Rete-Netzwerk-Red: Analyzing and Visualizing Scholarly Networks Using the Network Workbench Tool

Katy Börner[1], Bonnie (Weixia) Huang, Micah Linnemeier[2], Russell J. Duhon, Patrick Phillips, Ninali Ma[3], Angela Zoss, Hanning Guo, Mark A. Price

[1] katy@indiana.edu / [2] mwlinnem@indiana.edu (NWB) / [3] nianma@indiana.edu (SDB)
Cyberinfrastructure for Network Science Center, School of Library and Information Science
Indiana University, Bloomington, IN (USA)

Abstract

The enormous increase in digital scholarly data and computing power combined with recent advances in text mining, linguistics, network science, and scientometrics make it possible to scientifically study the structure and evolution of science on a large scale. This paper discusses the challenges of this 'BIG science of science' – also called 'computational scientometrics' research – in terms of data access, algorithm scalability, repeatability, as well as result communication and interpretation. It then introduces two infrastructures: (1) the Scholarly Database (SDB) (http://sdb.slis.indiana.edu), which provides free online access to 20 million scholarly records – papers, patents, and funding awards which can be cross-searched and downloaded as dumps, and (2) Scientometrics-relevant plug-ins of the open-source Network Workbench (NWB) Tool (http://nwb.slis.indiana.edu). The utility of these infrastructures is then exemplarily demonstrated in three studies: a comparison of the funding portfolios and co-investigator networks of different universities, an examination of paper-citation and co-author networks of major network science researchers, and an analysis of topic bursts in streams of text. The paper concludes with a discussion of related work that aims to provide practically useful and theoretically grounded cyberinfrastructure in support of computational scientometrics research, practice, and education.

Introduction

About 45 years ago, de Solla Price suggested studying science using the scientific methods of science (de Solla Price, 1963). Today, science of science studies draw from diverse fields such as scientometrics, informetrics, webometrics, history of science, sociology of science, psychology of the scientist, operational research on science, the economics of science, the analysis of the flow of scientific information, as well as the planning of science. They gather, handle, interpret, and predict a variety of features of the science and technology enterprise, such as scholarly communication, performance, development, and dynamics that are interesting for science (policy) decisions in academia, government, and industry.

Most studies use either Thomson Scientific's databases or Scopus, as they each constitute a multi-disciplinary, objective, internally consistent publication database. A number of recent studies have examined and compared the coverage of Thomson Scientific's Web of Science (WoS), Scopus, Ulrich's Directory, and Google Scholar (GS) (Meho & Yang, 2007; Pauly & Stergiou, 2005). It has been shown that the databases have a rather small overlap in records. In one study, the overlap between WoS and Scopus was only 58.2% while the overlap between GS and the union of WoS and Scopus was a mere 30.8%. While Scopus covers almost twice as many journals and conferences as WoS, it covers fewer journals in the arts and humanities. A comprehensive analysis requires access to more than one database, and more and more studies also correlate publication output with patent production, funding input, and other datasets.

While diverse tools exist to crawl, pre-process, analyze, or visualize scholarly data, most of the tools used in science of science studies today are proprietary or 'closed source', making it hard to impossible to replicate results, to compare new and old approaches, or to agree on standards.

Cyberinfrastructures, i.e., the programs, algorithms and computational resources required to support advanced data acquisition, storage, management, integration, visualization and analysis (Atkins et al., 2003), address the ever growing need to connect researchers and practitioners to the data, algorithms, and massive disk space and computing power that many computational sciences require (Emmott et al., 2006). At a time when efficient science and technology management is urgently needed; researchers need to make sense of massive amounts of data, knowledge, and expertise; and industry tries to overcome a major recession, access to an effective science of science cyberinfrastructure becomes highly desirable. The envisioned infrastructure would provide easy access to terabytes of scholarly data as well as advanced algorithms and tools running on powerful computers in support of 'BIG science of science' research – also called 'computational scientometrics', a term coined by C. Lee Giles (Giles, 2006). Ideally, the infrastructure would be free, i.e., available to anyone, and open source, i.e., anybody could see, improve, and add to the software code. It should have means to record analysis workflows so that others can rerun analyses and replicate results. It should support the effective communication and discussion of results.

This paper introduces the beginnings of such a science of science cyberinfrastructure: the Scholarly Database (SDB) and the Network Workbench (NWB) Tool. It starts with a general introduction of the system architecture, functionality, and user interface of SDB and NWB. It then demonstrates their utility in three original research studies that exemplify common workflows for the acquisition and preparation of bibliographic data, temporal data analysis, network analysis and visualization. The paper concludes with a discussion of related work.

Scholarly Database

The Scholarly Database (SDB) at Indiana University aims to serve researchers and practitioners interested in the analysis, modelling, and visualization of large-scale scholarly datasets. The motivation for this database and its previous implementation were presented in (LaRowe et al., 2007). The online interface at http://sdb.slis.indiana.edu provides access to four datasets: Medline papers, U.S. Patent and Trademark Office patents (USPTO), National Science Foundation (NSF) funding, and National Institutes of Health (NIH) funding – over 20 million records in total, see Table 1. Users can register for free to cross-search these databases and to download result sets as dumps for scientometrics research and science policy practice.

