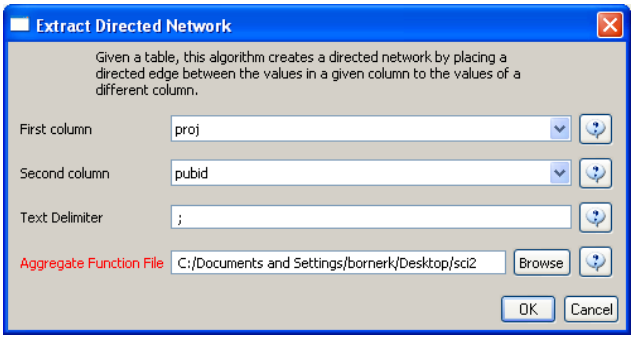
Three (planned) analyses:

1. Large network visualization of 119k grants to 157k pubs network to show the scalability.
2. Horizontal Bar Graph visualization of all NIH grants. *(need $ amounts)*
3. UCSD science map of publications for different institutes. *(need journal name)*



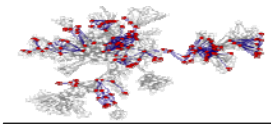
QVR Dataset – Large network visualization of 119k grants to 157k pubs network to show the scalability of the tool.

1. In original data file, delete all grants that have no associated publications.
2. Load resulting using 'File > Load > QVR-Bob-119239Grants.csv SAS-grants-pubs-simplified.csv' as csv file format.
3. Extract author bipartite grant to publications network using 'Data Preparation > Text Files > Extract Directed Network' using parameters:



Number of Publications by Institute
Based on NIH awards receiving funding betwee FY 1998 and FY 2009

IC	IC_ABBR	publications
NIH Total	NIH	947,903
AA	NIAAA	17,773
AG	NIA	47,087
AI	NIAID	100,092
AR	NIAMS	30,354
AT	NCCAM	3,291
CA	NCI	155,132
DA	NIDA	37,881
DC	NIDCD	19,130
DE	NIDCR	18,625
DK	NIDDK	96,295
EB	NIBIB	10,744
ES	NIEHS	24,472
EY	NEI	38,960
GM	NIGMS	152,378
HD	NICHD	55,104
HG	NHGRI	6,140
HL	NHLBI	130,150
LM	NLM	3,771
MD	NCMHD	1,430
MH	NIMH	63,363
NR	NINR	4,736
NS	NINDS	84,075
OD	OD	261
RG	CSR	3
RR	NCRR	39,873
TW	FIC	5,553



SAS Dataset Provided by Lindsey Pool

62,864 records, one per publication.

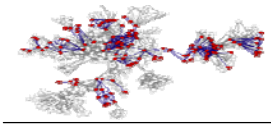
Replace missing values by NULL to load into Sci2 Tool

Load using 'File > Load > SAS-grants-pubs-simplified.csv' as csv file format.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Appl_Id	Inst_Zip	Grant_Sta	Grant_En	Grant_Titl	Grant_Abs	RCDC_Categories	Pub_Yr	Pub_Auth	Pub_Jour	PMID	Pub_title	Pub_abst	All_MeSH	
7527025	11203-	1-Aug-79	31-Jul-13	Brain Dysf	DESCRIPT	Alcoholism; Behaviora	1992	Zaninelli, R	Alcoholism	1558305	The Tridim	Cloninger I	Adult;Alcoholism	
7527025	11203-	1-Aug-79	31-Jul-13	Brain Dysf	DESCRIPT	Alcoholism; Behaviora	1991	Cohen, H L	Alcoholism	1755520	EEG chara	Baseline E	Adult;Alcoholism	
7527025	11203-	1-Aug-79	31-Jul-13	Brain Dysf	DESCRIPT	Alcoholism; Behaviora	1990	Porjesz, B	Alcohol (F	2222850	Event-relat	Visual evel	Adult;Alcoholism	
7527025	11203-	1-Aug-79	31-Jul-13	Brain Dysf	DESCRIPT	Alcoholism; Behaviora	1987	Porjesz, B	Electroenc	2431876	The N2 cor	The latenc	Adult;Alcoholism	
7527025	11203-	1-Aug-79	31-Jul-13	Brain Dysf	DESCRIPT	Alcoholism; Behaviora	1987	Brecher, M	Electroenc	2435516	The N2 cor	Event-relat	Adult;Brain;Elect	
7527025	11203-	1-Aug-79	31-Jul-13	Brain Dysf	DESCRIPT	Alcoholism; Behaviora	1988	Begleiter, I	Alcoholism	3056069	Potential biological m		Alcoholism;Biolo	
7527025	11203-	1-Aug-79	31-Jul-13	Brain Dysf	DESCRIPT	Alcoholism; Behaviora	1987	Brecher, M	Biological	3607113	Late positi	Abstinent	Adult;Alcoholism	
7527025	11203-	1-Aug-79	31-Jul-13	Brain Dysf	DESCRIPT	Alcoholism; Behaviora	1987	Begleiter, I	Alcohol (F	3620101	Auditory re	We have p	Adolescent;Alcol	
7527025	11203-	1-Aug-79	31-Jul-13	Brain Dysf	DESCRIPT	Alcoholism; Behaviora	1982	Porjesz, B	Alcoholism	6280509	Evoked brain potentia		Adult;Aged;Aging	
7527025	11203-	1-Aug-79	31-Jul-13	Brain Dysf	DESCRIPT	Alcoholism; Behaviora	1983	Begleiter, I	Psychophy	6828618	P3 and stimulus incer		Adult;Brain;Elect	

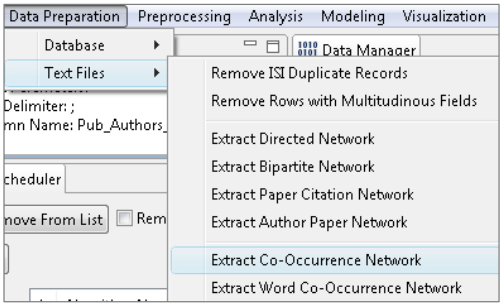
If you run out of Java heap space: Load using 'File > Load > SAS-grants-pubs-4columns.csv'

	A	B	C	D
1	Appl_Id	Pub_Authors_All	Pub_Journal	PMID
2	7527025	Zaninelli, R M; Porjesz, B; Begleiter, H	Alcoholism, clinical and experimental research	1558305
3	7527025	Cohen, H L; Porjesz, B; Begleiter, H	Alcoholism, clinical and experimental research	1755520
4	7527025	Porjesz, B; Begleiter, H	Alcohol (Fayetteville, N.Y.)	2222850
5	7527025	Porjesz, B; Begleiter, H; Bihari, B; Kissin, B	Electroencephalography and clinical neurophysiology	2431876
6	7527025	Brecher, M; Porjesz, B; Begleiter, H	Electroencephalography and clinical neurophysiology	2435516
7	7527025	Begleiter, H; Porjesz, B	Alcoholism, clinical and experimental research	3056069
8	7527025	Brecher, M; Porjesz, B; Begleiter, H	Biological psychiatry	3607113
9	7527025	Begleiter, H; Porjesz, B; Rawlings, R; Eckardt, M	Alcohol (Fayetteville, N.Y.)	3620101
10	7527025	Porjesz, B; Begleiter, H	Alcoholism, clinical and experimental research	6280509
11	7527025	Begleiter, H; Porjesz, B; Chou, C L; Aunon, J I	Psychophysiology	6828618
12	7527025	Brecher, M; Begleiter, H	Biological psychiatry	6871300
13	7527025	Begleiter, H; Porjesz, B; Tenner, M	Acta psychiatrica Scandinavica. Supplementum	6935921

