VIVO Data and Visualizations Preparations

Slides and sample data files are available at:
http://wiki.cns.iu.edu/display/PRES/VIVO+Presentation+Information

If you have not previously installed the Sci2 tool and downloaded the sample files, please ask for assistance. We have USB sticks with the needed files.

Please set cell phones to silent.

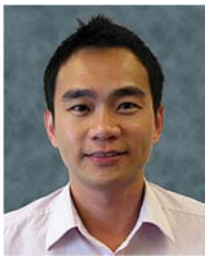
Check out http://scimaps.org for ideas on how to design visualizations.

Welcome and thank you for your interest!

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

VIVO Data and Visualizations

VIVO connect • share • discover



Robert Light and Chin Hua Kong
Cyberinfrastructure for Network Science Center
Department of Information and Library Science, School of Informatics and Computing
Indiana University Bloomington
http://cns.iu.edu

Presentation Overview

Part One

- Introduction
- VIVO Visualization Review
- VIVO Data Storage and Retrieval
  - Ontology
  - Basic SPARQL
  - Data Cleaning
- Introduction to Sci2
- Temporal Bar Graph
- Break

Part Two

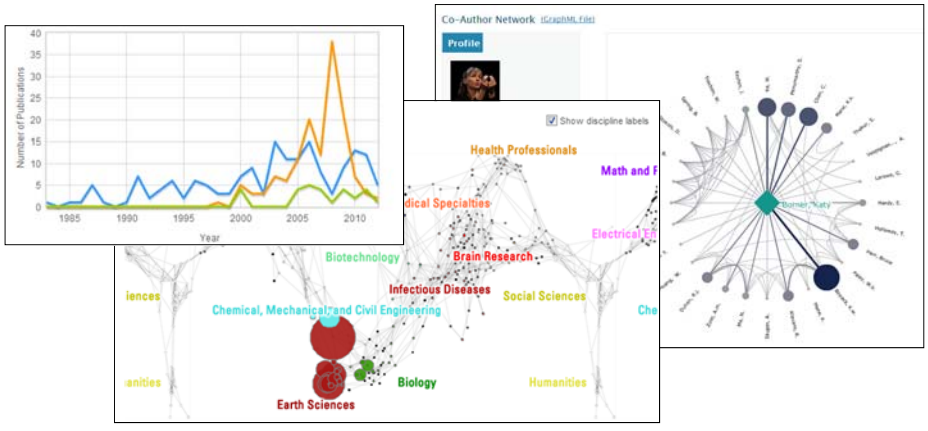
- Analyzing and visualizing VIVO data
  - Burst Detection
  - Map of Science
  - Network
- The next level of aggregated data
  - iNRN (http://nrn.cns.iu.edu)
  - VIVOSearch.org
- Continuing Education and Opportunities

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

VIVO Visualizations

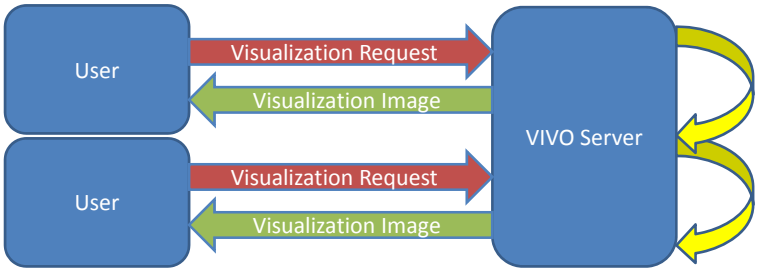
VIVO offers a treasure trove of data, but making sense of it all is often difficult when looked at one entity at a time.

Visualization brings large amounts of data together in a single graphic to allow for greater understanding.



VIVO Visualizations

The processing for these visualizations is done on the server-side, rather than on your desktop.



Minimal Processing Required

Handles Data Processing and Image Creation

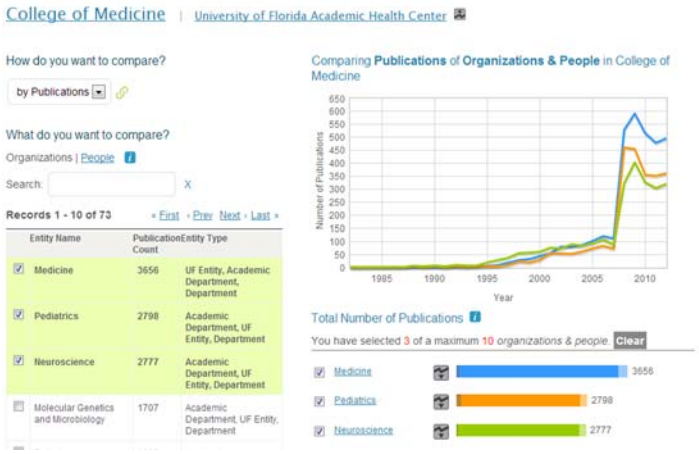
Visualization is a BLACK BOX

Requested Visualizations are Cached

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Trends over Time

Time is a major factor in many research questions, so it is accordingly a strong theme in visualization.



http://vivo.ufl.edu/vis/publication-graph/n166017

The UCSD Map Of Science

Originally created in 2006 and updated in 2011.

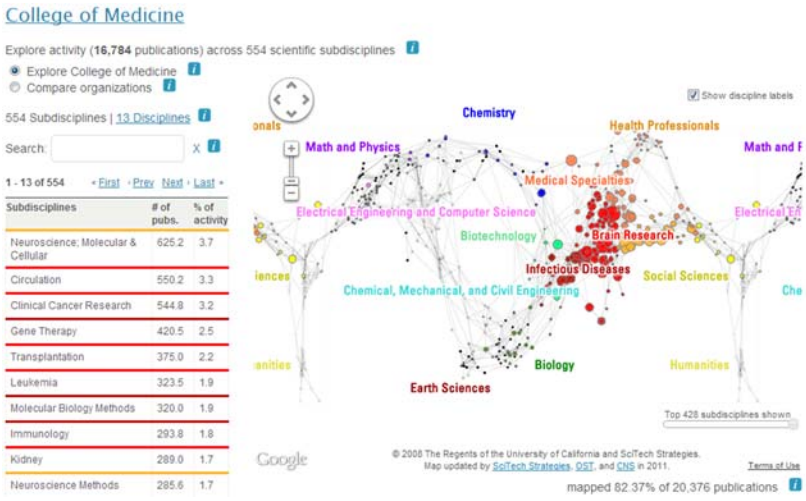
Created 554 clusters of journals called sub-disciplines based on citations and keywords, divided into 13 disciplines.

Maps over 25,000 journals from Scopus and Thomson Reuters' Web of Science as of 2011 update.

Börner, Katy, Richard Klavans, Michael Patek, Angela Zoss, Joseph R. Biberstine, Robert Light, Vincent Larivière, and Kevin W. Boyack (2012) Design and Update of a Classification System: The UCSD Map of Science. *PLoS ONE* 7(7): e39464. doi:10.1371/journal.pone.0039464

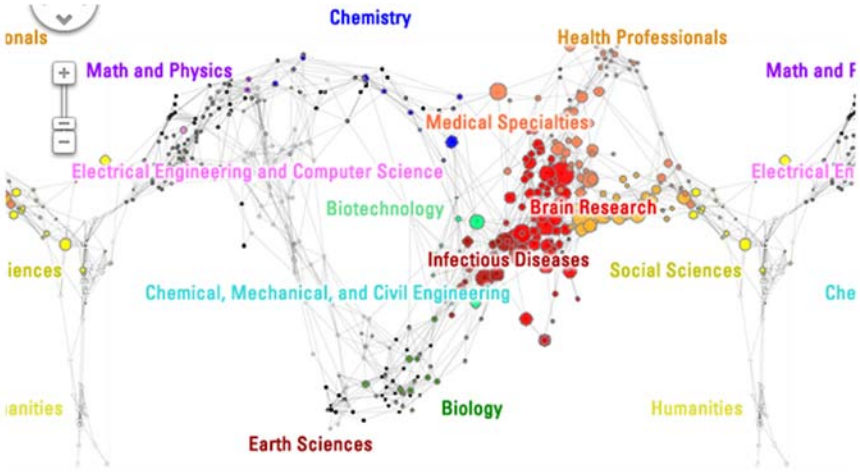
CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