Table 1. Names, number of records, years covered, and update information for datasets currently available via the Scholarly Database.

Dataset Name	# Records	Years Covered	Regular Update
Medline Publications	14,443,225	1898-2008	Yes
U.S. Patent and Trademark Office Patents	3,710,952	1976-2007	Yes
National Institutes of Health (NIH) Awards	1,043,804	1961-2002	No
National Science Foundation (NSF) Awards	174,835	1985-2004	No

SDB supports search across paper, patent, and funding databases. To search, select a year range and database(s) and enter search term(s) in creators (author/awardee/inventor), title, abstract, and full text (keywords and other text) fields; see Figure 1 (left).

The importance of a particular term in a query can be increased by putting a ^ and a number after the term. For instance, 'breast cancer^10' would increase the importance of matching the term 'cancer' by ten compared to matching the term 'breast'. Custom database queries can be run by contacting the SDB team lead and author of this paper, Nianli Ma. Search results

retrieved from different databases can be downloaded as data dumps in csv file format; see Figure 1 (right).

SDB stores all data in a PostgreSQL database (PostgreSQL Global Development Group, 2009). Full-text search is supported using Solr (The Apache Software Foundation, 2007) to index the contents of the database. Solr is an industry-standard, open source search server that can scale to very large amounts of data using replication and sharding. The online interface was developed in Django (Django Software Foundation, 2009). Django is a web framework written in the Python (Python Software Foundation, 2008) programming language with particularly good support for content-oriented web applications.
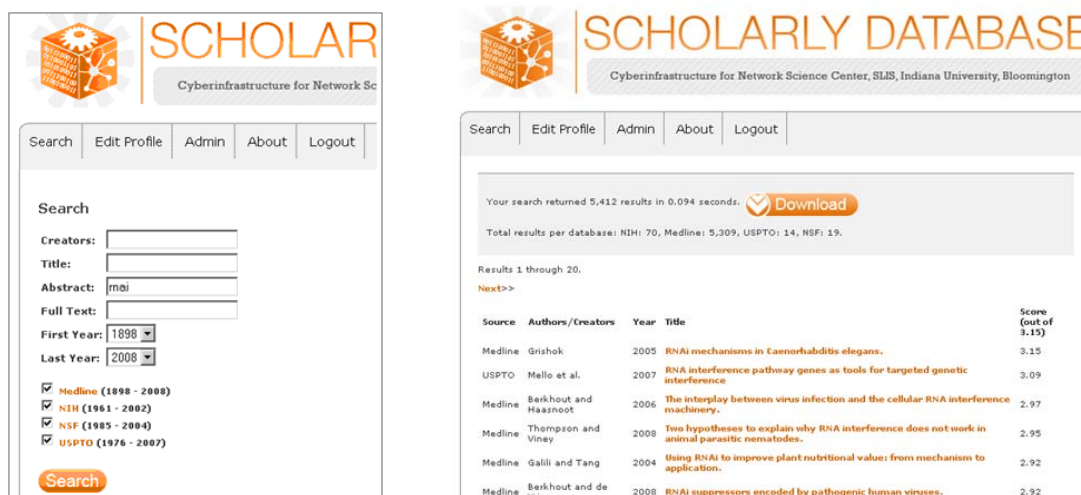

Figure 1. Partial search interface (left) and download interface (right) for the Scholarly Database. Note that the highest scoring record was retrieved from Medline while the second highest record comes from USPTO.

Network Workbench Tool

The Network Workbench (NWB) Tool (http://nwb.slis.indiana.edu) is a network analysis, modelling, and visualization toolkit for physics, biomedical, and social science research (Herr II et al., 2007). The basic interface comprises a 'Console', 'Data Manager', and 'Scheduler' Window as shown in Figure 2. The top menu provides easy access to relevant 'Preprocessing', 'Modeling', 'Analysis', 'Visualization', and 'Scientometrics' algorithms. Information on how to download, install, and run the NWB Tool can be found in the Network Workbench Tool User Manual (Cyberinfrastructure For Network Science Center, 2009).

NWB is built on Cyberinfrastructure Shell (CIShell) (Cyberinfrastructure for Network Science Center, 2008), an open source software framework for the easy integration and utilization of datasets, algorithms, tools, and computing resources. CIShell is based on the OSGi R4 Specification and Equinox implementation (OSGi-Alliance, 2008).

The Network Workbench Community Wiki (https://nwb.slis.indiana.edu/community) provides a one-stop online portal for researchers, educators, and practitioners interested in the study of networks. It is a place for users of the NWB Tool, CIShell, or any other CIShell-based program to get, upload, and request algorithms and datasets to be used in their tool so that it truly meet their needs and the needs of the scientific community at large.

Users of the NWB Tool can

- Access major network datasets online or load their own networks.
- Perform network analysis with the most effective algorithms available.
- Generate, run, and validate network models.

