The Changing Scientific Landscape

**Star Scientist -> Research Teams:** In former times, science was driven by key scientists. Today, science is driven by effectively collaborating co-author teams often comprising expertise from multiple disciplines and several geospatial locations (Börner, Dall’Asta, Ke, & Vespignani, 2005; Shneiderman, 2008).

**Users -> Contributors:** Web 2.0 technologies empower anybody to contribute to Wikipedia and to exchange images and videos via Flickr and YouTube. WikiSpecies, WikiProfessionals, or WikiProteins combine wiki and semantic technology in support of real time community annotation of scientific datasets (Mons et al., 2008).

**Cross-disciplinary:** The best tools frequently borrow and synergistically combine methods and techniques from different disciplines of science and empower interdisciplinary and/or international teams of researchers, practitioners, or educators to fine-tune and interpret results collectively.

**One Specimen -> Data Streams:** Microscopes and telescopes were originally used to study one specimen at a time. Today, many researchers must make sense of massive streams of multiple types of data with different formats, dynamics, and origin.

**Static Instrument -> Evolving Cyberinfrastructure (CI):** The importance of hardware instruments that are rather static and expensive decreases relative to software infrastructures that are highly flexible and continuously evolving according to the needs of different sciences. Some of the most successful services and tools are decentralized increasing scalability and fault tolerance.

**Modularity:** The design of software modules with well defined functionality that can be flexibly combined helps reduce costs, makes it possible to have many contribute, and increases flexibility in tool development, augmentation, and customization.

**Standardization:** Adoption of standards speeds up development as existing code can be leveraged. It helps pool resources, supports interoperability, but also eases the migration from research code to production code and hence the transfer of research results into industry applications and products.

**Open data and open code:** Lets anybody check, improve, or repurpose code and eases the replication of scientific studies.
Desirable Features of Plug-and-Play Macrosopes

**Division of Labor:** Ideally, labor is divided in a way that the expertise and skills of computer scientists are utilized for the design of standardized, modular, easy to maintain and extend “core architecture”. Dataset and algorithm plugins, i.e., the “filling”, are initially provided by those that care and know most about the data and developed the algorithms: the domain experts.

**Ease of Use:** As most plugin contributions and usage will come from non-computer scientists it must be possible to contribute, share, and use new plugins without writing one line of code. Wizard-driven integration of new algorithms and data sets by domain experts, sharing via email or online sites, deploying plugins by adding them to the ‘plugin’ directory, and running them via a Menu driven user interfaces (as used in Word processing systems or Web browsers) seems to work well.

**Plugin Content and Interfaces:** Should a plugin represent one algorithm or an entire tool? What about data converters needed to make the output of one algorithm compatible with the input of the next? Should those be part of the algorithm plugin or should they be packaged separately?

**Supported (Central) Data Models:** Some tools use a central data model to which all algorithms conform, e.g., Cytoscape, see Related Work section. Other tools support many internal data models and provide an extensive set of data converters, e.g., Network Workbench, see below. The former often speeds up execution and visual rendering while the latter eases the integration of new algorithms. In addition, most tools support an extensive set of input and output formats.

**Core vs. Plugins:** As will be shown, the “core architecture” and the “plugin filling” can be implemented as sets of plugin bundles. Answers to questions such as: “Should the graphical user interface (GUI), interface menu, scheduler, or data manager be part of the core or its filling?” will depend on the type of tools and services to be delivered.

**Supported Platforms:** If the software is to be used via Web interfaces then Web services need to be implemented. If a majority of domain experts prefers a stand-alone tool running on a specific operating system then a different deployment is necessary.

---

**Project Details**

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

**Software Team:** Lead: Micah Linnemeier

Members: Patrick Phillips, Russell Duhon, Tim Kelley & Ann McCranie

Previous Developers: Weixia (Bonnie) Huang, Bruce Herr, Heng Zhang, Duygu Balcan, Bryan Hook, Ben Markines, Santo Fortunato, Felix Terkhorn, Ramya Sabbineni, Vivek S. Thakre & Cesar Hidalgo

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

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

**Duration:** Sept. 2005 - Aug. 2009

**Website:** [http://nwb.slis.indiana.edu](http://nwb.slis.indiana.edu)
NWB Advisory Board:

Jason Leigh (CI)  [http://www.evl.uic.edu/spiff/](http://www.evl.uic.edu/spiff/)
Neo Martinez (Biology)  [http://online.sfsu.edu/~webhead/](http://online.sfsu.edu/~webhead/)
Michael Macy, Cornell University (Sociology)  [http://www.soc.cornell.edu/faculty/macy.shtml](http://www.soc.cornell.edu/faculty/macy.shtml)
Ulrik Brandes (Graph Theory)  [http://www.inf.uni-konstanz.de/~brandes/](http://www.inf.uni-konstanz.de/~brandes/)
Mark Gerstein, Yale University (Bioinformatics)  [http://bioinfo.mbb.yale.edu/](http://bioinfo.mbb.yale.edu/)
Tom Snijders, University of Groningen  [http://stat.gamma.rug.nl/snijders/](http://stat.gamma.rug.nl/snijders/)
Noshir Contractor, Northwestern University  [http://www.spcomm.uiuc.edu/nosh/](http://www.spcomm.uiuc.edu/nosh/)