The Map Of Science



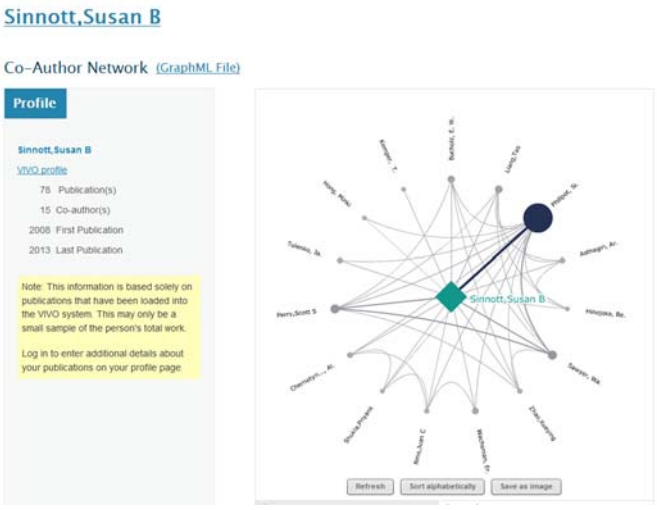
http://vivo.ufl.edu/vis/map-of-science/n166017

The Map Of Science



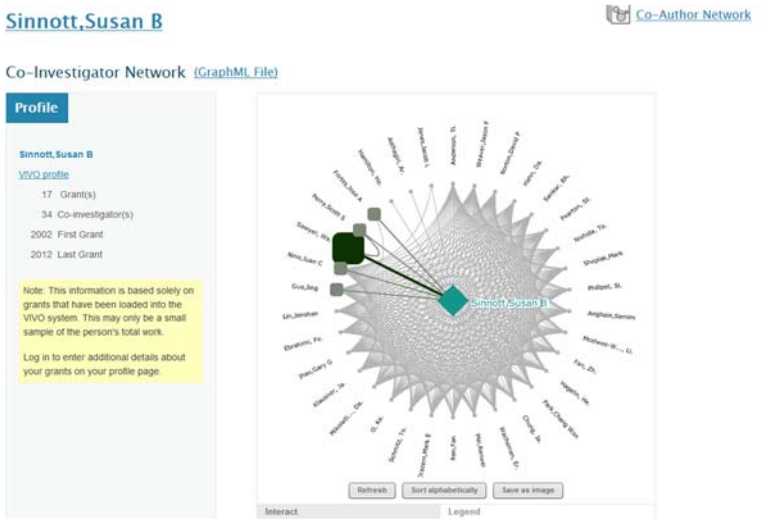
http://vivo.ufl.edu/vis/map-of-science/n166017

Co-Authorship



https://vivo.ufl.edu/vis/author-network/n1532

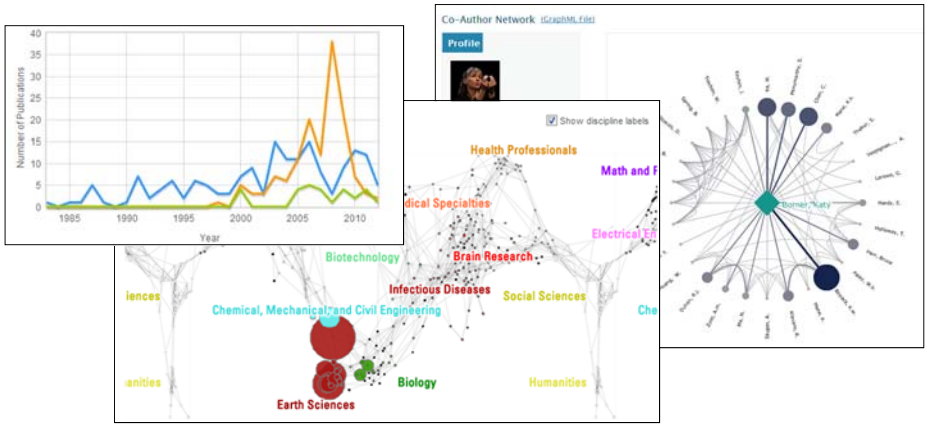
Co-Investigatorship



https://vivo.ufl.edu/vis/investigator-network/n1532

Going Beyond

These are certainly good visualizations and very useful, but do they fill every need?



Visualization Design

1. Identify your question
2. Identify the visualization that will answer it
3. Identify and acquire the data that you need to create the visualization
4. Visualize!

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Visualization Design

Good visualization starts with a question!

- How many of our grants are still active?
- How are the topics we're studying have changed over time?
- Which researchers are working with which funding agencies?
- How well is my department collaborating in terms of co-authorships?

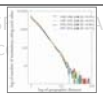
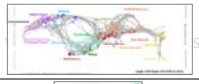
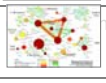
Visualization Design

What will answer your question?

- "When did…" – Look at timelines.
- "Where is…" – Physical location suggests a geomap.
- "With whom…" – You're looking at connections, so it'll probably be a network.
- "What subject areas…" – Topic mapping.

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Visualization Design

	Micro/Individual (1-100 records)	*Meso/Local (101–10,000 records)*	*Macro/Global (10,000 < records)*
Statistical Analysis/Profiling	Individual person and their expertise profiles	Larger labs, centers, universities, research domains, or states	All of NSF, all of scienc 
Temporal Analysis (When)	Funding portfolio of one individual	 opic bursts s of PNAS	113 Years of P Research 
Geospatial Analysis (Where)	Career trajectory of one individual	Mapping a st intellectual la pe 	PNAS pu 
Topical Analysis (What)	 s.	 flows in research	VxOrd/Topic m NIH funding 
Network Analysis (With Whom?)	NSF C rk of one in 	 netw 	NIH's core c y 

Visualization Design

All this leads up to deciding what DATA will power your visualization.

- People
- Organizations
- Article Names
- Article Venues
- Publication Dates
- Article Keywords
- Grant Names
- Funding Agencies
- Grant Subjects
- Grant Values
- Grant Durations
- Locations
- Affiliation Dates
- Courses Taught
- PhD Fellows
- Awards
- Titles
- Degrees

And, of course, the connections that link them all together.

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Image attributed to Jim Linwood (https://commons.wikimedia.org/wiki/File:Don't_Panic_Badge.jpg)

Data Selection and Query Design

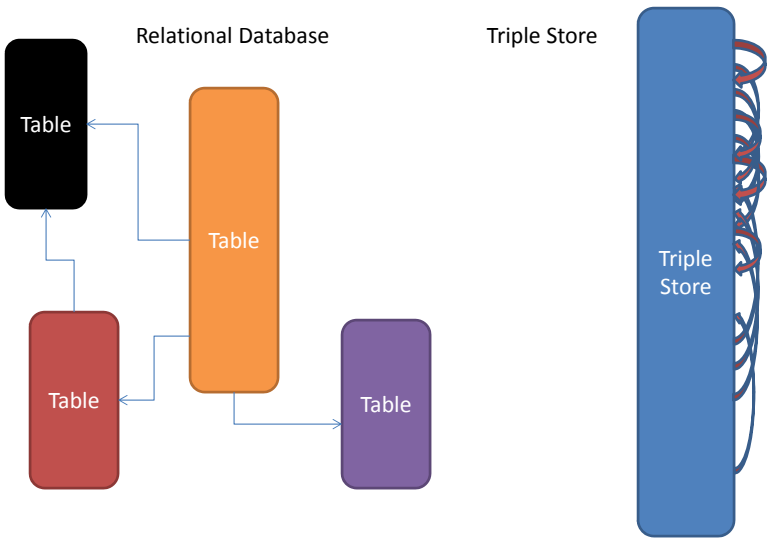
First, determine if you have the needed data. If it's not in your VIVO instance, it's not going to show up on your visualization, no matter what. If you do, then you can think about how to get at it.