- Use different visualizations to interactively explore and understand specific networks.
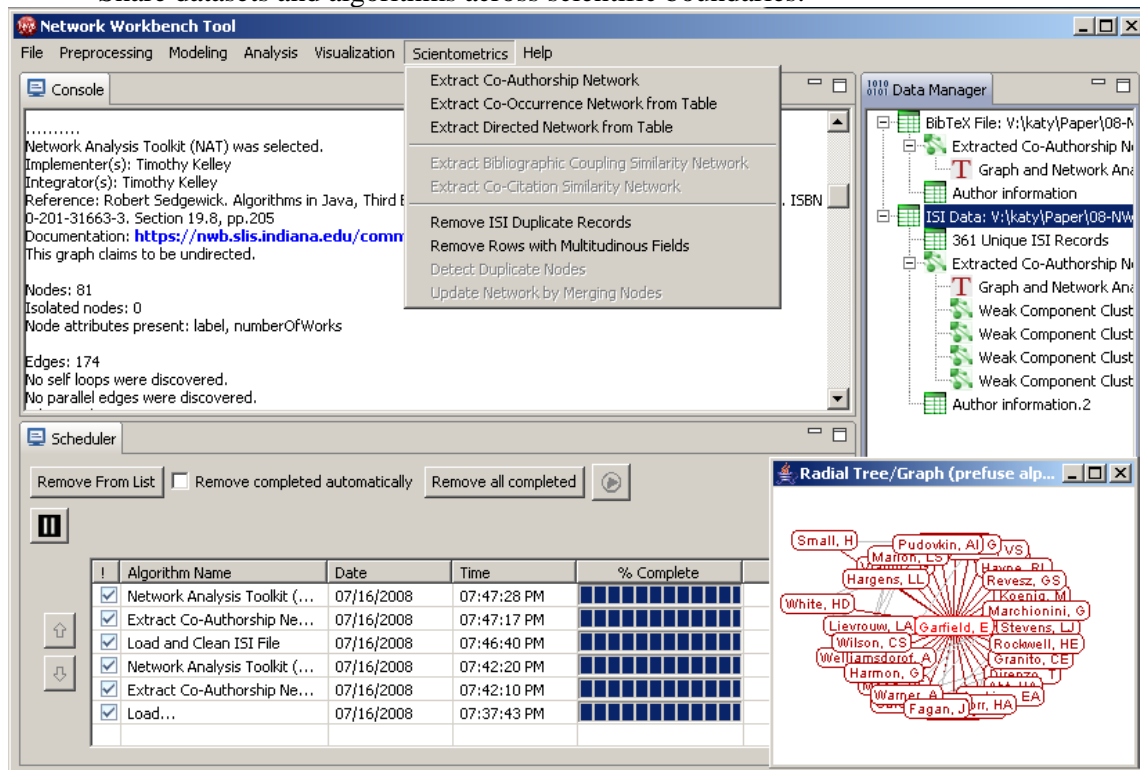- Share datasets and algorithms across scientific boundaries.

Figure 2. NWB Tool interface with Console, Data Manager, and Scheduler Windows, the Scientometrics menu, and a Radial Graph visualization of Garfield's co-author network.

In December 2008, the NWB Tool provides access to over 80 algorithms and 30 sample datasets for the study of networks. The loading, processing, and saving of seven file formats (NWB, GraphML, Pajek .net, Pajek .matrix, XGMML, TreeML, CSV) and an automatic conversion service among those formats are supported. Relevant for science of science studies, the NWB Tool can read data downloaded from SDB, Google Scholar, ISI Thomson Scientific Reuters, Scopus, and the NSF award database as well as EndNote and BibTeX formatted data.

Additional algorithms and data formats can be easily integrated into the NWB Tool using wizard driven templates. Although the CIShell and the NWB Tool are developed in JAVA, algorithms developed in other programming languages such as FORTRAN, C, and C++ can be integrated. Among others, JUNG (O'Madadhain et al., 2008) and Prefuse libraries (Heer et al., 2005) have been integrated into the NWB as plug-ins. NWB also supplies a plug-in that invokes the GnuPlot application (Williams & Kelley, 2008) for plotting data analysis results and the GUESS tool (Adar, 2007) for rendering network layouts. Support and advice in algorithm integration and custom tool development is provided by the NWB team lead and author of this paper, Micah Linnemeier.

Exemplary Workflows

This section aims to demonstrate the utility of the SDB and NWB Tool to answer specific research questions in an efficient and repeatable fashion. Detailed, step-by-step instructions on how to run these and many other analyses can be found in the Network Workbench Tool User Manual (NWB Team, 2008) and NWB Tutorial Slides (Börner, 2008).

NSF Funding Portfolios and Co-Investigator Networks of U.S. Universities

The first study aims to answer: What funding portfolios do major U.S. universities have, what scholarly co-investigator networks does funding inspire/support, and what roles do investigators play, e.g., gatekeeper, maximum degree, maximum funding amount?