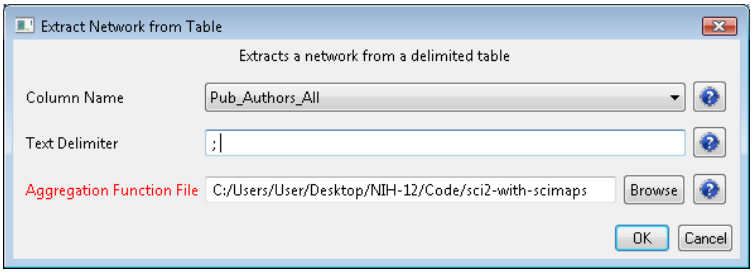


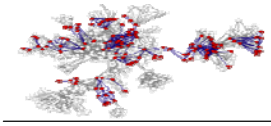
SAS Dataset – Extract Co-Author Network

Extract author co-occurrence network using 'Data Preparation > Text Files > Extract Co-Occurrence Network'



With parameters: *(ignore the Aggregate Function File but note the space after ;)*

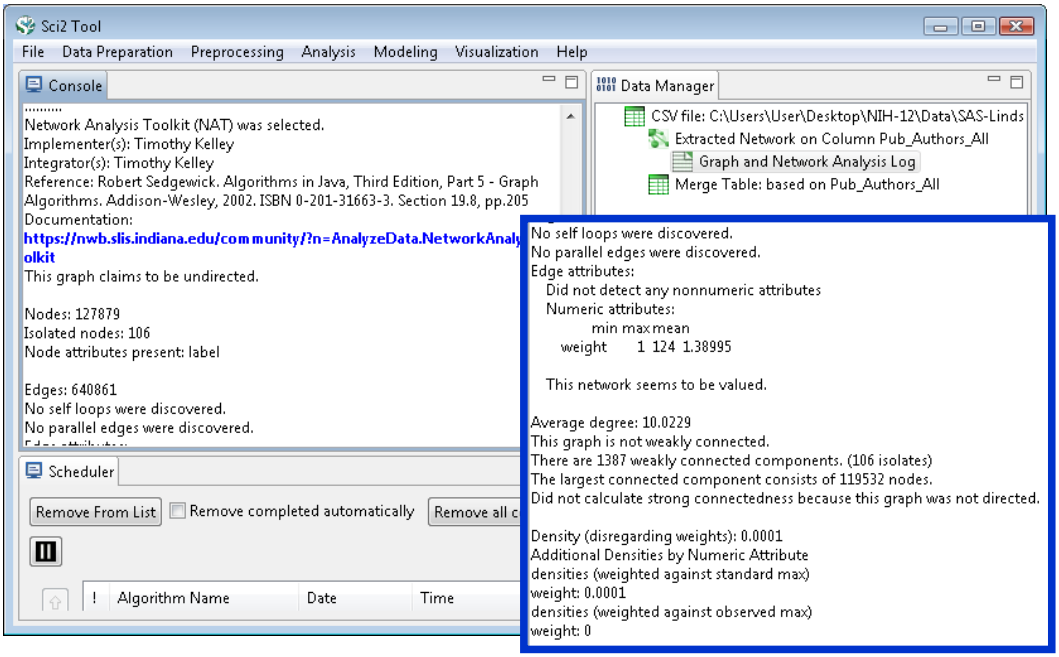




SAS Dataset – Extract Co-Author Network cont.

Nodes: 127,879 authors

Edges: 640,861 co-author relationships



[#09] Large Network Analysis and Visualization

- General Overview
- Designing Effective Network Visualizations
- Sci2-Reading and Modeling Networks
- Sci2-Analysing Large Networks
- Sci2-Visualizing Large Networks and Distributions
- Outlook
- Exercise: Identify Promising Large Network Analyses of NIH Data

Modeling the Co-Evolving Author-Paper Networks

Börner, Katy, Maru, Jeegar & Goldstone, Robert. (2004). The Simultaneous Evolution of Author and Paper Networks. PNAS. Vol. 101(Suppl. 1), 5266-5273.

The TARL Model (Topics, Aging, and Recursive Linking) incorporates

- A partitioning of authors and papers into topics,
- Aging, i.e., a bias for authors to cite recent papers, and
- A tendency for authors to cite papers cited by papers that they have read resulting in a rich get richer effect.

The model attempts to capture the roles of authors and papers in the production, storage, and dissemination of knowledge.

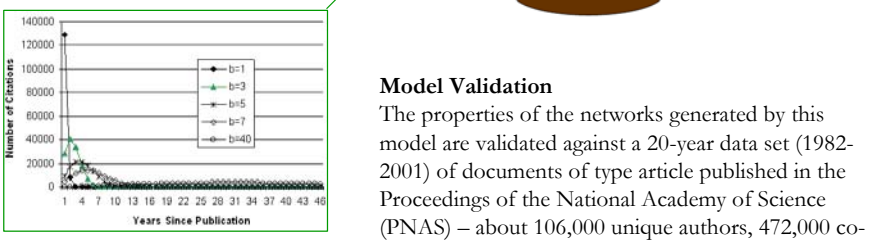
Model Assumptions

- Co-author and paper-citation networks co-evolve.
- Authors come and go.
- Papers are forever.
- Only authors that are 'alive' are able to co-author.
- All existing (but no future) papers can be cited.
- Information diffusion occurs directly via co-authorships and indirectly via the consumption of other authors' papers.

- Preferential attachment is modeled as an *emergent property* of the elementary, local networking activity of authors reading and citing papers, but also the references listed in papers.

```
-------------------------------------------
Model Parameters (0=without, 1=with)
-------------------------------------------
0/1  Topics
0/1  Co-Authors
0/1  Consider References
0    Aging Function
-------------------------------------------
Model Initialization Values
-------------------------------------------
2  # Years
5  # Authors in Start Year
5  # Papers in Start Year
2  # Papers Consumed (Referenced) per Paper
1  # Papers Produced per Author each Year
5  # Topics
1  # Co-Author(s) per Author
1   # Levels References are Considered
```

```
// Initialization
generate #_papers papers and assign a random topic to each paper;
generate #_authors authors and assign a random topic to each author;
randomly assign #_co-authors+1 authors to papers of the same topic;
// Simulation
for each year do {
    add #_new_authors new authors, deactivate authors older than #_author_age;
    for each topic do {
        randomly partition set of authors into author_groups of size #_co-authors+1;
        for each author_group do {
            for each new_paper to be produced, do {
                generate new_paper;
                randomly select #_read_ papers from existing papers;
                get all references of read_ papers up to #_reference_path_length;
                for each new_paper_reference do {
                    select a time_slice from (start year to curr_year-1) with probability given in aging_function;
                    randomly select a paper published or cited in this time_slice; as a new_paper_reference;
                    add the new_paper_reference to new_paper;
                }
            }
        }
    }
    add all new papers to the set of existing papers;
    add new links to author and paper information;
}
```



Aging function

Model Validation

The properties of the networks generated by this model are validated against a 20-year data set (1982-2001) of documents of type article published in the Proceedings of the National Academy of Science (PNAS) – about 106,000 unique authors, 472,000 co-author links, 45,120 papers cited within the set, and 114,000 citation references within the set.