Building a good visualization is often an iterative process. Your dataset will very likely evolve as your visualization does.

Always save your queries and data-cleaning steps so you can modify and improve them as your work evolves.

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Data Storage and Retrieval in VIVO



Data Storage in VIVO

All data is stored in the same format, via triples.

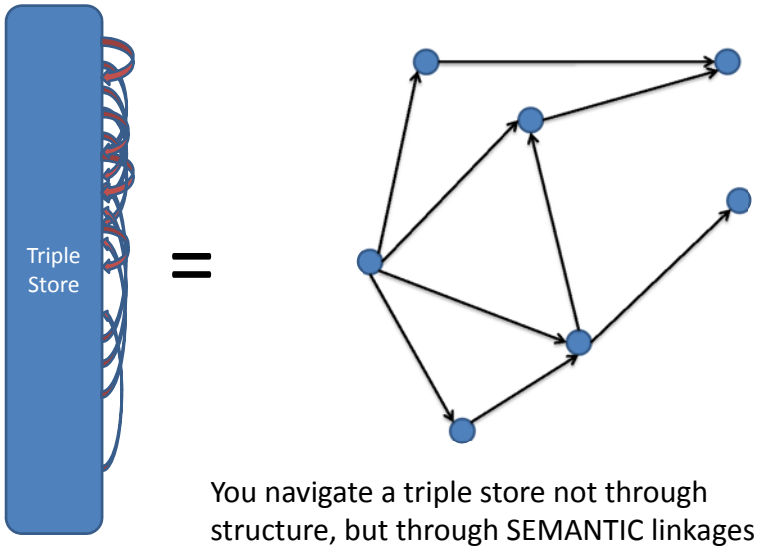
Subject – Predicate – Object

For example…

Robert Light – presented – VIVO Data Workshop
Katy Borner – wrote – *Atlas of Science*
Chin Hua Kong – slept through – whole trip here

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Data Storage and Retrieval in VIVO



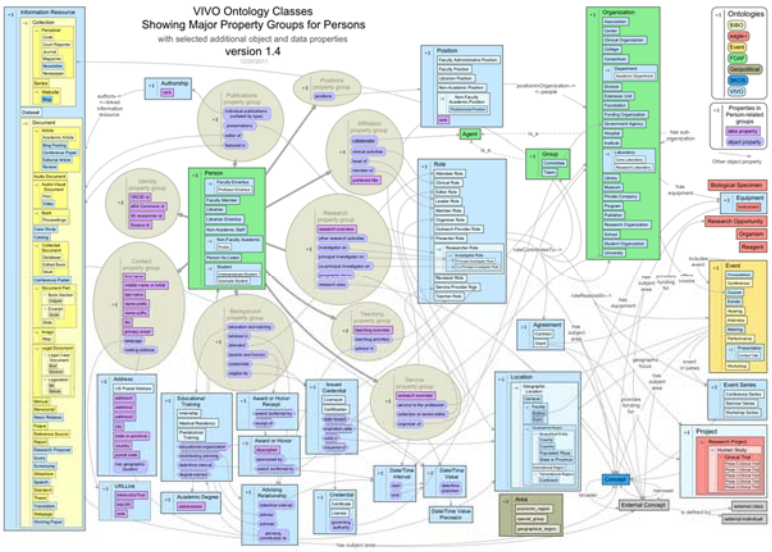
You navigate a triple store not through structure, but through SEMANTIC linkages.

Data Storage in VIVO

There is no table structure to guide data collection in a triple store. Instead, you have the **ontology**.

The ontology defines how data is linked together, and is critical when it comes to retrieving that data.

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations



https://wiki.duraspace.org/display/VIVO/VIVO+1.4+Ontology+overview+diagram

Image (again) attributed to Jim Linwood (https://commons.wikimedia.org/wiki/File:Don't_Panic_Badge.jpg)

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Data Storage in VIVO

You don't need to know the whole ontology, just how to find the entities and connections you need.

https://wiki.duraspace.org/display/VIVO/VIVO+Ontology+Classes+and+Properties.v1.4.1

Data Storage in VIVO

These are the types of links and data fields associated with a Person.

vivo	**adviseeIn**	vivo:AdvisingRelationship
vivo	**advisorIn**	vivo:AdvisingRelationship
vivo	awardOrHonor	vivo:AwardReceipt
vivo	**currentlyHeadOf**	foaf:Organization
vivo	domesticGeographicFocus	vivo:GeographicRegion
vivo	**editorOf**	bibo:Document
vivo	**educationalTraining**	vivo:EducationalTraining
vivo	**eligibleFor**	vivo:Credential
vivo	geographicFocus	vivo:GeographicRegion
vivo	hasCo-PrincipalInvestigatorRole	vivo:Co-PrincipalInvestigatorRole
vivo	**hasCredential**	vivo:IssuedCredential
vivo	**hasEditorRole**	vivo:EditorRole
vivo	**hasInvestigatorRole**	vivo:InvestigatorRole
vivo	**hasPrincipalInvestigatorRole**	vivo:PrincipalInvestigatorRole
vivo	**hasResearchArea**	owl:Thing
vivo	**hasReviewerRole**	vivo:ReviewerRole
vivo	**hasServiceProviderRole**	vivo:ServiceProviderRole
vivo	**hasTeacherRole**	vivo:TeacherRole
vivo	internationalGeographicFocus	vivo:GeographicRegion
vivo	**personInPosition**	vivo:Position

vivo	**eRACommonsId**
foaf	**firstName**
foaf	**lastName**
vivo	**middleName**
vivo	**orcidId**
vivo	**outreachOverview**
vivo	**overview**
vivo	**preferredTitle**
bibo	**prefixName**
vivo	**researcherId**
vivo	**researchOverview**
vivo	**scopusId**
bibo	**suffixName**
vivo	**teachingOverview**

https://wiki.duraspace.org/display/VIVO/VIVO+Ontology+Classes+and+Properties.v1.4.1

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Querying Data from VIVO

With some understanding of the ontology, we're ready to start writing SPARQL queries. This is the equivalent to SQL in a relational database.

We'll build a few SPARQL queries now, then later we'll use the results from those queries when we start working with the Sci2 tool.

Basic SPARQL

Remember that all data is in one table in the form of:

Subject – Predicate – Object

Unlike SQL, you don't have a table structure to guide your query design. But you DO have an ontology.

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Basic SPARQL

Goal: Build a query with publication data that we can use for visualization.

We'll need:

Titles – A meaningful identifier
Journals – For topic mapping
Publication Years – For burst mapping
Authors – For authorship networks

We'll start simply with the first two, then iterate down the list.

Basic SPARQL

```
PREFIX rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX vivo: <http://vivoweb.org/ontology/core#>
PREFIX bibo: <http://purl.org/ontology/bibo/>

SELECT ?AcademicArticle1 ?Venue1
WHERE
{
   ?AcademicArticle1 rdf:type bibo:AcademicArticle .
   ?AcademicArticle1 vivo:hasPublicationVenue ?Venue1 .
}
LIMIT 3
```

Prefixes serve a similar function to XML in that they let us avoid having to write out full URIs over and over.

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Basic SPARQL

```
PREFIX rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX vivo: <http://vivoweb.org/ontology/core#>
PREFIX bibo: <http://purl.org/ontology/bibo/>

SELECT ?AcademicArticle1 ?Venue1
WHERE
{
   ?AcademicArticle1 rdf:type bibo:AcademicArticle .
   ?AcademicArticle1 vivo:hasPublicationVenue ?Venue1 .
}
LIMIT 3
```

This is the critical section where you define two things:

1) What triples you want to examine (from the WHERE clause).
   Think of these like equations in algebra.
2) What information you want from them. (from the SELECT statement).