Funding data was downloaded from the *Award Search* site provided by the National Science Foundation (NSF) (http://www.nsf.gov/awardsearch). The site supports search by PI name, institution, and many other fields. Exemplarily, active NSF awards data from Indiana University, Cornell University, and University of Michigan Ann Arbor were downloaded on November 07, 2008. The files were loaded into the NWB Tool and co-investigator networks were extracted and visualized in GUESS. In these networks, nodes represent investigators and edges denote their co-occurrence, i.e., co-authorship. The co-investigator network of Cornell University without isolate nodes is shown in Figure 3, left. The largest connected component of this network is given in Figure 3, right. In both networks, the node size and color corresponds to the total award amount with smaller, darker nodes representing less money and larger, light green nodes denoting more funding and the top-50 nodes with the highest funding amounts are labeled.
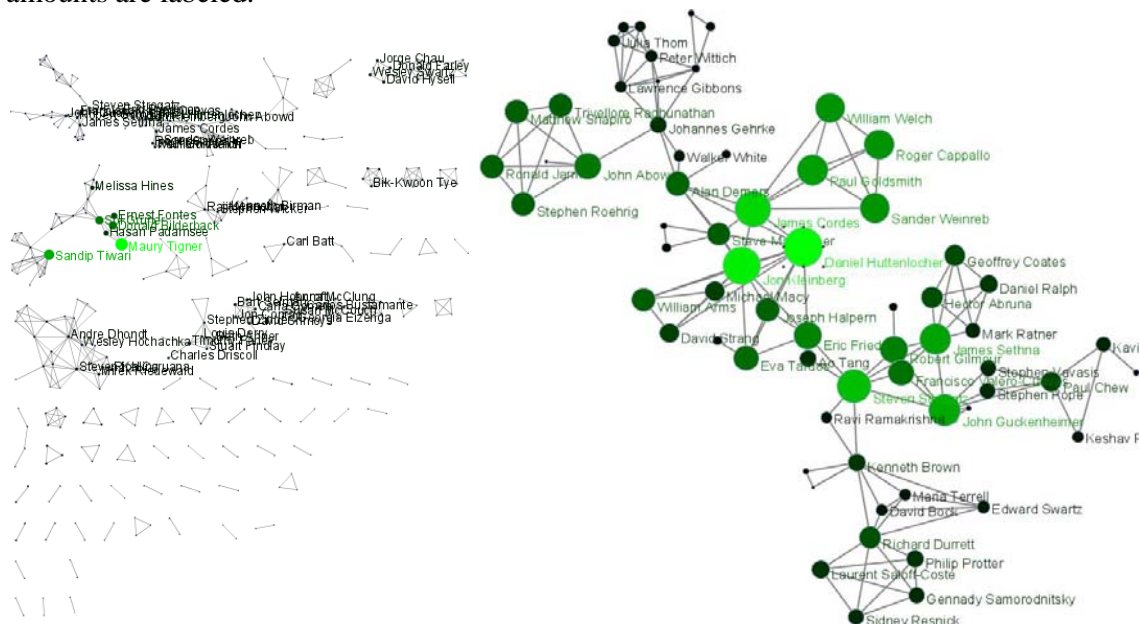

Figure 3 Complete network (left) and largest component (right) of Cornell University's co-investigator network (67 nodes).

The general characteristics of all three networks are given in Table 2. Note that the total award amount is attributed to the main investigator and his/her institution exclusively.

Table 2. Award properties and co-investigator network features for the three universities.

University	#Records / total award amount	Co-investigator network: #nodes / #edges / # components	Largest component: #nodes / #edges	Name, department, and $ amount of top investigator
Indiana University	257 / $100 million	223 / 312 / 52	19 / 37	Curtis Lively, Biology $7,436,828
Cornell University	501 / $546 million	375 / 573 / 78	67 / 143	Maury Tigner, Physics $107,216,976
Michigan University	619 / $305 million	497 / 672 / 117	55 / 105	Khalil Najafi, EECS $32,541,158

There are interesting differences in the funding portfolios of these universities. Michigan has clearly the largest number of currently active NSF awards totalling 497. With $546 million, Cornell has the highest total award amount. Cornell also has the largest component with 67 investigator nodes and 143 collaboration links, indicating much cross-fertilization across different disciplines. Cornell also happens to employ the investigator who currently has the highest total award amount: Maury Tigner. Note that being the main investigator on one major center grant and several campus equipment grants can easily result in multi-millions to spend over many years. Note also that the funding portfolios, networks, and top-investigators from other agencies, e.g., NIH, might look very different.

A closer examination of the largest component of the Cornell co-investigator network shown in Figure 3, right reveals that Steven Strogatz has the highest betweenness centrality (BC), effectively bridging between several disciplines, and Daniel Huttenlocher has the highest degree, i.e., the most collaborations with others in this network.

Future work should consider different means to associate award amounts to investigators and institutions. An analysis of the distribution of funding over scientific disciplines and departments is desirable. Co-investigator linkages among institutions deserve further attention.