Table 3 Statistics for SIM data

Year	#p	#a	#r	#c	a#ca
1981	1624	3953	0	756	8.21
1982	1040	5200	31200	112161	4
1983	1118	5590	33540	21397	4
1984	1197	5985	35910	10224	4
1985	1275	6375	38250	6184	4
1986	1353	6765	40590	4687	4
1987	1432	7160	42960	3573	4
1988	1510	7550	45300	2816	4
1989	1589	7945	47670	2219	4
1990	1667	8335	50010	1853	4
1991	1745	8725	52350	1634	4
1992	1824	9120	54720	1431	4
1993	1902	9510	57060	1167	4
1994	1981	9905	59430	1040	4
1995	2059	10295	61770	767	4
1996	2137	10685	64110	632	4
1997	2216	11080	66480	522	4
1998	2294	11470	68820	400	4
1999	2373	11865	71190	265	4
2000	2451	12255	73530	125	4
2001	2529	12645	75870	0	4
Total	**37316**		**1070760**	**173853**	

Table 2. PNAS Statistics

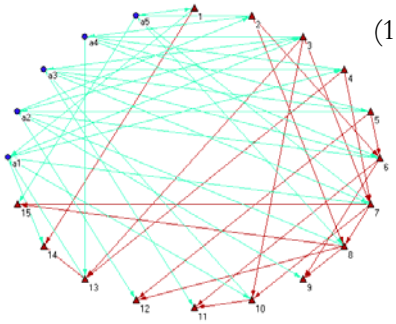
Year	#p	#a	#r	#c	a#ca
1982	1669	5201	46665	156690	3.92
1983	1611	5142	46685	161437	3.98
1984	1695	5583	49834	174161	4.22
1985	1846	6325	55662	191750	4.38
1986	2042	7209	64379	218229	4.76
1987	1924	7061	59110	207729	4.88
1988	2035	7471	63116	215227	4.8
1989	2088	7959	65883	215437	5.01
1990	2066	8031	66019	207138	5.15
1991	2382	9559	77740	223102	5.25
1992	2500	9812	80949	211238	5.29
1993	2413	9770	79848	193867	5.55
1994	2600	10656	86176	187353	5.56
1995	2476	10429	82021	151249	5.66
1996	2765	11803	99061	148622	5.96
1997	2618	11255	96788	122908	6.12
1998	2711	12328	100973	107764	6.48
1999	2603	12182	97018	76080	6.69
2000	2501	12201	94181	44131	7.6
2001	2575	13038	97450	16357	8.5
Total	**45120**		**1509558**	**3230469**	

The TARL Model: Pseudo Code

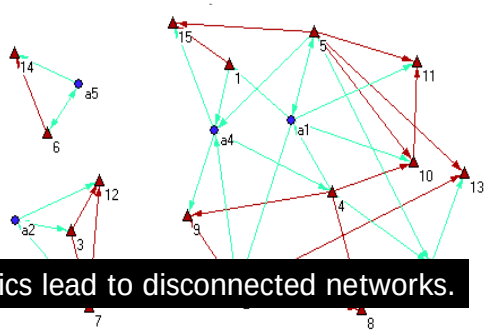
```
// Initialization
generate #_papers papers and assign a random topic to each paper;
generate #_authors authors and assign a random topic to each author;
randomly assign #_co-authors+1 authors to papers of the same topic;
// Simulation
for each year do {
    add #_new_authors new authors, deactivate authors older than #_author_age;
    for each topic do {
        randomly partition set of authors into author_groups of size #_co-authors+1;
        for each author_group do {
            for each new_paper to be produced, do {
                generate new_paper;
                randomly select #_read_ papers from existing papers;
                get all references of read_ papers up to #_reference_path_length;
                for each new_paper_reference do {
                    select a time_slice from (start year to curr_year-1) with probability given in aging_function;
                    randomly select a paper published or cited in this time_slice; as a new_paper_reference;
                    add the new_paper_reference to new_paper;
                }
            }
        }
    }
    add all new papers to the set of existing papers;
    add new links to author and paper information;
}
```

The TARL Model: The Effect of Parameters

(0000)

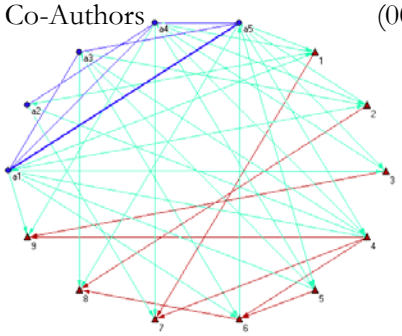


(1000) Topics

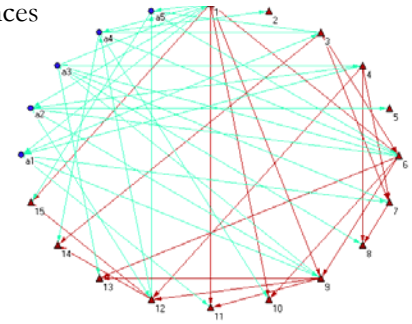


Topics lead to disconnected networks.

(0100) Co-Authors



(0010) References



Co-authoring leads to fewer papers.


```
----------------------------------------
Model Parameters (0=without, 1=with)
----------------------------------------
0/1  Topics
0/1  Co-Authors
0/1  Consider References
0    Aging Function
----------------------------------------
Model Initialization Values
----------------------------------------
2  # Years
5  # Authors in Start Year
5  # Papers in Start Year
2  # Papers Consumed (Referenced) per Paper
1  # Papers Produced per Author each Year
5  # Topics
1  # Co-Author(s) per Author
1   # Levels References are Considered
```

```
// Initialization
generate #_papers papers and assign a random topic to each paper;
generate #_authors authors and assign a random topic to each author;
randomly assign #_co-authors+1 authors to papers of the same topic;
// Simulation
for each year do {
    add #_new_authors new authors, deactivate authors older than #_author_age;
    for each topic do {
        randomly partition set of authors into author_groups of size #_co-authors+1;
        for each author_group do {
            for each new_paper to be produced, do {
                generate new_paper;
                randomly select #_read_papers from existing papers;
                get all references of read_papers up to #_reference_path_length;
                for each new_paper_reference do {
                    select a time_slice from (start year to curr_year-1) with probability given in aging_function;
                    randomly select a paper published or cited in this time_slice; as a new_paper_reference;
                    add the new_paper_reference to new_paper;
                }
            }
        }
    }
    add all new papers to the set of existing papers;
    add new links to author and paper information;
}
```

Input Script

Model

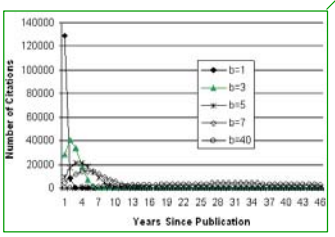
Simulated Networks

Model Validation

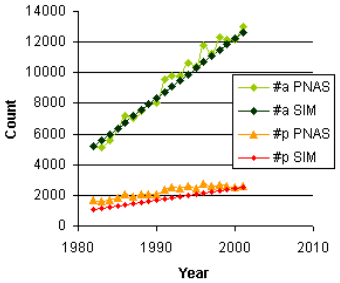
PNAS Data Set

Simple Statistics
Network Properties
N, <k>, l, C , γ

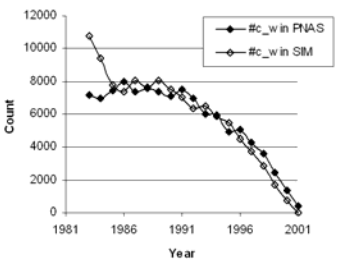
Aging function

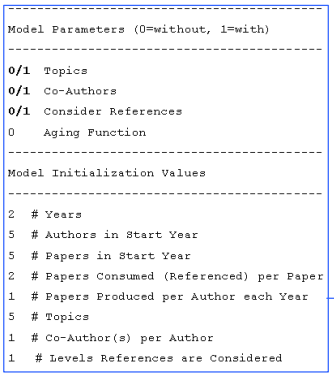


Counts for Papers and Authors



Counts for Citations





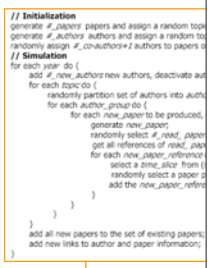
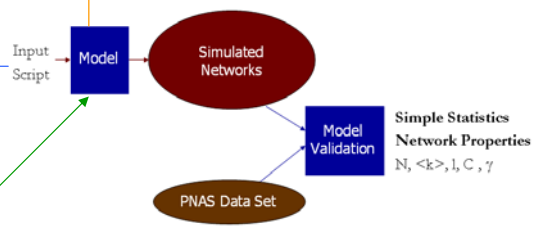


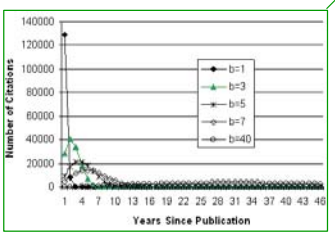
Table 2. Properties of co-author & paper citation networks comprising number of nodes n, average node degree <k>, path length l, cluster coefficient C, and power law exponent γ. Source references are given in the left column.