Resources

Publications
- [http://nwb.slis.indiana.edu/pub.html](http://nwb.slis.indiana.edu/pub.html)

Community Wiki, Tutorials, FAQ
- [https://nwb.slis.indiana.edu/community](https://nwb.slis.indiana.edu/community)
- [http://nwb.slis.indiana.edu/doc.html](http://nwb.slis.indiana.edu/doc.html)

Software
- [http://cishell.org](http://cishell.org)
- [http://nwb.slis.indiana.edu/download.html](http://nwb.slis.indiana.edu/download.html)

Developer Resources
- [http://cns-trac.slis.indiana.edu/trac/nwb](http://cns-trac.slis.indiana.edu/trac/nwb)
1. Exemplary Network Science Research by NWB PIs
   • Computational Proteomics
   • Computational Economics
   • Computational Social Science
   • Computational Scientometrics
   • Computational Epidemics

2. NWB Tool Challenges and Opportunities

3. NWB Tool Overview

4. NWB Tool for Scientometrics Research

5. Discussion of Future Work

Computational Proteomics

What relationships exist between protein targets of all drugs and all disease-gene products in the human protein–protein interaction network?

Computational Economics

Does the type of product that a country exports matter for subsequent economic performance?


Fig. 1. The product space. (A) Hierarchically clustered proximity (a) matrix representing the 775 SITC-4 product classes exported in the 1999–2000 period. (B) Network representation of the product space. Links are color coded with their proximity value. The sizes of the nodes are proportional to world trade, and their colors are chosen according to the classification introduced by Lerner.

Computational Social Science

Studying large scale social networks such as Wikipedia

Vizzards 2007 Entry

Computational Scientometrics

113 Years of Physical Review

Bruce W. Herr II and Russell Duhon (Data Mining & Visualization), Elisha F. Hardy (Graphic Design), Shashikant Penumarthi (Data Preparation) and Katy Börner (Concept)

Computational Epidemics

Forecasting (and preventing the effects of) the next pandemic.


2. NWB Challenges and Opportunities

- **Data**
  - Different data formats
  - Different data models

- **Algorithms**
  - Different research purposes (preprocessing, modeling, analysis, visualization, clustering)
  - Different implementations of the same algorithm
  - Different programming languages
  - Algorithm developers/users are not computer scientists

- **Different tools** (Pajek, UCINet, Guess, Cytoscape, R, …)

- **Different communities, practices, cultures**

---

### NWB Deliverables

**Network Workbench (NWB) Tool**
- A network analysis, modeling, and visualization toolkit for physics, biomedical, and social science research.
- Install and run on multiple Operating Systems.
- Supports many file formats.
- Easy integration of new algorithms thanks to CIShell/OSGi.

**Cyberinfrastructure Shell (CIShell)**
- An open source, software framework for the integration and utilization of datasets, algorithms, tools, and computing resources.
- Extends OSGi industry standard.
 CIShell – Serving Non-CS Algorithm Developers & Users

CIShell is built upon the Open Services Gateway Initiative (OSGi) Framework.

**OSGi** ([http://www.osgi.org](http://www.osgi.org)) is

- A standardized, component oriented, computing environment for networked services.
- Successfully used in the industry from high-end servers to embedded mobile devices since 8 years.
- Alliance members include IBM (Eclipse), Sun, Intel, Oracle, Motorola, NEC and many others.
- Widely adopted in open source realm, especially since Eclipse 3.0 that uses OSGi R4 for its plugin model.

**Advantages of Using OSGi**

- Any CIShell algorithm is a service that can be used in any OSGi-framework based system.
- Using OSGi, running CIShells/tools can connected via RPC/RMI supporting peer-to-peer sharing of data, algorithms, and computing power.

Ideally, CIShell becomes a standard for creating OSGi Services for algorithms.
The Network Workbench (NWB) tool supports researchers, educators, and practitioners interested in the study of biomedical, social and behavioral science, physics, and other networks.

In Aug. 2009, the tool provides more 160 plugins that support the preprocessing, analysis, modeling, and visualization of networks.

More than 40 of these plugins can be applied or were specifically designed for S&T studies.