Paper-Citation and Co-Author Networks of Major Network Science Researchers

The second study addresses the questions: Do researchers which come from different domains of science but contribute major works to one and the same domain, e.g., network science, grow different collaboration networks? How much do their publication, citation, and h-index dynamics differ?

Exemplarily, four major network science researchers were selected: Eugene Garfield and three principal investigators of the Network Workbench project. Data for all four male researchers was downloaded from Reuters/Thomson Scientific in December 2007. Their names, ages (retrieved from Library of Congress), number of citations for highest cited paper, h-index (Bornmann, 2006), and number of papers and citations over time as calculated by the Web of Science by Thomson Scientific (Thomson-Reuters, 2008) are given in Table 3. Note that this dataset does not capture books or Conference proceedings by these authors.

Table 3. Names, age, number of papers, number of citations for highest cited paper, and h-index for four major network science researchers.

Author Name	Department	Age in '07	# Papers	# Cites	h-Index
Eugene Garfield	IS, Scientometrics	82	672	672	31
Stanley Wasserman	Sociology, Psychology, Statistics	56	35	122	17
Alessandro Vespignani	Physics	42	101	451	33
Albert-László Barabási	Physics, Biology	40	126	2,218	47 (Dec 07)
		41	*159*	*3,488*	*52 (Dec 08)*

The older an author the more papers, citations, and the higher an h-index are expected. Yet, Vespignani and Barabási publishing in physics and biology manage to attract citation counts and have h-indexes that are impossible to achieve in social science domains in such a short time frame. To give a concrete example, in Dec. 2007, Garfield's highest cited paper on "Citation Analysis as a Tool in Journal Evaluation" published in 1972 had 672 counts. Barabási's highest cited paper published in 1999 has 2,218 counts; in December 2008 the same paper has 3,488 citation counts. Within one single year, Barabási's h-index increased by 5 to 52.

Similarly, there are major differences in the structure of the collaboration networks in which these four authors are embedded in. Figure 4 shows the joint co-author network of all records retrieved for the four authors as rendered in GUESS. Each node represents an author and is color and size coded by the number of papers per author. Edges represent co-author relationships and are color and thickness coded by the number of times two authors wrote a paper together. The top-50 authors with the most papers are labelled.

While Barabási's and Vespignani's co-author network are strongly interlinked with Stanley and Vazquez being major connectors with high BC values, Garfield's and Wasserman's networks are unconnected to the networks of any of the three other researchers. The size and density of the networks differs considerably. When extracting the co-author network from all of Barabási's papers, a co-author network with 128 nodes and density of .07517 results; for Vespignani the network has 72 nodes and .08646 density, Wasserman has 18 and .20261, and Garfield has 33 and .11932. The top-3 strongest co-author linkages for Barabási are Vicsek, Jeong, Albert; for Vespignani are Zapperi, Pastor-Satorras,Pietronero, for Wasserman are Galaskiewicz, Iacobucci, Anderson; and for Garfield are Pudovkin, Welljamsdorof, and Sher is tied for third place with four other authors. The paper with the most authors is entitled "Experimental determination and system level analysis of essential genes in Escherichia coli MG1655" (2003) and has 21 authors.
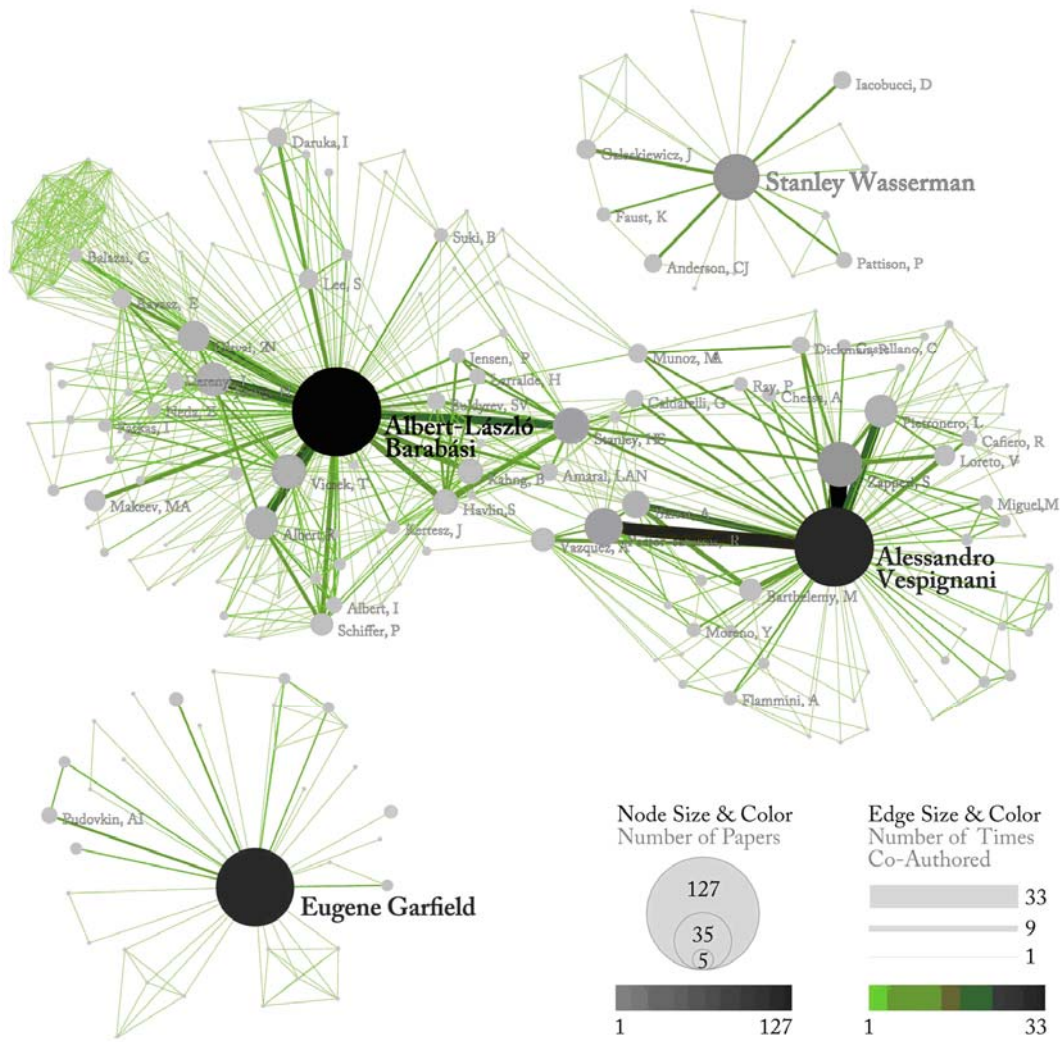