Network	n	<k>	l	C	γ	Reference
Co-authorship networks						
LANL	52,909	9.7	5.9	0.43	--	Newman, (2001a; 2001b; 2001c)
MEDLINE	1,520,251	18.1	4.6	0.066	--	
SPIRES	56,627	1.73	4.0	0.726	1.2	
NCSTRL	11,994	3.59	9.7	0.496	--	
Math.	70,975	3.9	9.5	0.59	2.5	Barabasi et al., (2002)
Neurosci.	209,293	11.5	6	0.76	2.1	
PNAS	105,915	8.97	5.89	0.399	2.54	
Paper-citation networks						
ISI	783,339	8.57	--	--	3	Redner, (1998)
PhysRev	24,296	14.5	--	--	3	
PNAS	45,120	3.53	--	0.081	2.29	
SIM	37,114	2.13	--	0.074	2.05	

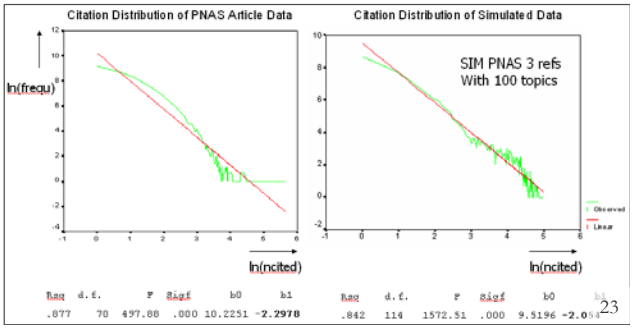
Co-Author and Paper-Citation Network Properties

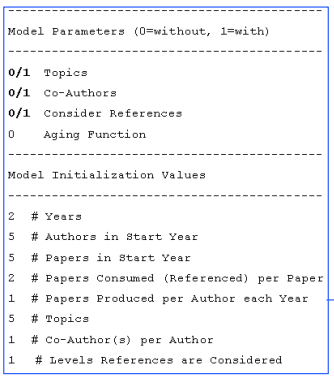


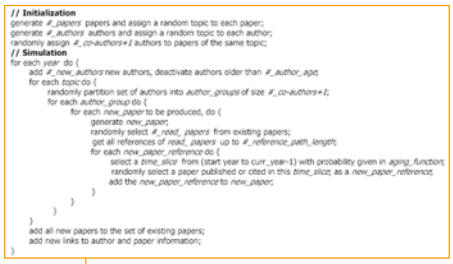
Aging function

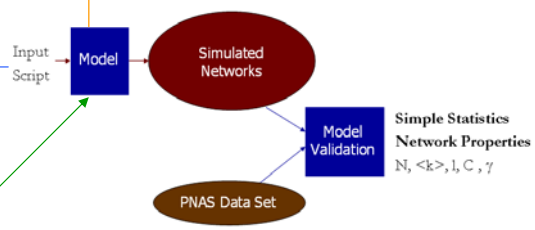


Power Law Distributions

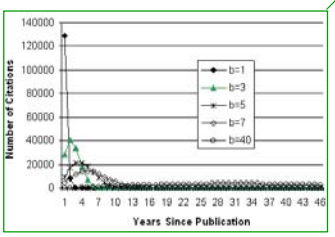








Aging function



Topics: The number of topics is linearly correlated with the clustering coefficient of the resulting network: C= 0.000073 * #topics. Increasing the number of topics increases the power law exponent as authors are now restricted to cite papers in their own topics area.

Aging: With increasing b, and hence increasing the number of older papers cited as references, the clustering coefficient decreases. Papers are not only clustered by topic, but also in time, and as a community becomes increasingly nearsighted in terms of their citation practices, the degree of temporal clustering increases.

References/Recursive Linking: The length of the chain of paper citation links that is followed to select references for a new paper also influences the clustering coefficient. Temporal clustering is ameliorated by the practice of citing (and hopefully reading!) the papers that were the earlier inspirations for read papers.

[#09] Large Network Analysis and Visualization

- General Overview
- Designing Effective Network Visualizations
- Sci2-Reading and Modeling Networks
- Sci2-Analyzing Large Networks
- Sci2-Visualizing Large Networks and Distributions
- Outlook
- Exercise: Identify Promising Large Network Analyses of NIH Data

Network Analysis and Visualization – General Workflow

Original Data

	A	B
1	Source Node	Target Nodes
2	A	1;2;3
3	B	3;4
4	C	2;3
5	D	1

	A	B
1	Source Node	Target Nodes
2	A	1
3	A	2
4	A	3
5	B	3
6	B	4
7	C	2
8	C	3
9	D	1

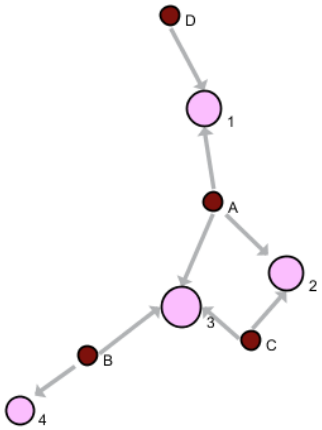
Extract Network

Extract Bipartite Network was selected.
Input Parameters:
First column: Source Node
Text Delimiter: ;
Second column: Target Nodes

Calculate Node Attributes

```
*Nodes
id*int label*string  bipartitetype*string indegree*int outdegree*int
1     "A" "Source Node" 0 3
2     "3" "Target Nodes" 3 0
3     "2" "Target Nodes" 2 0
4     "1" "Target Nodes" 2 0
5     "B" "Source Node" 0 2
6     "4" "Target Nodes" 1 0
7     "C" "Source Node" 0 2
8     "D" "Source Node" 0 1
*DirectedEdges
source*int target*int
1     2
1     4
1     3
5     6
5     2
7     2
7     3
8     4
```

Visualization/Layout



Large Network Analysis & Visualization – General Workflow

Original Data

Millions of records, in 100s of columns.

SAS and Excel might not be able to handle these files.

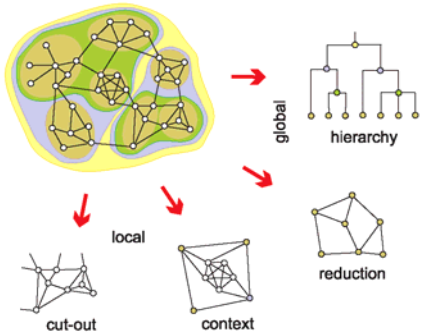
Files are shared between DB and tools as delimited text files (.csv).

Extract Network

It might take several hours to extract a network on a laptop or even on a parallel cluster.

Derived Statistics

Degree distributions
Number of components and their sizes
Extract giant component, subnetworks for further analysis



Visualizations

It is typically not possible to layout the network.
DrL scales to 10 million nodes.

[#09] Large Network Analysis and Visualization

- General Overview
- Designing Effective Network Visualizations
- Sci2-Reading and Modeling Networks
- Sci2-Analysing Large Networks
- Sci2-Visualizing Large Networks and Distributions
- Outlook
- Exercise: Identify Promising Large Network Analyses of NIH Data

DrL Large Network Layout

See Section 4.9.4.2 in Sci2 Tutorial,
http://sci.slis.indiana.edu/registration/docs/Sci2_Tutorial.pdf

DrL is a force - directed graph layout toolbox for real - world large - scale graphs up to 2 million nodes. It includes:

- Standard force - directed layout of graphs using algorithm based on the popular VxOrd routine (used in the VxInsight program).
- Parallel version of force - directed layout algorithm.
- Recursive multilevel version for obtaining better layouts of very large graphs.
- Ability to add new vertices to a previously drawn graph.