It has been downloaded more than 35,000 times since Dec. 2006.
NWB Tool: Supported Data Formats

**Personal Bibliographies**
- Bibtex (.bib)
- Endnote Export Format (.enw)

**Data Providers**
- Web of Science by Thomson Scientific/Reuters (.isi)
- Scopus by Elsevier (.scopus)
- Google Scholar (access via Publish or Perish save as CSV, Bibtext, EndNote)
- Awards Search by National Science Foundation (.nsf)

**Scholarly Database** (all text files are saved as .csv)
- Medline publications by National Library of Medicine
- NIH funding awards by the National Institutes of Health (NIH)
- NSF funding awards by the National Science Foundation (NSF)
- U.S. patents by the United States Patent and Trademark Office (USPTO)
- Medline papers – NIH Funding

**Network Formats**
- NWB (.nwb)
- Pajek (.net)
- GraphML (.xml or .graphml)
- XGMML (.xml)

**Burst Analysis Format**
- Burst (.burst)

**Other Formats**
- CSV (.csv)
- Edgelist (.edge)
- Pajek (.mat)
- TreexML (.xml)
NWB Tool: Algorithms (July 1st, 2008)
See https://nwb.slis.indiana.edu/community and handout for details.

### NWB Tool: Output Formats

- NWB tool can be used for data conversion. Supported output formats comprise:
  - CSV (.csv)
  - NWB (.nwb)
  - Pajek (.net)
  - Pajek (.mat)
  - GraphML (.xml or .graphml)
  - XGMML (.xml)

- GUESS
  Supports export of images into common image file formats.
  - Horizontal Bar Graphs
  saves out raster and ps files.
NWB Tool Overview

1. Download, install, and run.
2. Load, view, convert, save data.
3. Read and visualize a directory hierarchy.
4. Load a network, compute its basic properties, and explore it in GUESS.
NWB Tool Overview

1. Download, install, and run.
2. Load, view, convert, save data.
3. Read and visualize a directory hierarchy.
4. Load a network, compute its basic properties, and explore it in GUESS.

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

NWB Tool 1.0.0
Select your operating system from the pull down menu.

Save as *.jar file.

Install and run.

Session log files are stored in ‘*yournwbdirectory*/logs’ directory.
NWB Tool Interface Components

File, Preprocessing, Modeling, and Visualization Menus
Analysis Menu and Submenus

Integrated Tools

Gnuplot
portable command-line driven interactive data and function plotting utility
http://www.gnuplot.info/.

GUESS
exploratory data analysis and visualization tool for graphs and networks.
https://nwb.slis.indiana.edu/community/?n=VisualizeData.GUESS.
In November 2008, the NWB tool supports loading the following input file formats:

- GraphML (*.xml or *.graphml)
- XGMML (*.xml)
- Pajek .NET (*.net) & Pajek .Matrix (*.mat)
- NWB (*.nwb)
- TreeML (*.xml)
- Edge list (*.edge)
- CSV (*.csv)
- ISI (*.isi)
- Scopus (*.scopus)
- NSF (*.nsf)
- BibTeX (*.bib)
- EndNote (*.enw)

and the following network file output formats:

- GraphML (*.xml or *.graphml)
- Pajek .MAT (*.mat)
- Pajek .NET (*.net)
- NWB (*.nwb)
- XGMML (*.xml)
- CSV (*.csv)

These formats are documented at
[https://nwb.slis.indiana.edu/community/?n=DataFormats.HomePage](https://nwb.slis.indiana.edu/community/?n=DataFormats.HomePage)

NWB Ecology of Data Formats and Converters

Not shown are 15 sample datasets, 45 data preprocessing, analysis, modeling and visualization algorithms, 9 services.

13 Supported data formats

6 Output formats for diverse visualization algorithms

8 Intermediate data formats

Supported by 35 data converters.
Network Workbench Marketplace: An Ecology of Data Formats, Converters, and Algorithms
Sample Datasets

The ‘*yournwbdirectory*/sampledata’ directory provides sample datasets from the biology, network, scientometrics, and social science research domains:

/biology
/network
/scientometrics
    /scientometrics/bibtex
    /scientometrics/csv
    /scientometrics/endnote
    /scientometrics/isi
    o FourNetSciResearchers.isi
    /scientometrics/nsf
    o Cornell.nsf
    o Indiana.nsf
    o Michigan.nsf
    /scientometrics/scopus
    /socialscience
    o florentine.nwb

The blue ones are used in this tutorial.

Property Files and Python Scripts

The ‘*yournwbdirectory*/’ directory also contains

/sampledata/scientometrics/properties  // Used to extract networks and merge data
    o bibtexCoAuthorship.properties
    o endnoteCoAuthorship.properties
    o isiCoAuthorship.properties
    o isiCoCitation.properties
    o isiPaperCitation.properties
    o mergeBibtexAuthors.properties
    o mergeEndnoteAuthors.properties
    o mergeIsiAuthors.properties
    o mergeNsfPIs.properties
    o mergeScopusAuthors.properties
    o nsfCoPI.properties
    o scopusCoAuthorship.properties

/sampledata/scripts/GUESS  // Used to do color/size/shape code networks
    o co-author-nw.py
    o co-PI-nw.py
    o paper-citation-nw.py
    o reference-co-occurrence-nw.py
NWB Tool Overview

1. Download, install, and run.
2. Load, view, convert, save data.
3. Read and visualize a directory hierarchy.
4. Load a network, compute its basic properties, and explore it in GUESS.