Basic SPARQL

AcademicArticle1	Venue1
http://vivo-trunk.indiana.edu/individual/Article869	http://vivo-trunk.indiana.edu/individual/Proceedings639
http://vivo-trunk.indiana.edu/individual/Article579	http://vivo-trunk.indiana.edu/individual/Proceedings265
http://vivo-trunk.indiana.edu/individual/Article673	http://vivo-trunk.indiana.edu/individual/Proceedings692

Ewww… This doesn't look like what we wanted.

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

URIs

VIVO does not store information based on the name, as we would consider it. Instead, it assigns every entity a Uniform Resource Identifier (or URI).

Why? To eliminate ambiguity. There are over 300 people named Smith in the University of Florida VIVO instance, many of whom share first names. In this environment, "Smith, J." is just too vague.

Basic SPARQL

```
PREFIX rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX vivo: <http://vivoweb.org/ontology/core#>
PREFIX bibo: <http://purl.org/ontology/bibo/>

SELECT ?title ?journal
WHERE
{
   ?AcademicArticle1 rdf:type bibo:AcademicArticle .
   ?AcademicArticle1 vivo:hasPublicationVenue ?Venue1 .
   ?AcademicArticle1 rdfs:label ?title .
   ?Venue1 rdfs:label ?journal .
}
LIMIT 3
```

Let's try this again, but let's specify that we want the label associated with the URI. We've added two new conditions to our WHERE clause.

What's the catch?

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Basic SPARQL

title	journal
Evidence for genes on chromosome 2 contributing to alcohol dependence with conduct disorder and suicide attempts	American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics
Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature	Journal of Informetrics
Facility-level factors influencing chronic heart failure care process performance in a national integrated health delivery system	Medical Care

Now we're getting somewhere!

Basic SPARQL

```
PREFIX rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX vivo: <http://vivoweb.org/ontology/core#>
PREFIX bibo: <http://purl.org/ontology/bibo/>

SELECT ?title ?journal ?dt1
WHERE
{
   ?AcademicArticle1 rdf:type bibo:AcademicArticle .
   ?AcademicArticle1 vivo:hasPublicationVenue ?Venue1 .
   ?AcademicArticle1 rdfs:label ?title .
   ?Venue1 rdfs:label ?journal .
   ?AcademicArticle1 vivo:dateTimeValue ?DateTimeValue1 .
   ?DateTimeValue1 vivo:dateTime ?dt1 .
}
LIMIT 3
```

Let's add a level. For Academic Articles, the dateTimeValue is the Publication Date.

Again, remember that if the paper has no publication date listed, it will vanish.

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Basic SPARQL

```
PREFIX rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX vivo: <http://vivoweb.org/ontology/core#>
PREFIX bibo: <http://purl.org/ontology/bibo/>

SELECT ?title ?journal ?dt1
WHERE
{
   ?AcademicArticle1 rdf:type bibo:AcademicArticle .
   ?AcademicArticle1 vivo:hasPublicationVenue ?Venue1 .
   ?AcademicArticle1 rdfs:label ?title .
   ?Venue1 rdfs:label ?journal .
   OPTIONAL {
   ?AcademicArticle1 vivo:dateTimeValue ?DateTimeValue1 .
   ?DateTimeValue1 vivo:dateTime ?dt1 .
   }
}
LIMIT 3
```

Putting clauses into an OPTIONAL structure does about what you'd expect, it makes these clauses optional. If this data is associated, get it, but don't throw out records that don't have it.

Basic SPARQL

title	journal	dt1
Evidence for genes on chromosome 2 contributing to alcohol dependence with conduct disorder and suicide attempts	American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics	2010-01-01T00:00:00
Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature	Journal of Informetrics	
Facility-level factors influencing chronic heart failure care process performance in a national integrated health delivery system	Medical Care	2007-01-01T00:00:00

This is starting to look like a dataset.

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

(Not So) Basic SPARQL

```
PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX bibo: <http://purl.org/ontology/bibo/>
PREFIX vivo: <http://vivoweb.org/ontology/core#>

SELECT STR(?title) STR(?dt1) STR(?journal) (group_concat(STR(?author) ; separator = "|") as
?authorlist)
WHERE
{
   ?AcademicArticle1 rdf:type bibo:AcademicArticle .
   ?AcademicArticle1 rdfs:label ?title .
   ?AcademicArticle1 vivo:hasPublicationVenue ?Venue1 .
   ?Venue1 rdfs:label ?journal .
   ?AcademicArticle1 vivo:informationResourceInAuthorship ?a .
   ?Person1 vivo:authorInAuthorship ?a .
   ?Person1 rdfs:label ?author .
   OPTIONAL {
   ?DateTimeValue1 vivo:dateTime ?dt1 .
   ?AcademicArticle1 vivo:dateTimeValue ?DateTimeValue1 .
   }
}
GROUP BY ?title ?dt1 ?journal
```

(Not So) Basic SPARQL

title	dt1	journal	authorlist
Network structure of cerebral cortex shapes functional connectivity on multiple time scales	2007-01-01T00:00:00	Proceedings of the National Academy of Sciences of the United States of America	Breakspear, M.|Sporns, Olaf|Kotter, R.|Honey, C.J.
Proximity and precedence in arithmetic	2010-01-01T00:00:00	Quarterly Journal of Experimental Psychology	Goldstone, R.L.|Landy, D.
K-core decomposition of internet graphs: Hierarchies, self-similarity and measurement biases	2008-01-01T00:00:00	Networks and Heterogeneous Media	Alvarez-Hamelin, J.I.|Barrat, A.|Dall'asta, L.|Vespignani, Alessandro

The full version of this set, taken from http://vivo-netsci.cns.iu.edu/vivo12/ is the file publication_results.csv in the download packet. We'll be using it later this afternoon.

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

```
PREFIX rdf:  <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:  <http://www.w3.org/2000/01/rdf-schema#>
PREFIX vivo: <http://vivoweb.org/ontology/core#>

SELECT STR(?title) STR(?agency) STR(?start) STR(?end) (group_concat(STR(?pi) ; separator = "|") as
?pilist) STR(?amount)
WHERE
{
   ?grant rdf:type vivo:Grant .
   ?grant vivo:dateTimeInterval ?Interval1 .
   ?Interval1 vivo:start ?startDate .
   ?startDate vivo:dateTime ?start .
   ?Interval1 vivo:end ?endDate .
   ?endDate vivo:dateTime ?end .
   ?grant vivo:totalAwardAmount ?amount .
   ?grant rdfs:label ?title .
   ?grant vivo:grantAwardedBy ?funder .
   ?funder rdfs:label ?agency .
   { ?person vivo:hasPrincipalInvestigatorRole ?piRole .
   ?piRole vivo:roleIn ?grant . }
   UNION
   { ?person vivo:hasCo-PrincipalInvestigatorRole ?coPiRole .
   ?coPiRole vivo:roleIn ?grant . }
   ?person rdfs:label ?pi .

}
GROUP BY ?title ?agency ?amount ?start ?end
LIMIT 5
```

Another example

This grant query is structured fairly similarly to the publication query we just did, with a few differences.

The UNION clause lets you merge two different graphs, in this case we want a list of anyone who is a PI or a co-PI on a grant. We then go on to get the labels for those people and add them to the output.

```
{ ?person vivo:hasPrincipalInvestigatorRole ?piRole .
  ?piRole vivo:roleIn ?grant . }
  UNION
  { ?person vivo:hasCo-PrincipalInvestigatorRole ?coPiRole .
  ?coPiRole vivo:roleIn ?grant . }
```

Sample output is on the next page.