The version of DrL included in Sci2 only does the standard force - directed layout (no recursive or parallel computation).

Davidson, G. S., B. N. Wylie and K. W. Boyack (2001). "Cluster stability and the use of noise in interpretation of clustering." Proc. IEEE Information Visualization 2001: 23-30.

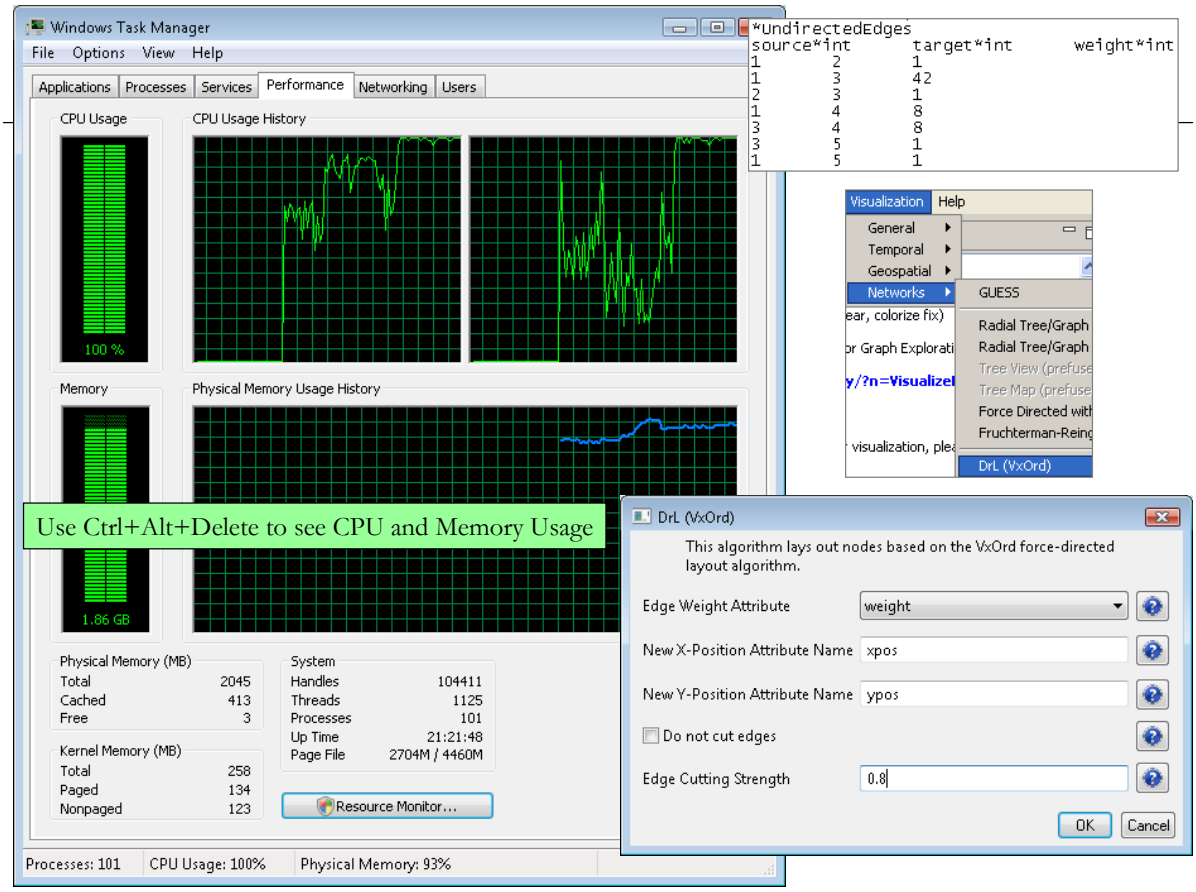
DrL Large Network Layout

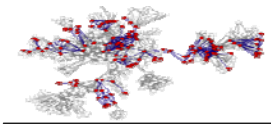
See Section 4.9.4.2 in Sci2 Tutorial,
http://sci.slis.indiana.edu/registration/docs/Sci2_Tutorial.pdf

How to use: DrL expects the edges to be *weighted* and *undirected* where the non - zero weight denotes how similar the two nodes are (higher is more similar). Parameters are as follows:

- The **edge cutting parameter** expresses how much automatic edge cutting should be done. 0 means as little as possible, 1 as much as possible. Around .8 is a good value to use.
- The **weight attribute parameter** lets you choose which edge attribute in the network corresponds to the similarity weight. The X and Y parameters let you choose the attribute names to be used in the returned network which corresponds to the X and Y coordinates computed by the layout algorithm for the nodes.

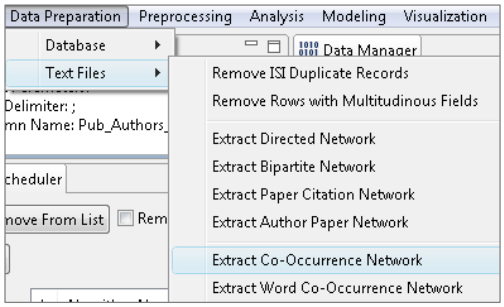
DrL is commonly used to layout large networks, e.g., those derived in co - citation and co - word analyses. In the Sci2 Tool, the results can be viewed in either GUESS or *'Visualization > Specified (prefuse alpha)'.*
See also https://nwb.slis.indiana.edu/community/?n=VisualizeData.DrL



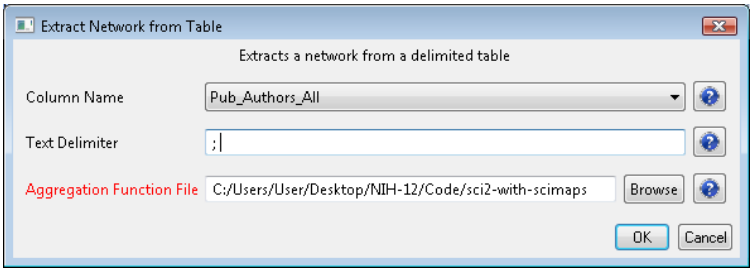


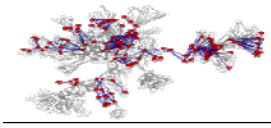
SAS Dataset – Extract Co-Author Network

Extract author co-occurrence network using 'Data Preparation > Text Files > Extract Co-Occurrence Network'



With parameters: *(ignore the Aggregate Function File but note the space after ;)*

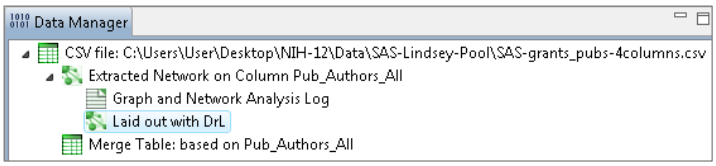




DrL Run & Output

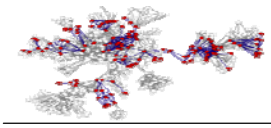
DrL (VxOrd) was selected.
Author(s): S. Martin, W. M. Brown, K. Boyack
Implementer(s): S. Martin, W. M. Brown, K. Boyack
Integrator(s): Bruce Herr
Reference: S. Martin, W. M. Brown, K. Boyack, "Dr. L: Distributed Recursive (Graph) Layout," in preparation for Journal of Graph Algorithms and Applications. (**http://citeseer.ist.psu.edu/davidson01cluster.html)**
Documentation: **https://nwb.slis.indiana.edu/community/?n=VisualizeData.DrL**

Input Parameters:
Edge Cutting Strength: 0.8
New X-Position Attribute Name: xpos
Edge Weight Attribute: weight
Do not cut edges: false
New Y-Position Attribute Name: ypos