Load, View and Save (Convert) Data

Use 'File > Load File' to load florentine.nwb in sample datasets in ‘*yournwdirectory*/sampledata/socialscience’.

The loaded file will appear in the Data Manager window.

Right click loaded file to save, view, rename, or discard.
Data Converter Graph

There is no central data format.

Instead, data formats used in different communities and required by the different algorithms are supported.

NWB Tool Overview

1. Download, install, and run.
2. Load, view, convert, save data.
3. Read and visualize a directory hierarchy.
4. Load a network, compute its basic properties, and explore it in GUESS.

Network Workbench (http://nwb.slis.indiana.edu)
Use ‘File > Read Directory Hierarchy’ with parameters

Visualize resulting ‘Directory Tree - Prefuse (Beta) Graph’ using
- ‘Visualization > Tree View (prefuse beta)’
- ‘Visualization > Tree Map (prefuse beta)’
- ‘Visualization > Balloon Graph (prefuse alpha)’
- ‘Visualization > Radial Tree/Graph (prefuse alpha)’

Different views of the /nwb directory hierarchy.

Note the size of the /plugin directory.
NWB Tool Overview

1. Download, install, and run.
2. Load, view, convert, save data.
3. Read and visualize a directory hierarchy.
4. Load a network, compute its basic properties, and explore it in GUESS.

Select florentine.nwb in Data Manager.
- Run ‘Analysis > Network Analysis Toolkit (NAT)’ to get basic properties.

This graph claims to be undirected.
Nodes: 34
Isolated nodes: 1
Node attributes: present: label, wealth, totalties, priorities
Ndgps: 27
No self loops were discovered.
no parallel edges were discovered.
edge attributes:
Numeric attributes: Example value
marriage... T
business... T

did not detect any numeric attributes.
This network does not seem to be a valued network.

Average degree: 3.375
This graph is not weakly connected.
There are 2 weakly connected components. (1 isolates)
The largest connected component consists of 15 nodes.
Did not calculate strong connectedness because this graph was not directed.

Density (disregarding weights): 0.025

- Select network and run ‘Visualization > GUESS’ to open GUESS with file loaded.
- Apply ‘Layout -> GEM’.
Pan:
“grab” the background by holding left-click and moving your mouse.

Zoom:
Using scroll wheel, press the “+” and “-“ buttons in the upper-left hand corner, or right-click and move the mouse left or right. Center graph by selecting ‘View -> Center’.

Select to select/move single nodes. Hold down ‘Shift’ to select multiple.

Right click to modify Color, etc.
Graph Modifier:
Select “all nodes” in the Object drop-down menu and click ‘Show Label’ button.

Select “nodes based on ->”, then select “wealth” from the Property drop-down menu, “=>” from the Operator drop-down menu, and finally type “50” into the Value box. Then a color/size/shape code.

Network Workbench (http://nwb.slis.indiana.edu).

Interpreter:
Uses Jython a combination of Java and Python.

Try
```
colorize(wealth, white, red)
```
```
resizeLinear(sitebetweenness, 5, 25)
```
Workflow Design Primer

Modularity at data preprocessing/analysis/modeling level.

Modularity at visualization level:
➢ ‘Data Layers’ are used in GIS systems to support the visual layering and coordination of different datasets, e.g., water pipes, streets, electricity lines, etc.
➢ ‘Design Layers’ supported by graphic design software such as Photoshop or Dreamweaver enable the separate design and modular composition of design elements.
➢ ‘Visualization Layers’ define distinct parts with very specific functionality that collectively define a visualization.
Exemplary Analyses and Visualizations

Individual Level
A. Loading ISI files of major network science researchers, extracting, analyzing and visualizing paper-citation networks and co-author networks.
B. Loading NSF datasets with currently active NSF funding for 3 researchers at Indiana U

Institution Level

Scientific Field Level
D. Extracting co-author networks, patent-citation networks, and detecting bursts in SDB data.
Data Acquisition from Web of Science

Download all papers by
- Eugene Garfield
- Stanley Wasserman
- Alessandro Vespignani
- Albert-László Barabási

from
- Science Citation Index Expanded (SCI-EXPANDED) --1955-present
- Social Sciences Citation Index (SSCI) --1956-present
- Arts & Humanities Citation Index (A&HCI) --1975-present

Comparison of Counts
No books and other non-WoS publications are covered.

<table>
<thead>
<tr>
<th>Age</th>
<th>Total # Cites</th>
<th>Total # Papers</th>
<th>H-Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eugene Garfield</td>
<td>82</td>
<td>1,525</td>
<td>672</td>
</tr>
<tr>
<td>Stanley Wasserman</td>
<td>122</td>
<td>122</td>
<td>35</td>
</tr>
<tr>
<td>Alessandro Vespignani</td>
<td>42</td>
<td>451</td>
<td>101</td>
</tr>
<tr>
<td>Albert-László Barabási</td>
<td>40</td>
<td>2,218</td>
<td>126</td>
</tr>
<tr>
<td></td>
<td>41</td>
<td>16,920</td>
<td>159</td>
</tr>
</tbody>
</table>
Network Extraction

Sample paper network (left) and four different network types derived from it (right)

From ISI files, about 30 different networks can be extracted.