Figure 4. Joint co-author network of all four network science experts.

Bursts of Activity in Streams of Text

The third study aims to answer: What topic burst exist in an emerging area of research, e.g., RNA interference (RNAi) research? Exactly what topics are active and when?

Using the Scholarly Database, 5,309 Medline papers with 'rnai' in the abstract field were retrieved and downloaded; see Figure 1 for interface snapshots. The first paper was published in 1978 and the number of papers on this topic increases considerably in recent years. The sequence of papers published over time can be seen as a discrete time series of words. Kleinberg's burst detection algorithm (Kleinberg, 2002) identifies sudden increases in the usage frequency of words over time. Rather than using plain frequencies of the occurrences of words, the algorithm employs a probabilistic automaton whose states correspond to the frequencies of individual words. State transitions correspond to points in time around which the frequency of a word changes significantly. The algorithm generates a ranked list of the word bursts in the document stream, together with the intervals of time in which they occurred. Using the burst detection algorithm available via the NWB Tool, the abstracts of the 5,309 Medline papers were analyzed. First, all words occurring in the abstract were normalized: they were stemmed, i.e., words such as 'scientific', 'science' were reduced to 'scien' and stop words such as 'a' or 'the' were removed. The result is a 1,224 row table with all bursting words, their burst length, weight, and the start and end years of the bursts. The table was sorted by burst weight and four words: 'protein', 'result', 'use', and 'function' top the list with an infinite burst weight. The subsequent top-7 most highly bursting words are given in Table 4. Since the burst detection algorithm was run with 'bursting state = 1', i.e., only one burst per word, the burst weight is identical to the burst strength in this output and only the burst weight is shown in Table 4. Interestingly, all these words burst rather early in 1998 or 1999. Many of them have a rather long burst duration with 'embryo' bursting over 6 years.

Table 4. Words and their burst length and weight as well as start and end years.

Word	Length	Weight	Start	End
elegan	5	105.4307	1998	2002
strand	4	70.72182	1999	2002
doubl	5	62.88262	1999	2003
embryo	6	42.70386	1998	2003
caenorhabd	5	36.59776	1998	2002
drosophila	5	33.98245	1999	2003
phenotyp	3	31.08153	1999	2001

The result shows the words and concepts important to the events being studied that increased in usage, were more active for a period of time, and then faded away.

Related Work and Discussion

A discussion of the unique features of the Scholarly Database and it relation to similar efforts was provided in (LaRowe et al., 2007). Here features and related work of the NWB Tool are discussed. Table 5 provides an overview of existing tools used in scientometrics research; see also (Börner et al., 2007; Fekete & Börner, 2004). The tools are sorted by the date of their creation. Domain refers to the field in which they were originally developed, such as social science (SocSci), scientometrics (Scientom), biology (Bio), geography (Geo), and computer science (CS).

Table 5. Network analysis and visualization tools commonly used in scientometrics research.

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)
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)

GUESS	2007	Networks	A tool for visual graph exploration that integrates a scripting environment.	Yes	All Major	(Adar, 2007)
Publish or Perish	2007	Scientom	Harvests and analyzes data from Google Scholar, focusing on measures of research impact.	No	Windows, Linux	(Harzing, 2008)

Many of these tools are specialized and very capable. For instance, BibExcel and Publish or Perish are great tools for bibliometric data acquisition and analysis. HistCite and CiteSpace each support very specific insight needs – from studying the history of science to the identification of scientific research frontiers. The S&T Dynamics Toolbox provides many algorithms commonly used in scientometrics research and it provides bridges to more general tools. Pajek and UCINET are very versatile, powerful network analysis tools that are widely used in social network analysis. Cytoscape is excellent for working with biological data and visualizing networks.