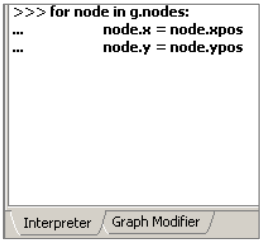
Entering liquid stage .....
Liquid stage completed in 317 seconds, total energy = 9.55681e+013.
Entering expansion stage .....
Finished expansion stage in 324 seconds, total energy = 4.29353e+009.
Entering cool-down stage .....
Completed cool-down stage in 321 seconds, total energy = 1.33472e+009.
Entering crunch stage .....
Finished crunch stage in 79 seconds, total energy = 1.49297e+009.
Entering simmer stage .....
Finished simmer stage in 98 seconds, total energy = 22.5252.
Layout calculation completed in 1139 seconds (not including I/O).
Writing out solution to inFile.icoord ...
Total Energy: 22.4969.
Program terminated successfully.

```
*Nodes
id*int  label*string    xpos*real       ypos*real
1       "Begleiter, H"  252.803 385.732
2       "Zaninelli, R M"259.491 394.488
3       "Porjesz, B"    253.609 385.524
4       "Cohen, H L"    252.678 387.526
5       "Kissin, B"     259.513 381.005
6       "Bihari, B"     255.578 392.549
7       "Brecher, M"    253.478 380.96
8       "Eckardt, M"    260.747 376.859
```

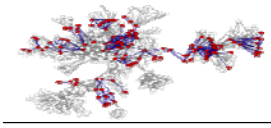
DrL Output Visualization

I saved file as SAS-Co-Author-DrL-Layout.nwb. Visualize network using GUESS by selecting the network file, running 'Network > Visualizing > GUESS', then run the following commands in the GUESS Interpreter:



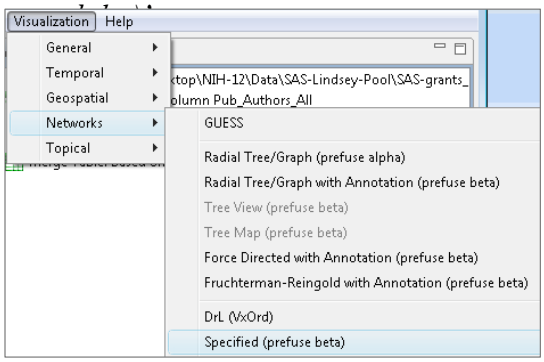
> for node in g.nodes:
... node.x = node.xpos
... node.y = node.ypos

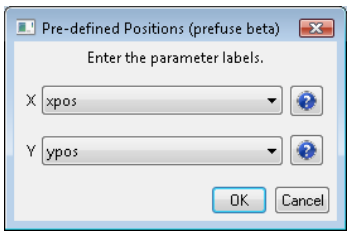
to position the nodes at the x and y position calculated by DrL.

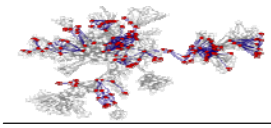


DrL Output Visualization

Visualize SAS-Co-Author-DrL-Layout.nwb using *'Visualization > Specified (prefuse*

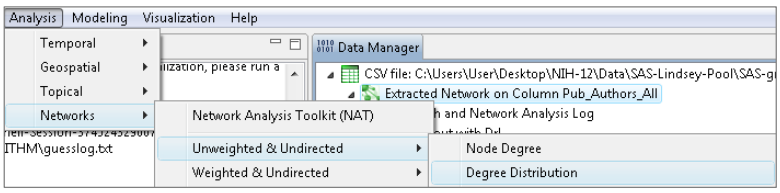


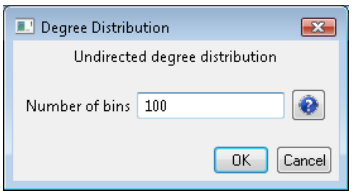


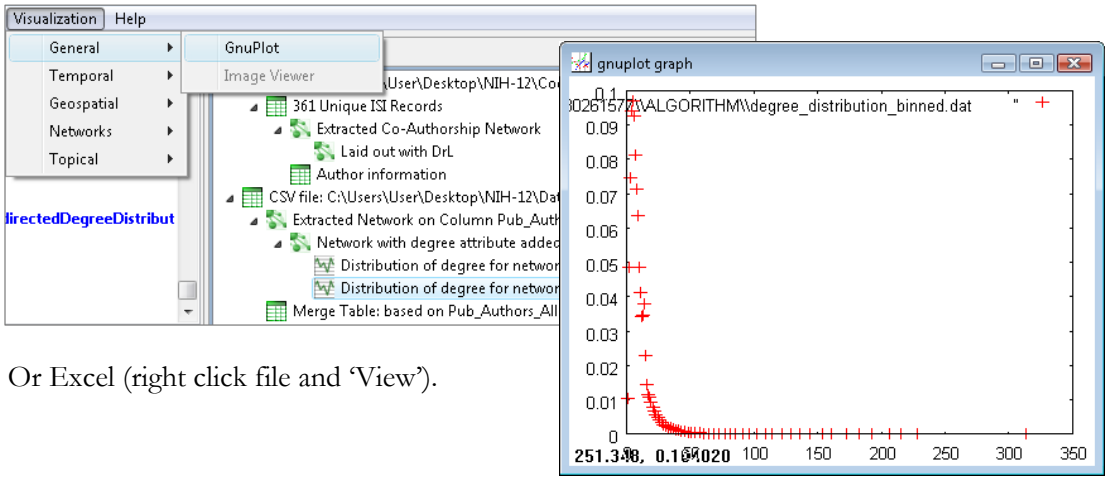


DrL Output – Plot Node Degree Distribution

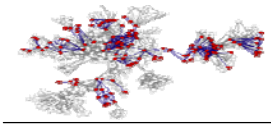
Calculate degree distribution using and plot using 'Visualization > General > Gnuplot'





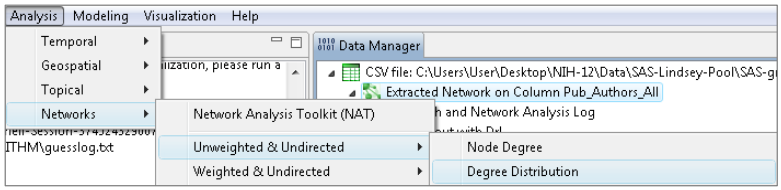


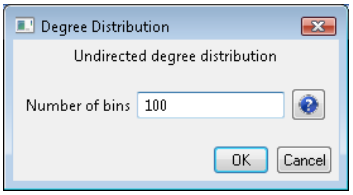
Or Excel (right click file and 'View').