Extract Co-Author Network

Load *yourworkingdirectory*/sampledata/scientometrics/isi/FourNetSciResearchers.isi’ using ‘File > Load and Clean ISI File’.

To extract the co-author network, select the ‘361 Unique ISI Records’ table and run ‘Scientometrics > Extract Co-Author Network’ using isi file format:

The result is an undirected network of co-authors in the Data Manager. It has 247 nodes and 891 edges.

To view the complete network, select the network and run ‘Visualization > GUESS > GEM’. Run Script > Run Script…. And select Script folder > GUESS > co-author-nw.py.
Comparison of Co-Author Networks

Eugene Garfield

Stanley Wasserman

Alessandro Vespignani

Albert-László Barabási

Joint Co-Author Network of all Four NetsSci Researchers
Load `yournwbdirectory*/sampledata/scientometrics/isi/FourNetSciResearchers.isi` using 'File > Load and Clean ISI File'.

To extract the paper-citation network, select the ‘361 Unique ISI Records’ table and run 'Scientometrics > Extract Directed Network' using the parameters:

The result is a directed network of paper citations in the Data Manager. It has 5,335 nodes and 9,595 edges.

To view the complete network, select the network and run 'Visualization > GUESS'. Run 'Script > Run Script …' and select `yournwbdirectory*/script/GUESS/paper-citation-nw.py'.
Exemplary Analyses and Visualizations

Individual Level
A. Loading ISI files of major network science researchers, extracting, analyzing and visualizing paper-citation networks and co-author networks.
B. Loading NSF datasets with currently active NSF funding for 3 researchers at Indiana U

Institution Level

Scientific Field Level
D. Extracting co-author networks, patent-citation networks, and detecting bursts in SDB data.

NSF Awards Search via http://www.nsf.gov/awardsearch

Save in CSV format as *name*.nsf
NSF Awards Search Results

<table>
<thead>
<tr>
<th>Name</th>
<th># Awards</th>
<th>First A. Starts</th>
<th>Total Amount to Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geoffrey Fox</td>
<td>27</td>
<td>Aug 1978</td>
<td>12,196,260</td>
</tr>
<tr>
<td>Michael McRobbie</td>
<td>8</td>
<td>July 1997</td>
<td>19,611,178</td>
</tr>
<tr>
<td>Beth Plale</td>
<td>10</td>
<td>Aug 2005</td>
<td>7,224,522</td>
</tr>
</tbody>
</table>

**Disclaimer:**

Only NSF funding, no funding in which they were senior personnel, only as good as NSF's internal record keeping and unique person ID. If there are 'collaborative' awards then only their portion of the project (award) will be included.

Using NWB to Extract Co-PI Networks

- Load into NWB, open file to count records, compute total award amount.
- Run ‘Scientometrics > Extract Co-Occurrence Network’ using parameters:
  - Select ‘Extracted Network ..’ and run ‘Analysis > Network Analysis Toolkit (NAT)’
  - Remove unconnected nodes via ‘Preprocessing > Delete Isolates’.
  - ‘Visualization > GUESS’, layout with GEM
  - Run ‘co-PI-nw.py’ GUESS script to color/size code.
Exemplary Analyses and Visualizations

Individual Level
A. Loading ISI files of major network science researchers, extracting, analyzing and visualizing paper-citation networks and co-author networks.
B. Loading NSF datasets with currently active NSF funding for 3 researchers at Indiana U

Institution Level

Scientific Field Level
D. Extracting co-author networks, patent-citation networks, and detecting bursts in SDB data.
Active NSF Awards on 11/07/2008:

- Indiana University 257
  (there is also Indiana University at South Bend Indiana University Foundation, Indiana University Northwest, Indiana University-Purdue University at Fort Wayne, Indiana University-Purdue University at Indianapolis, Indiana University-Purdue University School of Medicine)

- Cornell University 501
  (there is also Cornell University – State, Joan and Sanford I. Weill Medical College of Cornell University)

- University of Michigan Ann Arbor 619
  (there is also University of Michigan Central Office, University of Michigan Dearborn, University of Michigan Flint, University of Michigan Medical School)

Active NSF Awards on 09/10/2009:

- Stanford University 429

Save files as csv but rename into .nsf.
Or simply use the saved in "*yourwbdirectory*/sampledata/scientometrics/ nsf/".
Extracting Co-PI Networks

Load NSF data, selecting the loaded dataset in the Data Manager window, run ‘Scientometrics > Extract Co-Occurrence Network’ using parameters:

Two derived files will appear in the Data Manager window: the co-PI network and a merge table. In the network, nodes represent investigators and edges denote their co-PI relationships. The merge table can be used to further clean PI names.