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

title	agency	start	end	pilist	amount
Towards an Institutional Theory of Collective Action	NATIONAL SCIENCE FOUNDATION	1987-07-01T00:00:00	1990-06-30T00:00:00	Ostrom, Elinor	200,921
Cross-Situational Statistical Word Learning: Behaviors, Mechanisms, and Constraints	NATIONAL INSTITUTE OF CHILD HEALTH, HUMAN DEVL.	2009-07-01T00:00:00	2011-06-30T00:00:00	SMITH, LINDA B.	370,033
The Decentralized Reforms and Property Rights: Potentials and Puzzles for Forest Sustainability and Livelihoods	VIRGINIA POLYTECHNIC INSTITUTE & STATE UNIVERSITY	2006-01-01T00:00:00	2010-03-31T00:00:00	Ostrom, Elinor	595,335
Genomic Studies in Bipolar Affective Disorder	NATIONAL INSTITUTE OF MENTAL HEALTH	1995-04-01T00:00:00	1999-03-31T00:00:00	Nurnberger, John	1,275,514
Doctoral Dissertation Research in Political Science: The Scramble for Property Rights: Renegotiating Livelihoods and Sustainability after Uganda's Forest Governance Reform""	NATIONAL SCIENCE FOUNDATION	2006-09-01T00:00:00	2007-08-31T00:00:00	Ostrom, Elinor	11,760

The full version of this set is also in the download packet, grants_results.csv. We'll be using it here in just a few minutes.

Data Cleaning

Generally, the data you get from a query needs a little work before it's ready for use in your favorite visualization tool.

Record your cleaning steps as you do them, in case you need to rebuild your dataset for some reason.

You will not do the above, at least not until you've been burned a few times.

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

title	agency	start	end	pilist	amount
Towards an Institutional Theory of Collective Action	NATIONAL SCIENCE FOUNDATION	1987-07-01T00:00:00	1990-06-30T00:00:00	Ostrom, Elinor	200,921
Cross-Situational Statistical Word Learning: Behaviors, Mechanisms, and Constraints	NATIONAL INSTITUTE OF CHILD HEALTH, HUMAN DEVL.	2009-07-01T00:00:00	2011-06-30T00:00:00	SMITH, LINDA B.	370,033
The Decentralized Reforms and Property Rights: Potentials and Puzzles for Forest Sustainability and Livelihoods	VIRGINIA POLYTECHNIC INSTITUTE & STATE UNIVERSITY	2006-01-01T00:00:00	2010-03-31T00:00:00	Ostrom, Elinor	595,335
Genomic Studies in Bipolar Affective Disorder	NATIONAL INSTITUTE OF MENTAL HEALTH	1995-04-01T00:00:00	1999-03-31T00:00:00	Nurnberger, John	1,275,514
Doctoral Dissertation Research in Political Science: The Scramble for Property Rights: Renegotiating Livelihoods and Sustainability after Uganda's Forest Governance Reform""	NATIONAL SCIENCE FOUNDATION	2006-09-01T00:00:00	2007-08-31T00:00:00	Ostrom, Elinor	11,760

For the visualization we're about to do, I want the ranges for the grants to be normal dates.

Data Cleaning

title	agency	start	end	pilist	amount	start_date	end_date
Towards an Institutional Theory of Collective Action	NATIONAL SCIENCE FOUNDATION	1987-07-01T00:00:00	1990-06-30T00:00:00	Ostrom, Elinor	200,921	07/01/1987	06/30/1990
Cross-Situational Statistical Word Learning: Behaviors, Mechanisms, and Constraints	NATIONAL INSTITUTE OF CHILD HEALTH, HUMAN DEVL.	2009-07-01T00:00:00	2011-06-30T00:00:00	SMITH, LINDA B.	370,033	07/01/2009	06/30/2011
The Decentralized Reforms and Property Rights: Potentials and Puzzles for Forest Sustainability and Livelihoods	VIRGINIA POLYTECHNIC INSTITUTE & STATE UNIVERSITY	2006-01-01T00:00:00	2010-03-31T00:00:00	Ostrom, Elinor	595,335	01/01/2006	03/31/2010
Genomic Studies in Bipolar Affective Disorder	NATIONAL INSTITUTE OF MENTAL HEALTH	1995-04-01T00:00:00	1999-03-31T00:00:00	Nurnberger, John	1,275,514	04/01/1995	03/31/1999
Doctoral Dissertation Research in Political Science: The Scramble for Property Rights: Renegotiating Livelihoods and Sustainability after Uganda's Forest Governance Reform""	NATIONAL SCIENCE FOUNDATION	2006-09-01T00:00:00	2007-08-31T00:00:00	Ostrom, Elinor	11,760	09/01/2006	08/31/2007

This can be achieved fairly easily in Excel, but be careful that the csv doesn't get garbled by commas in the titles or missing data. This can lead to unwanted offsets in the data.

Also, note that derived columns are added to the set, rather than replacing raw data.

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Tools for Data Cleaning

Excel is generally the one that most people are familiar with, but it is rather limited in the ability to replicate work easily.

R is easily scripted to make any task quickly replicable and is extendable to cover most tasks, but the early learning curve can be steep.

Most any spreadsheet or statistical software can work, provided you can see and manipulate your data. Straight programming is also an option.

Visualization Tools

These are numerous and diverse. Once again, Excel is probably the most commonly used, but as before, it is rather limited.

We will be using the Science of Science (Sci2) Tool, a **free** software set designed by the programming team at the CNS Center.

Register at http://sci2.cns.iu.edu for access.

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Introduction to Sci2

The Science of Science (Sci2) Tool is a free and open-source modular toolset originally designed for the study of science. However it has many uses that support temporal, geospatial, topical, and network analysis and visualization of scholarly datasets.

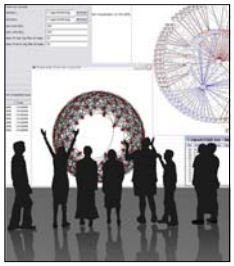
Macroscopes

Decision making in science, industry, and politics, as well as in daily life, requires that we make sense of the massive amounts of data that result from complex systems.

Rather than making things larger or smaller, **macroscopes let us observe what is too great, slow, or complex for us to comprehend or sometimes even notice.**

Microscopes

Telescopes



Macroscopes

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Plug-and-Play Macroscopes

While microscopes and telescopes are physical instruments, macroscopes are **continuously changing bundles of software plugins**

Macroscopes make it easy to

- Simply drop plugins into the tool and they appear in the menu, ready to use
- Sharing algorithm components, tools, or novel interfaces becomes as easy as sharing images on Flickr or videos on YouTube

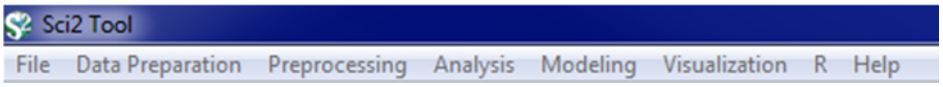
Sci2 Tool – Supported Data Formats

Input:
- Network Formats
- GraphML (*.xml or *.graphml)
- XGMML (*.xml)
- Pajek .NET (*.net)
- NWB (*.nwb)
- Scientometric Formats
- ISI (*.isi)
- Bibtex (*.bib)
- Endnote Export Format (*.enw)
- Scopus csv (*.scopus)
- NSF csv (*.nsf)
- Other Formats
- Pajek Matrix (*.mat)
- TreeML (*.xml)
- Edgelist (*.edge)
- CSV (*.csv)

Output:
- Network File Formats
- GraphML (*.xml or *.graphml)
- Pajek .MAT (*.mat)
- Pajek .NET (*.net)
- NWB (*.nwb)
- XGMML (*.xml)
- CSV (*.csv)
- JPEG (*.jpg)
- PDF (*.pdf)
- PostScript (*.ps)

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

A Quick Tour of Sci2



The basic menu sets out the general structure of a workflow.

File – Load data.
Data Preparation – Extract the needed elements from your data.
Preprocessing – Do any needed manipulations.
Analysis – Perform algorithms or mapping to your data.
Modeling – network modeling (rarely used)
Visualization - Create your visualization!