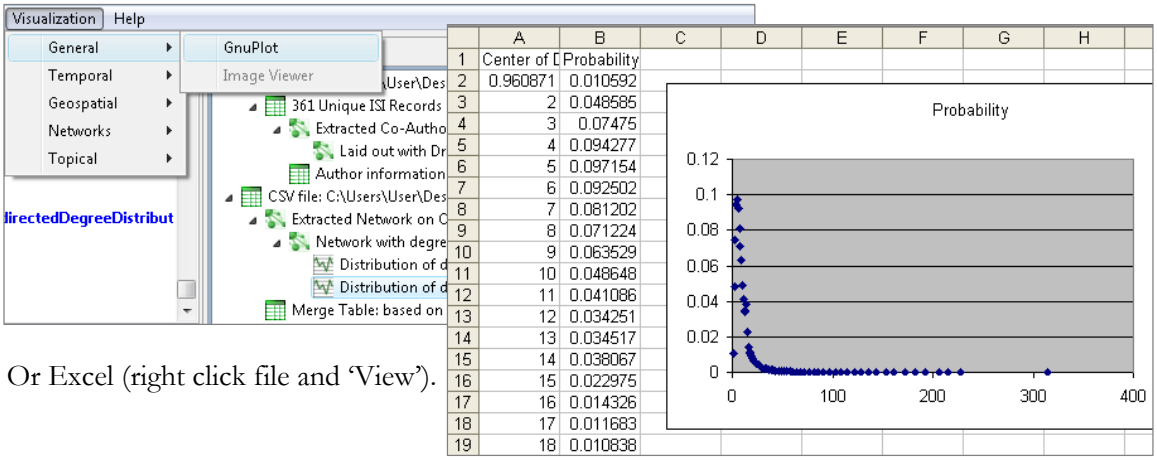


DrL Output – Plot Node Degree Distribution

Calculate degree distribution using and plot using 'Visualization > General > Gnuplot'







	A	B	C	D	E	F	G	H
1	Center of D	Probability						
2	0.960871	0.010592						
3	2	0.048585						
4	3	0.07475						
5	4	0.094277						
6	5	0.097154						
7	6	0.092502						
8	7	0.081202						
9	8	0.071224						
10	9	0.063529						
11	10	0.048648						
12	11	0.041086						
13	12	0.034251						
14	13	0.034517						
15	14	0.038067						
16	15	0.022975						
17	16	0.014326						
18	17	0.011683						
19	18	0.010838						

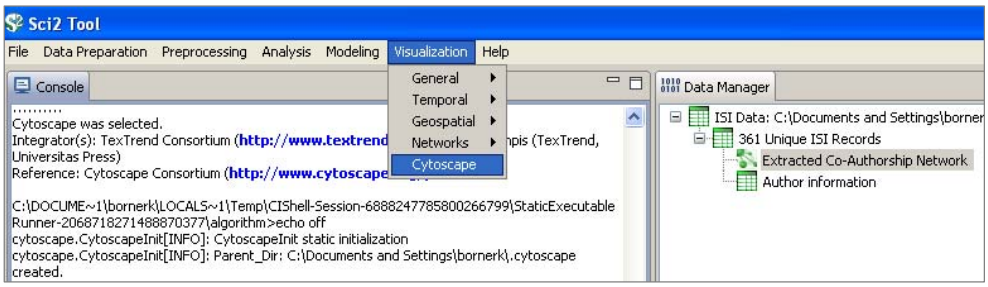
Or Excel (right click file and 'View').

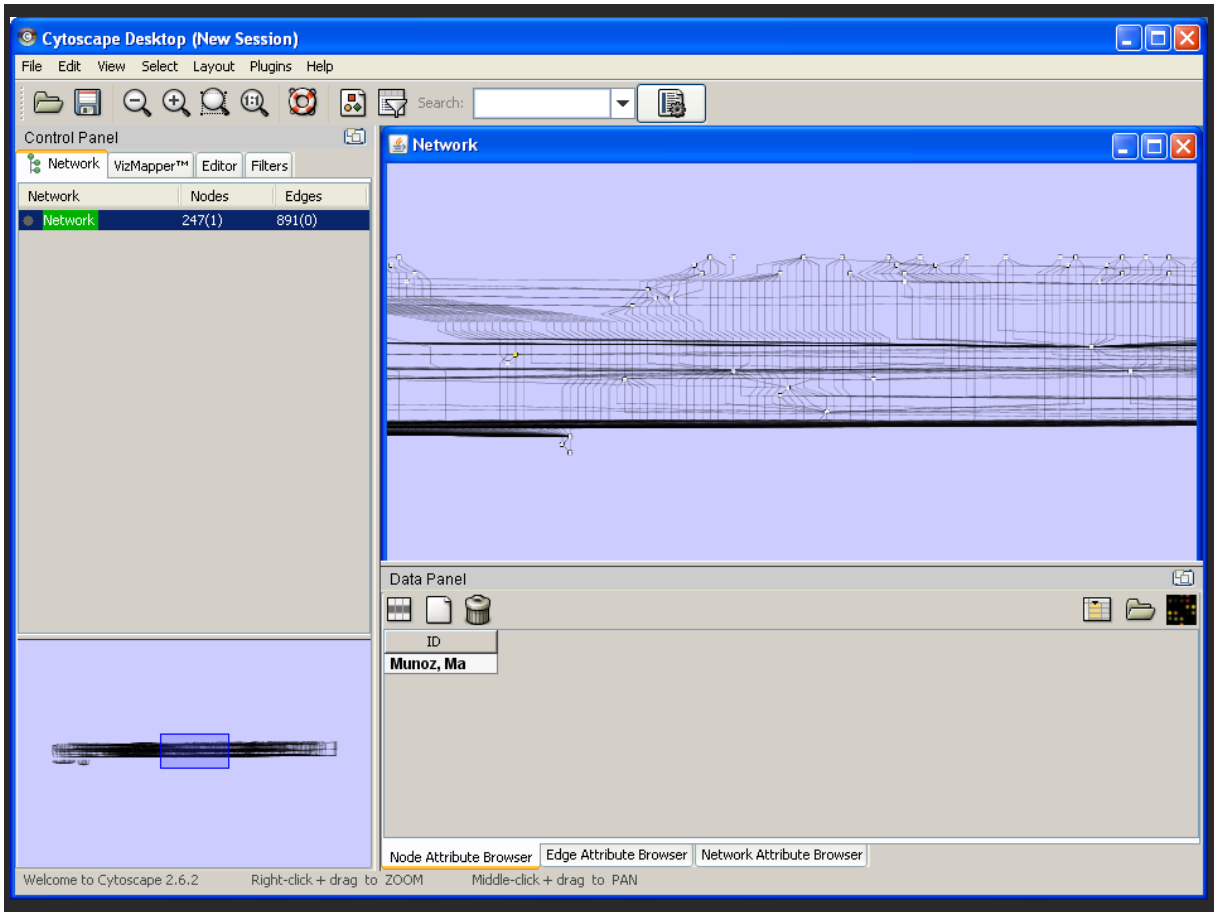
[#09] Large Network Analysis and Visualization