Running the ‘Analysis > Network Analysis Toolkit (NAT)’ reveals that the number of nodes and edges but also of isolate nodes that can be removed running ‘Preprocessing > Delete Isolates’.

Select ‘Visualization > GUESS’ to visualize. Run ‘co-PI-nw.py’ script.

Indiana U: 223 nodes, 312 edges, 52 components

U of Michigan: 497 nodes, 672 edges, 117 components

Cornell U: 375 nodes, 573 edges, 78 components
**Extract Giant Component**

Select network after removing isolates and run *Analysis > Unweighted and Undirected > Weak Component Clustering* with parameter

![Weak Component Clustering](image)

Indiana’s largest component has 19 nodes, Cornell’s has 67 nodes, Michigan’s has 55 nodes.

Visualize Cornell network in GUESS using same .py script and save via *File > Export Image* as jpg.

![Largest component of Cornell U co-PI network](image)

Node size/color ~ total award money
Top-50 total award money nodes are labeled.
Top-10 Investigators by Total Award Money

```python
for i in range(0, 10):
    print str(nodesbytotalawardmoney[i].label) + ": " +
    str(nodesbytotalawardmoney[i].totalawardmoney)
```

**Indiana University**
- Curtis Lively: 7,436,828
- Frank Lester: 6,402,330
- Maynard Thompson: 6,402,330
- Michael Lynch: 6,361,796
- Craig Stewart: 6,216,352
- William Snow: 5,434,796
- Douglas V. Houweling: 5,068,122
- James Williams: 5,068,122
- Miriam Zolan: 5,000,627
- Carla Caceres: 5,000,627

**Cornell University**
- Maury Tigner: 107,216,976
- Sandip Tiwari: 72,094,578
- Sol Gruner: 48,469,991
- Donald Bilderback: 47,360,053
- Ernest Fontes: 29,380,053
- Hasan Padamsee: 18,292,000
- Melissa Hines: 13,099,545
- Daniel Huttenlocher: 7,614,326
- Timothy Fahey: 7,223,112
- Jon Kleinberg: 7,165,507

**Michigan University**
- Khalil Najafi: 32,541,158
- Kensing Wise: 32,164,404
- Jacquelyne Eccles: 25,890,711
- Georg Raithel: 23,832,421
- Roseanne Sension: 23,812,921
- Theodore Norris: 23,35,0921
- Paul Berman: 23,350,921
- Roberto Merlin: 23,350,921
- Robert Schoeni: 21,991,140
- Wei-Jun Jean Yeung: 21,991,140

**Stanford University**
429 active NSF awards on 09/10/2009
Largest component
39 nodes

Stanford U:
218 nodes, 285 edges, 49 components
157 isolate nodes were deleted
Top-10 Investigators by Total Award Money

for i in range(0, 10):
    print str(nodesbytotalawardmoney[i].label) + "
    print str(nodesbytotalawardmoney[i].totalawardmoney)

Stanford University
Dan Boneh: 11,837,800
Rajeev Motwani: 11,232,154
Hector Garcia-Molina: 10,577,906
David Goldhaber-Gordon: 9,792,029
Kathryn Moler: 7,870,029
John C. Mitchell: 7,290,668
Alfred Spormann: 6,803,000
Gordon Brown: 6,158,000
Jennifer Widom: 5,661,311

Search Results
Bad:
Results are sorted by award date, with the most recent awards at the top. Click on a column heading to re-sort the results. The up/down arrows at the right of each column title control whether the sort is ascending or descending.
To view the abstract, click on the award number or title. Click on the data in other columns to perform a new search with that parameter.
Refine Search
344 awards found, displaying 1 to 50.
[First/Prev] 1, 5, 10, 15, 20, 25, 30, 35 [Next/Last]

Search for all active NSF awards by Northwestern University on 9/2/2009 via
http://www.nsf.gov/awardsearch
Nodes: 323, Edges: 313, Average degree: 1.9, 149 weakly connected components. (107 isolates)

Giant component has 63 nodes,
Color and size coding by total award money

3. Exemplary Analyses and Visualizations

Individual Level
A. Loading ISI files of major network science researchers, extracting, analyzing and visualizing paper-citation networks and co-author networks.
B. Loading NSF datasets with currently active NSF funding for 3 researchers at Indiana U

Institution Level

Scientific Field Level
D. Extracting co-author networks, patent-citation networks, and detecting bursts in SDB data.
Goto: http://sdb.slis.indiana.edu
Email: nwb@indiana.edu
Password: nwb
Open and Preprocess SDB zip file

Load `medline_medline_master.csv` to NWB.

Run 'Preprocessing > Normalize Text' with a space as New Separator.
Run ‘Analysis > Textual > Burst Detection’ with parameters:
and space as a separator.