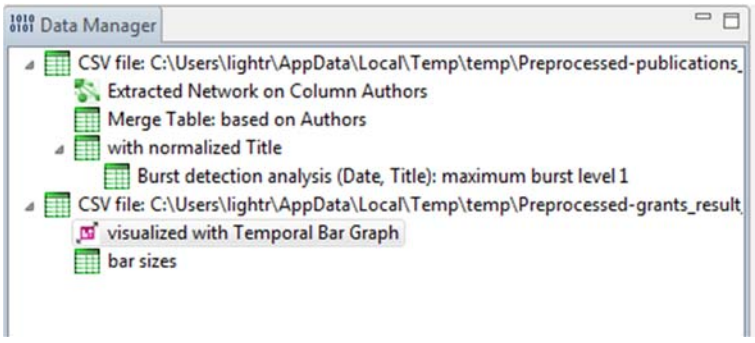
This first example is very simple, so we'll be skipping a lot of these steps.

A Quick Tour of Sci2

The console is where most communication from the system takes place. This is where you will see warnings and alerts, as well as certain statistics.

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

A Quick Tour of Sci2



The newly opened file appears in the Data Manager. As you perform work on a file, those steps create derived files in a tree structure.

From here you can choose which file you are working on, as well as discard files that are no longer of use.

A Quick Tour of Sci2

The Scheduler shows what has been done, when it was done and what is still outstanding. Long tasks will either show a percent completed, or a rolling green bar to show that they are still working.

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Documentation and Support

Sci2 is documented via a Wiki maintained at http://sci2.wiki.cns.iu.edu

When in trouble, check the console. It will often offer a link to the relevant Wiki page.

"But that didn't help at all!"

The Sci2 page does offer "Ask An Expert" for problems and questions not covered by the Wiki.

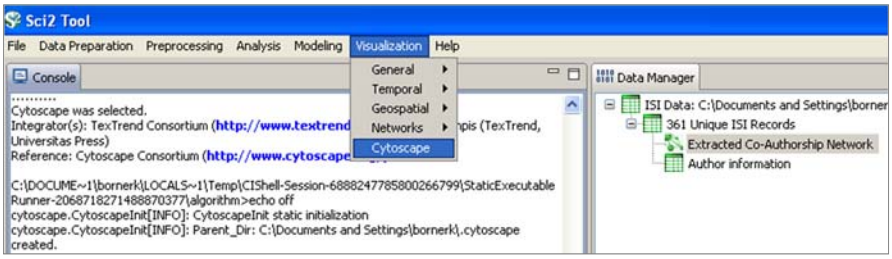
Supported Tools

Adding more layout algorithms and network visualization interactivity via Cytoscape http://www.cytoscape.org.

Simply add *org.textrend.visualization.cytoscape_0.0.3.jar* into your /plugin directory.

Restart Sci² Tool

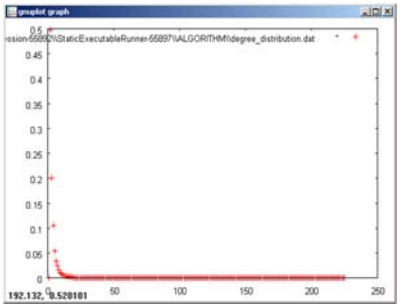
Cytoscape now shows in the Visualization Menu

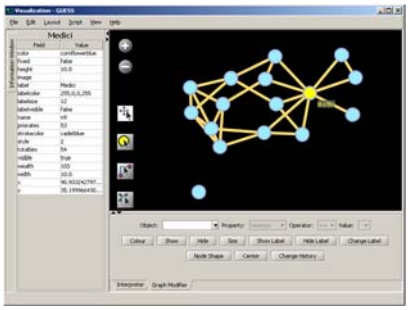


Select a network in Data Manager, run Cytoscape and the tool will start with this network loaded.

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Supported Tools





Gnuplot

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

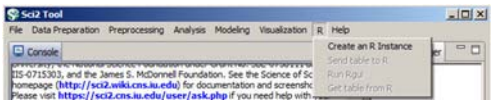
GUESS

exploratory data analysis and visualization tool for graphs and networks.

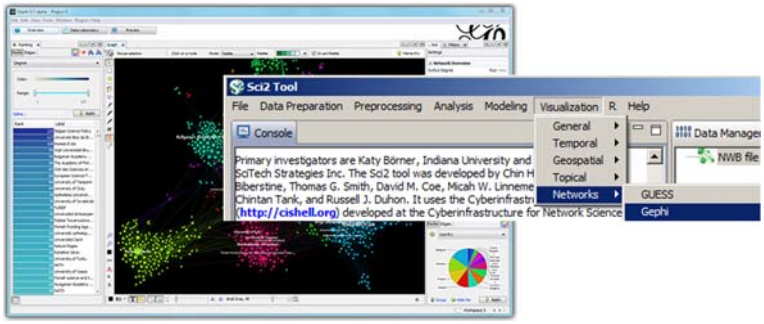
http://nwb.cns.iu.edu/?n=VisualizeData.GUESS.

Bridged Tools

R statistical tool
http://www.r-project.org/



Gephi visualization tool
https://gephi.org/



CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Temporal Bar Graph

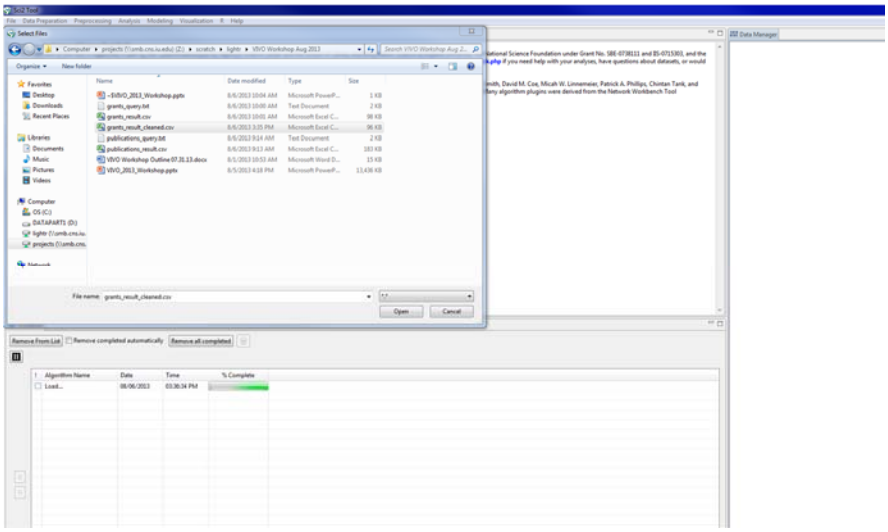
Question: How does our funding look over time in terms of number of grants as well as amount of money?

Visualization: A bar graph over time, with one bar per grant and bar area sized to the amount of money.

Data: Grant titles, date ranges (in years) and amount awarded.

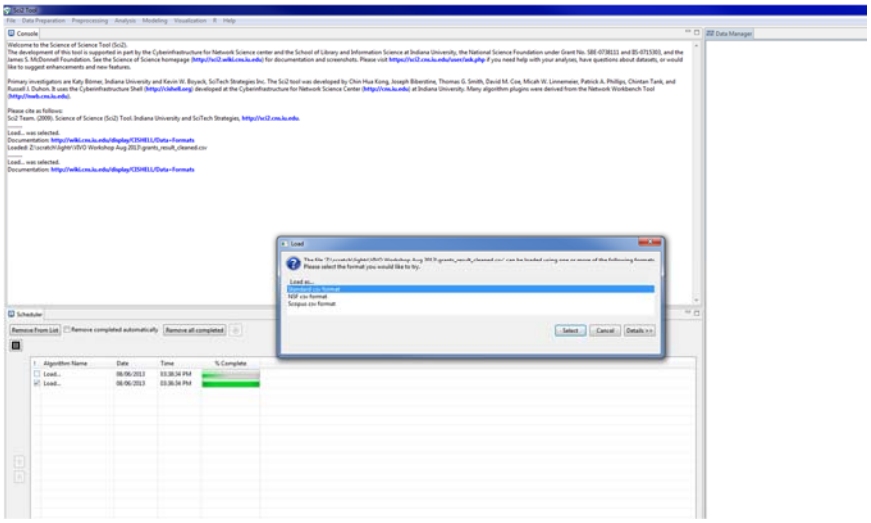
We have all the required data in grant_results_cleaned.csv in the download packet.