The NWB Tool has fewer analysis algorithms than Pajek and UCINET and less flexible visualizations than Cytoscape. However, it is open source, highly flexible, and scalable to very large networks. Plus, it makes it much easier for researchers and algorithm authors to integrate new and existing algorithms and tools that take in diverse data formats. This is made possible by the OSGi component architecture and CIShell algorithm architecture built on top of OSGi.

The Cytoscape team recently decided to adopt an architecture based on OSGi. This will make it possible for Cytoscape to use many of the NWB analysis and modelling algorithms while the NWB Tool can benefit from Cytoscape's visualization capabilities. Other software development teams are exploring an adoption of OSGi. Ultimately, a true marketplace-like cyberinfrastructure might result that makes it easy to share and use datasets, algorithms, and tools across scientific boundaries. Ideally, this marketplace is free for anybody to use and contribute to, enabling harnessing the power of millions of minds for studies in biology, physics, social science, and many other disciplines but also for the study of science itself.

Acknowledgments

References

Adar, Eytan. (2007). *Guess: The Graph Exploration System.* http://graphexploration.cond.org/ (accessed on 4/22/08).

AT&T-Research-Group. (2008). *Graphviz-Graph Visualizaiton Software*. http://www.graphviz.org/Credits.php (accessed on 7/17/08).

Atkins, Daniel E., Kelvin K. Drogemeier, Stuart I. Feldman, Hector Garcia-Molina, Michael L. Klein, David G. Messerschmitt, Paul Messian, Jeremiah P. Ostriker, Margaret H. Wright. (2003). Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure. Arlington, VA: National Science Foundation.

Auber, David. (2003). Tulip: A Huge Graph Visualisation Framework. In Petra Mutzel & Michael Jünger (Eds.), *Graph Drawing Softwares, Mathematics and Visualization* (pp. 105-126). Berlin: Springer-Verlag.

Batagelj, Vladimir, Andrej Mrvar. (1998). Pajek-Program for Large Network Analysis. *Connections, 21*(2), 47-57

Borgatti, S.P., M. G. Everett, L.C. Freeman. (2002). *Ucinet for Windows: Software for Social Network Analysis*. http://www.analytictech.com/ucinet/ucinet_5_description.htm (accessed on 7/15/08).

Börner, Katy. (2008).Network Workbench Tool: For Large Scale Network Analysis, Modeling, and Visualization. Presented at I-School, University of Michigan, http://ivl.slis.indiana.edu/km/pres/2008-borner-nwb-ws.pdf (accessed on 1/13/2009).

Börner, Katy, Bruce W. Herr II, Jean-Daniel Fekete. (2007, July 3, 2007). *IV07 Software Infrastructures Workshop.* IV07 Zurich, Switzerland.

Bornmann, Lutz. (2006). *H Index: A New Measure to quantify the Research Output of Individual Scientists*. http://www.forschungsinfo.de/iq/agora/H_Index/h_index.asp (accessed on 7/17/08).

Brandes, Ulrik, Dorothea Wagner. (2008). *Analysis and Visualization of Social Networks*. http://visone.info/ (accessed on 7/15/08).

Chen, Chaomei. (2006). CiteSpace II: Detecting and Visualizing Emerging Trends and Transient patterns in Scientific Literature. *JASIST, 54*(5), 359-377.

Csárdi, Gábor, Tamás Nepusz. (2006). *The igraph software package for complex network research*. http://necsi.org/events/iccs6/papers/c1602a3c126ba822d0bc4293371c.pdf (accessed on 7/17/08).

Cyberinfrastructure for Network Science Center. (2008). Cyberinfrastructure Shell. http://cishell.org/ (accessed on 7/17/08).

Cyberinfrastructure For Network Science Center. (2009). *Network Workbench Tool: User Manual, 1.0.0 beta*. http://nwb.slis.indiana.edu/Docs/NWB-manual-1.0.0beta.pdf (accessed on 04/13/2009).

Cytoscape-Consortium. (2008). *Cytoscape*. http://www.cytoscape.org/index.php (accessed on 7/15/08).

de Solla Price, Derek J. (1963). *Little Science, Big Science.* Unpublished Manuscript.

Django Software Foundation. (2009). Django: The Web Framework for Perfectionists with Deadlines. http://www.djangoproject.com/contact/foundation/ (accessed on 1/13/2009).

Emmott, Stephen, Stuart Rison, Serge Abiteboul, Christopher Bishop, José Blakeley, René Brun, Søren Brunak, Peter Buneman, Luca Cardelli. (2006). *Towards 2020 Science*. The Microsoft Research Group and the 2020 Science Group. http://research.microsoft.com/en-us/um/cambridge/projects/towards2020science/downloads/T2020S_ReportA4.pdf (accessed on 1/13/2009).