Sort result by burst weight

<table>
<thead>
<tr>
<th>Word</th>
<th>Length</th>
<th>Weight</th>
<th>Strength</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>care</td>
<td>1</td>
<td>Infinity</td>
<td>Infinity</td>
<td>1988</td>
<td>1988</td>
</tr>
<tr>
<td>water</td>
<td>1</td>
<td>29.8883</td>
<td>29.8883</td>
<td>2002</td>
<td>2002</td>
</tr>
<tr>
<td>countri</td>
<td>10</td>
<td>27.03612</td>
<td>27.03612</td>
<td>1990</td>
<td>1999</td>
</tr>
<tr>
<td>protect</td>
<td>1</td>
<td>26.88557</td>
<td>26.88557</td>
<td>2002</td>
<td>2002</td>
</tr>
<tr>
<td>farm</td>
<td>1</td>
<td>23.32114</td>
<td>23.32114</td>
<td>2005</td>
<td>2005</td>
</tr>
<tr>
<td>villag</td>
<td>2</td>
<td>23.273</td>
<td>40.65081</td>
<td>2008</td>
<td></td>
</tr>
<tr>
<td>crop</td>
<td>2</td>
<td>22.33649</td>
<td>30.42535</td>
<td>2008</td>
<td></td>
</tr>
<tr>
<td>blood</td>
<td>5</td>
<td>22.12166</td>
<td>22.12166</td>
<td>1996</td>
<td>2000</td>
</tr>
</tbody>
</table>

Network Workbench (http://nwb.slis.indiana.edu)
Top-10 burst terms from abstracts of the AI search results.

**Medline**

<table>
<thead>
<tr>
<th>Word</th>
<th>Length</th>
<th>Weight</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>medical</td>
<td>17</td>
<td>299.7924</td>
<td>1993</td>
<td>1999</td>
</tr>
<tr>
<td>knowledge</td>
<td>5</td>
<td>293.6575</td>
<td>1991</td>
<td>1993</td>
</tr>
<tr>
<td>knowledge</td>
<td>6</td>
<td>215.2407</td>
<td>1997</td>
<td>2002</td>
</tr>
<tr>
<td>expert</td>
<td>13</td>
<td>171.0443</td>
<td>1994</td>
<td>1997</td>
</tr>
<tr>
<td>systems</td>
<td>15</td>
<td>170.3306</td>
<td>1995</td>
<td>1999</td>
</tr>
<tr>
<td>intelligence</td>
<td>21</td>
<td>123.8794</td>
<td>1991</td>
<td>2001</td>
</tr>
<tr>
<td>patent</td>
<td>21</td>
<td>123.9297</td>
<td>1962</td>
<td>2002</td>
</tr>
<tr>
<td>care</td>
<td>12</td>
<td>106.5222</td>
<td>1999</td>
<td>2001</td>
</tr>
<tr>
<td>registration</td>
<td>5</td>
<td>104.8139</td>
<td>2005</td>
<td></td>
</tr>
<tr>
<td>knowledge-based</td>
<td>16</td>
<td>98.83778</td>
<td>1987</td>
<td>2002</td>
</tr>
</tbody>
</table>

**NIH**

<table>
<thead>
<tr>
<th>Word</th>
<th>Length</th>
<th>Weight</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>choice</td>
<td>3</td>
<td>117.2205</td>
<td>1993</td>
<td>2000</td>
</tr>
<tr>
<td>commercial</td>
<td>9</td>
<td>87.79158</td>
<td>1995</td>
<td></td>
</tr>
<tr>
<td>proposed</td>
<td>6</td>
<td>87.79158</td>
<td>1995</td>
<td></td>
</tr>
<tr>
<td>review</td>
<td>9</td>
<td>83.89652</td>
<td>1978</td>
<td>1980</td>
</tr>
<tr>
<td>protein</td>
<td>1</td>
<td>72.12888</td>
<td>1988</td>
<td>1989</td>
</tr>
<tr>
<td>networks</td>
<td>4</td>
<td>71.2252</td>
<td>1993</td>
<td>1996</td>
</tr>
<tr>
<td>patterns</td>
<td>3</td>
<td>66.44826</td>
<td>1977</td>
<td>1979</td>
</tr>
<tr>
<td>being</td>
<td>8</td>
<td>66.29254</td>
<td>1971</td>
<td>1978</td>
</tr>
<tr>
<td>reasoning</td>
<td>5</td>
<td>65.68178</td>
<td>1984</td>
<td>1985</td>
</tr>
<tr>
<td>expert</td>
<td>4</td>
<td>60.49975</td>
<td>1987</td>
<td>1990</td>
</tr>
</tbody>
</table>