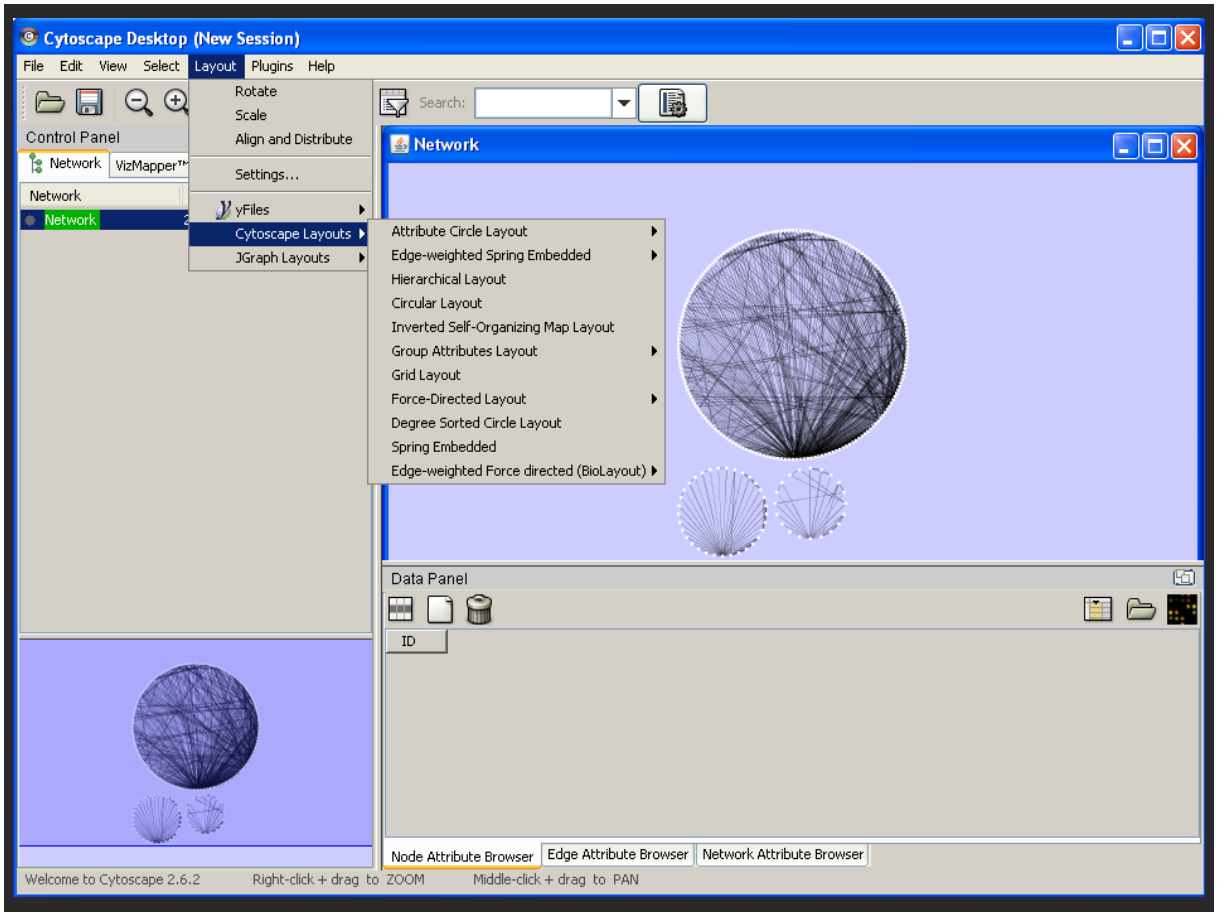
- General Overview
- Designing Effective Network Visualizations
- Sci2-Reading and Modeling Networks
- Sci2-Analysing Large Networks
- Sci2-Visualizing Large Networks and Distributions
- Outlook
- Exercise: Identify Promising Large Network Analyses of NIH Data

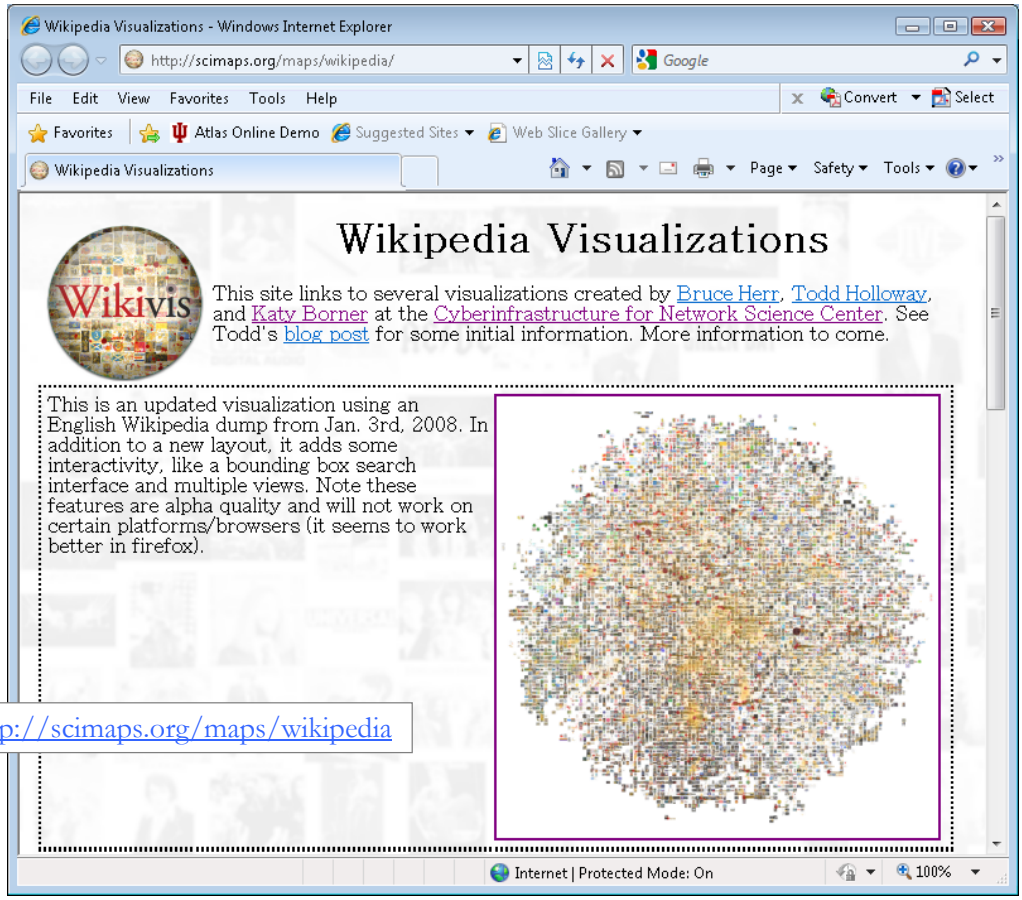
Planned Work

- Add (scalable) clustering algorithms to Sci2 Tool.
- Advanced network reduction algorithms.
- Visual language that helps communicate patterns, trends, activity bursts, etc.
- More interactivity, e.g., by opening networks in Cytoscape http://www.cytoscape.org.









http://scimaps.org/maps/wikipedia

[#09] Large Network Analysis and Visualization

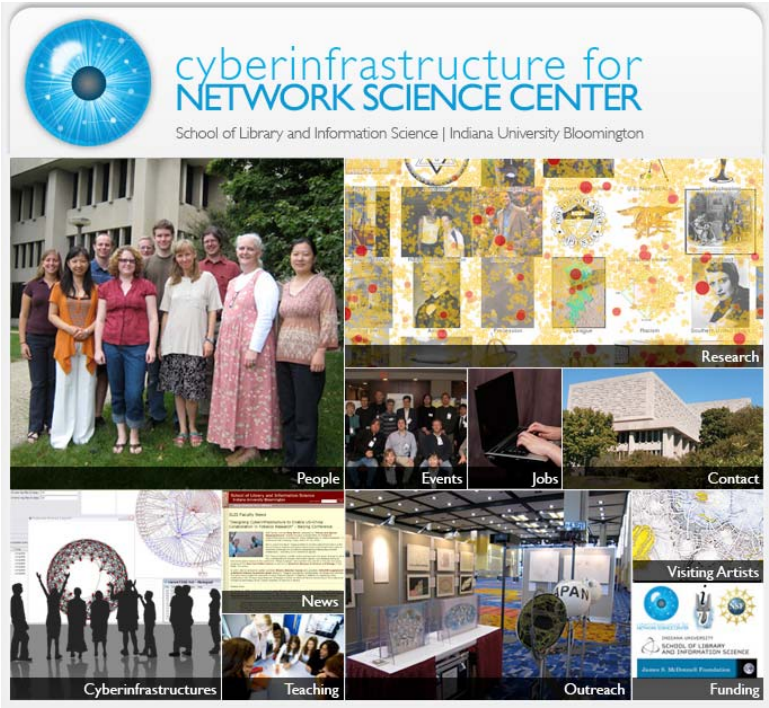
- General Overview
- Designing Effective Network Visualizations
- Sci2-Reading and Modeling Networks
- Sci2-Analysing Large Networks
- Sci2-Visualizing Large Networks and Distributions
- Outlook
- Exercise: Identify Promising Large Network Analyses of NIH Data

Exercise

Please identify a promising large network analysis of NIH data.

Document it by listing

- Project title
- User, i.e., who would be most interested in the result?
- Insight need addressed, i.e., what would you/user like to understand?
- Data used, be as specific as possible.
- Analysis algorithms used.
- Visualization generated. Please make a sketch with legend.



All papers, maps, cyberinfrastructures, talks, press are linked from http://cns.slis.indiana.edu