Fekete, Jean-Daniel, Katy Börner. (2004). *Workshop on Information Visualization Software Infrastructures* Austin, Texas.

Garfield, Eugene. (2008). HistCite: Bibliometric Analysis and Visualization Software (Version 8.5.26). Bala Cynwyd, PA: HistCite Software LLC. http://www.histcite.com/ (accessed on 7/15/08).

Giles, C. Lee. (2006). The Future of CiteSeeer: CiteSeer[x], *Lecture Notes in Computer Science* (Vol. 4213). Berlin/Heidelberg: Springer.

Harzing, Anne-Wil. (2008). *Publish or Perish: A citation analysis software program*. http://www.harzing.com/resources.htm (accessed on 4/22/08).

Heer, Jeffrey, Stuart K. Card, James A. Landay. (2005). *Prefuse: A toolkit for interactive information visualization.* Conference on Human Factors in Computing Systems Portland, OR: New York: ACM Press, pp. 421-430.

Herr II, Bruce W., Weixia (Bonnie) Huang, Shashikant Penumarthy, Katy Börner. (2007). Designing Highly Flexible and Usable Cyberinfrastructures for Convergence. In William S. Bainbridge & Mihail C. Roco (Eds.), *Progress in Convergence: Technologies for Human Wellbeing* (Vol. 1093, pp. 161-179). Boston, MA: Annals of the New York Academy of Sciences.

Huang, Weixia (Bonnie), Bruce Herr, Russell Duhon, Katy Börner. (2007). Network Workbench--Using Service-Oriented Architecture and Component-Based Development to Build a Tool for Network Scientists. Presented at International Workshop and Conference on Network Science, Queens, NY.

Ihaka, Ross, Robert Gentleman. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics, 5*(3), 299-314. http://www.amstat.org/publications/jcgs/ (accessed on 7/17/08).

Kleinberg, J. M. (2002). *Bursty and Hierarchical Structure in Streams.* 8th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining: ACM Press, pp. 91-101.

Krebs, Valdis. (2008). *Orgnet.com: Software for Social Network Analysis and Organizational Network Analysis*. http://www.orgnet.com/inflow3.html (accessed

LaRowe, Gavin, Sumeet Ambre, John Burgoon, Weimao Ke, Katy Börner. (2007). The Scholarly Database and Its Ultility for Scientometrics Research. Presented at Proceedings of the 11th International Conference on Scientometrics and Informetrics, Madrid, Spain, June 25-27. 457-462.

Leydesdorff, Loet. (2008). *Software and Data of Loet Leydesdorff*. http://users.fmg.uva.nl/lleydesdorff/software.htm (accessed on 7/15/2008).

Meho, Lokman I., Kiduk Yang. (2007). Impact of Data Sources on Citation Counts and Rankings of LIS Faculty: Web of Science Versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology, 58*(13), 2105-2125. http://dlist.sir.arizona.edu/1733/01/meho-yang-03.pdf (accessed on 04/01/2009).

NWB Team. (2008). Network Workbench--A Workbench for Network Scientists: Documentation. http://nwb.slis.indiana.edu/doc.html (accessed on 1/13/2009).

O'Madadhain, Joshua, Danyel Fisher, Tom Nelson. (2008). Jung: Java Universal Network/Graph Framework: University of California, Irvine. http://jung.sourceforge.net/ (accessed on 04/01/2009).

OSGi-Alliance. (2008). *OSGi Alliance*. http://www.osgi.org/Main/HomePage (accessed on 7/15/08).

Pauly, Daniel, Konstantinos I. Stergiou. (2005). Equivalence of Results from two Citation Analyses: Thomson ISI's Citation Indx and Google Scholar's Service. *Ethics in Science and Environmental Politics, 5*, 33-35.

Persson, Olle. (2008). Bibexcel. Umeå, Sweden: Umeå University. http://www.umu.se/inforsk/Bibexcel/ (accessed on 7/15/08).

PostgreSQL Global Development Group. (2009). PostgreSQL: The World's Most Advanced Open Source Database. http://www.postgresql.org/about/ (accessed on 1/13/2009).

Python Software Foundation. (2008). Python Programming Language--Official Website. http://www.python.org/ (accessed on 1/13/2009).

Siek, Jeremy, Lie-Quan Lee, Andrew Lumsdaine. (2002). *The Boost Graph Library: User Guide and Reference Manual.* New York: Addison-Wesley.

Takatsuka, M., M. Gahegan. (2002). GeoVISTA Studio: A Codeless Visual Programming Environment for Geoscientific Data Analysis and Visualization. *The Journal of Computers & Geosciences, 28*(10), 1131-1144

The Apache Software Foundation. (2007). Apache Solr. http://lucene.apache.org/solr/ (accessed on 1/13/2009).

Thomson-Reuters. (2008). *Web of Science*. http://scientific.thomsonreuters.com/products/wos/ (accessed on 7/17/08).

Williams, Thomas, Colin Kelley. (2008). *gnuplot homepage*. http://www.gnuplot.info/ (accessed on 7/17/08).