Temporal Bar Graph

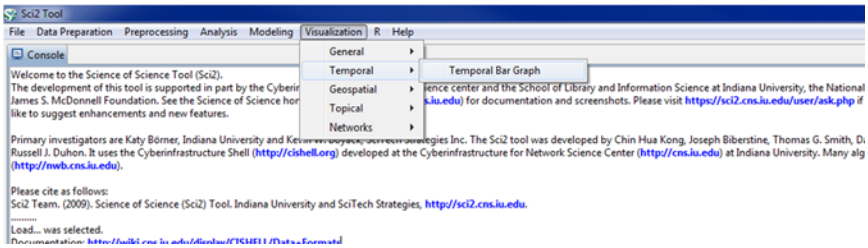


CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Temporal Bar Graph



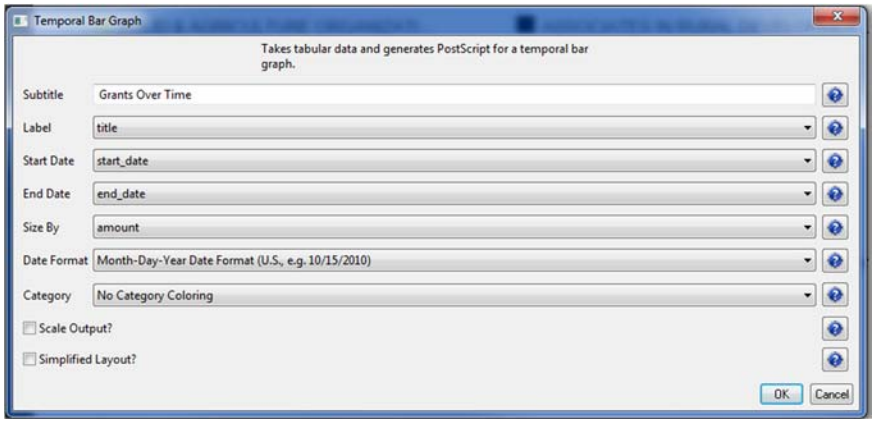
Temporal Bar Graph



As noted earlier, there isn't a lot of preparation or preprocessing needed here, so we'll go straight to Visualization.

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Temporal Bar Graph



We get a screen asking for the details of our visualization. Note that each entry has a ? next to it to offer help.

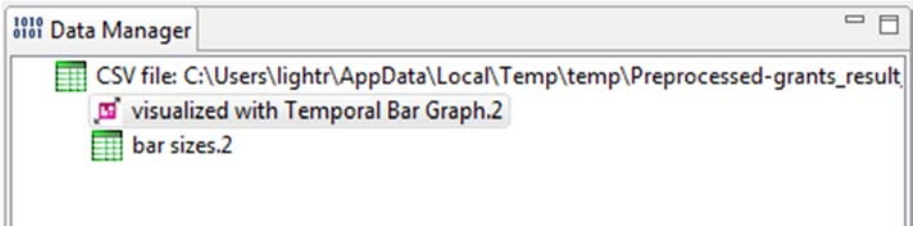
Temporal Bar Graph

Subtitle – This will show on the visualization, explaining what you've done.
Label – What defines your bars
Start Date – Self-explanatory
End Date – Self-explanatory
Size By – What should influence the area of your bars?
Date Format – How are your dates structured? This is why we had to clean up the VIVO dates.
Category – Should we color our bars by category?*

*You can only get 10 different colors and we have a lot more agencies than that, so we won't bother with this.

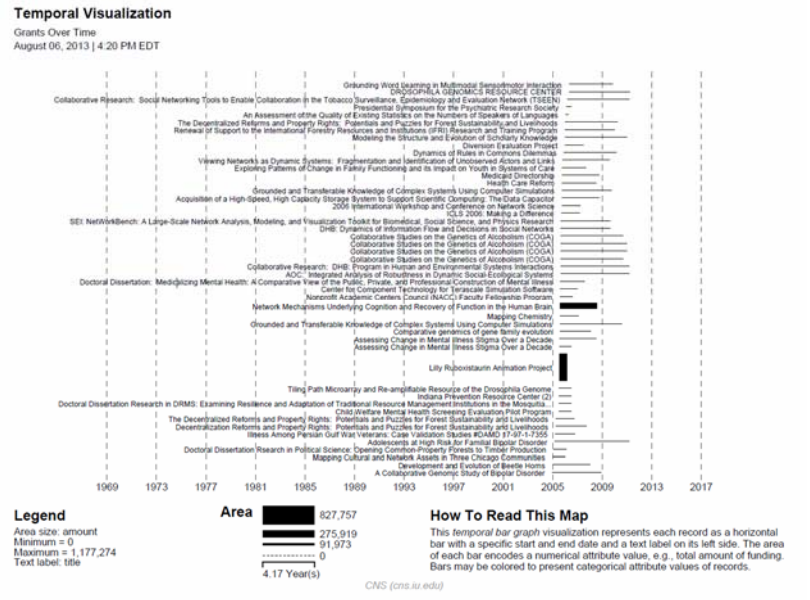
CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Temporal Bar Graph



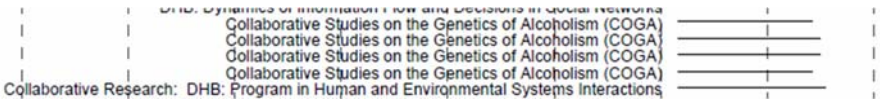
In most cases, visualizations are returned as PostScript (.ps) files. To convert these to PDF:

Windows – Adobe or GhostScript & GSView
Mac OS – Preview
Linux – ps2pdf comes on most distributions
Anyone: Visit http://ps2pdf.com for an online tool.



Verification

WAIT! Just because you have your picture doesn't mean you're done yet. Look it over, does it answer your question? Do the results make sense? Do you see any gaps or oddities in the data?



Here is the core data behind this, see the problem?
http://vivo-netsci.cns.iu.edu/vivo12/display/Grant78086

Verification

In this case, it looks like there is some duplicated data. This may be because a grant was renewed or updated or data may have inadvertently been entered twice. You'd have to go back and do some research to verify what this data really means. If it is in error, you may have to make corrections, then repeat the process of data acquisition, cleaning and visualization. If it's not, you may have to modify your queries to accommodate it.

In the end, your final output can only be as good as the data that went into it.

Questions?

Intermission

Fifteen minute break

Please let us know if you have any difficulty running Sci2

Check out these links for more information on CNS activities!
http://cns.iu.edu
http://scimaps.org
http://sci2.cns.iu.edu

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Presentation Overview

Part One

- Introduction
- VIVO Visualization Review
- VIVO Data Storage and Retrieval
  - Ontology
  - Basic SPARQL
  - Data Cleaning
- Introduction to Sci2
- Temporal Bar Graph
- Break

Part Two

- Analyzing and visualizing VIVO data
  - Burst Detection
  - Map of Science
  - Network
- The next level of aggregated data
  - iNRN (http://nrn.cns.iu.edu)
  - VIVOSearch.org
- Continuing Education and Opportunities

What is Burst Detection?

A burst is something suddenly appearing or become popular. For example, when iPhone 5 released in September 2012, tweets related to iPhone and Apple became much more frequent for several months, then slowly returned to normal levels. Running a burst detection algorithm over Twitter data in this timeframe will show bursts on words such as iPhone, Apple, iOS, iTunes, etc.

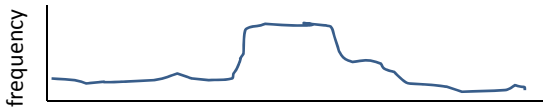
Why use burst detection? To find trends and emergent topics, to study events happening over the time, find out the main stream topics, and to further predict future trends as they emerge.