**NSF**

<table>
<thead>
<tr>
<th>Word</th>
<th>Length</th>
<th>Weight</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>flow</td>
<td>6</td>
<td>47.05997</td>
<td>1999</td>
<td></td>
</tr>
<tr>
<td>gray</td>
<td>2</td>
<td>28.19508</td>
<td>2000</td>
<td>2004</td>
</tr>
<tr>
<td>learning</td>
<td>2</td>
<td>27.47028</td>
<td>1997</td>
<td>1999</td>
</tr>
<tr>
<td>human</td>
<td>5</td>
<td>25.4525</td>
<td>2000</td>
<td></td>
</tr>
<tr>
<td>control</td>
<td>2</td>
<td>24.02877</td>
<td>1992</td>
<td>1993</td>
</tr>
<tr>
<td>knowledge</td>
<td>1</td>
<td>21.48756</td>
<td>1993</td>
<td>1998</td>
</tr>
<tr>
<td>students</td>
<td>1</td>
<td>21.07674</td>
<td>1997</td>
<td>1997</td>
</tr>
<tr>
<td>problems</td>
<td>2</td>
<td>20.17755</td>
<td>1998</td>
<td>1999</td>
</tr>
<tr>
<td>more</td>
<td>2</td>
<td>19.56109</td>
<td>2000</td>
<td>2001</td>
</tr>
<tr>
<td>use</td>
<td>1</td>
<td>19.83503</td>
<td>2001</td>
<td>2001</td>
</tr>
</tbody>
</table>

**USPTO**

<table>
<thead>
<tr>
<th>Word</th>
<th>Length</th>
<th>Weight</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>3</td>
<td>19.03947321</td>
<td>2004</td>
<td>2006</td>
</tr>
<tr>
<td>video</td>
<td>3</td>
<td>15.3273542</td>
<td>1998</td>
<td>2000</td>
</tr>
<tr>
<td>disclosed</td>
<td>2</td>
<td>14.0569471</td>
<td>1999</td>
<td>2000</td>
</tr>
<tr>
<td>neural</td>
<td>3</td>
<td>13.30105906</td>
<td>2004</td>
<td>2006</td>
</tr>
<tr>
<td>&quot;current&quot;</td>
<td>2</td>
<td>12.4328047</td>
<td>1999</td>
<td>2000</td>
</tr>
<tr>
<td>unit</td>
<td>2</td>
<td>12.35743838</td>
<td>2003</td>
<td>2004</td>
</tr>
<tr>
<td>material</td>
<td>1</td>
<td>12.09487303</td>
<td>2000</td>
<td>2000</td>
</tr>
<tr>
<td>feedback</td>
<td>1</td>
<td>12.0778195</td>
<td>2000</td>
<td>2000</td>
</tr>
<tr>
<td>role</td>
<td>1</td>
<td>12.0778195</td>
<td>2000</td>
<td>2000</td>
</tr>
<tr>
<td>elevator</td>
<td>4</td>
<td>11.83351857</td>
<td>1991</td>
<td>1994</td>
</tr>
</tbody>
</table>
Bonus: Sci² Tool

Sci² Tool
Blondel Community Detection
Circular Hierarchy
RefMapper Output

Please read the documentation at http://sci.sis.indiana.edu/refmapper.pdf for information on how to get started with the tool, to learn how journal names are identified in PDF files of references and science map overlays are generated, and for guidance on how to interpret the resulting science map overlays presented in this document.

Date and time of analysis: June 5, 2009 10:18:47 AM EDT
Input directory: C:\Users\User\Desktop\RefMapper\Data\Scuip-2008\Implementing-science-policy

6 PDF files found
112 references identified in all files.
107 references with journal names.

The references with identified journal names are shown below:

Overlay on the UCSD Map of Science
Provided by the Cyberinfrastructure for Network Science Center at Indiana University.

Introduction
L. G. Wilson writes in Confront: The Unity of Knowledge (1988) that the features that distinguish science from pseudo sciences are repeatability, accuracy, instrumentation, and correlation.

Needs Analysis
An early part of the "The State of Science Policy: A Macroscopic View" NSF STP-07FR1 is a series of interviews with policy makers conducted to identify what science of science research has been productive and effective. In total, 50 formal, one-hour interviews have been conducted with science policy makers at university, campus level, program officer level, and division director level for governmental, state, and private foundations. Data compilation will start in October 2008 and resulting report can be ordered by sending a request to Mark Price (mark.price@indiana.edu).

Conceptualization of Science
A "science of science" requires a universally generally and personally social investigation of the structure and evaluation of science. A special journal issue entitled "Science of Science: Conceptualizations and Models of Science" is edited by R. Schrank, Indiana University & Andreas Schreiber, Royal Netherlands Academy of Arts and Sciences. Invitations to contribute are invited. The special issue will be published in the Journal of Information 31(1) in January 2009.

Scholarly Database
The Scholarly Database (SDB) at Indiana University aims to serve researchers and practitioners interested in the analysis, modeling, and visualization of large-scale scholarly datasets. The database currently provides access to over 50 million papers, patents and grants. Resulting datasets can be downloaded in bulk. Register for free access at https://sdb.slis.indiana.edu.

Cyberinfrastructures
The Cyberinfrastructures of the Network Science Center (NSC) provide a unique distributed, shared research environment for large-scale network analysis, modeling, and visualization. Theses Scientific SE, Open Science and Scholar, Data, 

http://sci.slis.indiana.edu

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