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

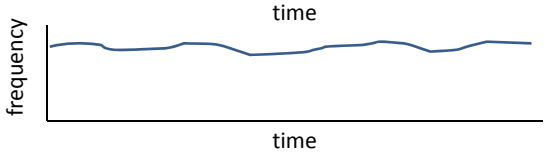
What ISN'T Burst Detection?

Burst detection is NOT designed to simply pick up popular words. It is designed to detect a CHANGE in the frequency of a word's occurrence in the data.

This is a burst:



This is not:



Even though the second graph shows a word with more frequent usage, there is no change in frequency, so no burst would be shown.

Burst Detection

Question: What are the research topics that have emerged at our university? What was popular that may have recently faded? What has become more relevant recently?

Visualization: Temporal Bar Graph of Burst Detection.

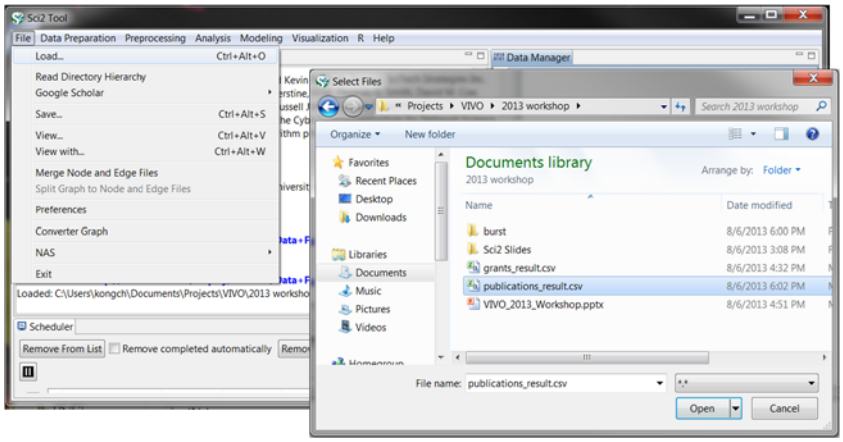
Data: Publication titles, and Year.

We have all the required data in publication_result.csv in the download packet.

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

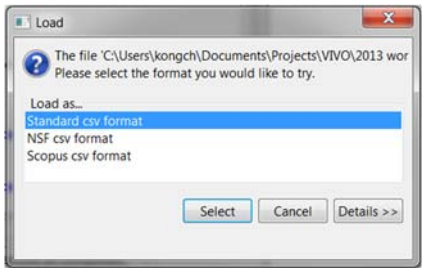
Burst Detection

Load publication_result.csv file into Sci2



Burst Detection

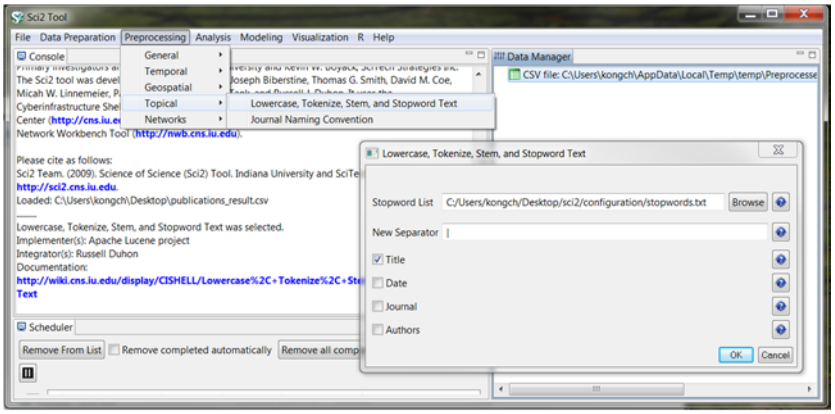
Load the CSV file as **Standard csv format**. For other supported data formats in Sci2, see http://sci2.wiki.cns.iu.edu/display/SCI2TUTORIAL/2.3+Data+Formats



CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Burst Detection

Normalizing publication title *by using* **Lowercase, Tokenize, Stem, Stopword Text** plugin.



Burst Detection

Lowercase: *United States -> united states*

Tokenize: Split text content into words with delimiter. E.g. *The|VIVO|visualization|workshop*

Stem: Normalizes past and present tense and common suffixes. Example: *managed, manage, manages, and managing* are stemmed into *manag*

Stopword: Defined in stopword.txt. Example: *the, is, are, from, while…* You can build your own file as well, for special cases.

CNS Cyberinfrastructure for Network Science Center VIVO Data and Visualizations

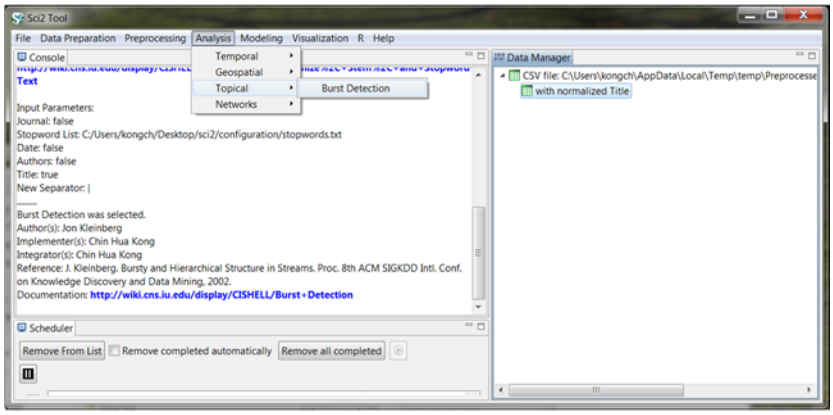
Burst Detection

	A
1	Title
2	network|structur|cerebr|cortex|shape|function|connect|multipl|time|s
3	proxim|preced|arithmet
4	k|core|decomposit|internet|graph|hierarchi|self|similar|measur|bias
5	speci|lifetim|distribut|simpl|model|ecolog
6	xcat|scienc|portal
7	epidem|predict|meta|popul|model|heterogen|coupl|impact|diseas|par
8	strateg|network|restor
9	role|geographi|traffic|structur|complex|network
10	horn|possess|reduc|maneuver|horn|polyphen|beetl|onthophagus|nigri
11	est|microarray|analysi|horn|develop|onthophagus|beetl
12	comput|simul|multi|pit|corros|process|materi
13	virtual|world|petri|dish|rat|maze|supercollid

Sample result

Burst Detection

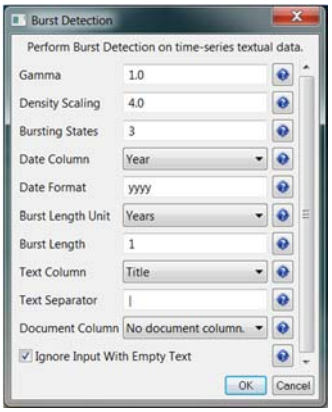
Invoking Burst Detection algorithm



CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Burst Detection

For parameter details, please refer to
http://sci2.wiki.cns.iu.edu/display/CISHELL/Burst+Detection



Burst Detection

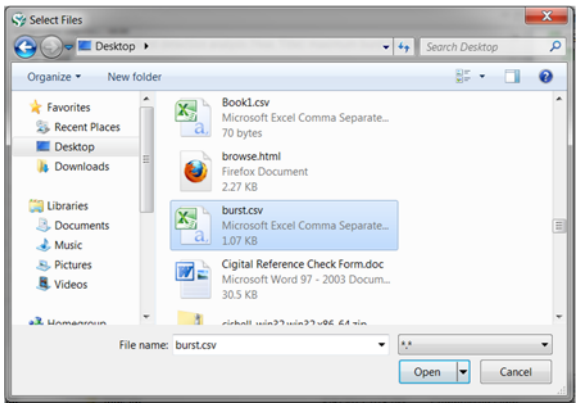
Open the burst detection result table, set the empty End fields to the latest year (2011) in the dataset

Word	Level	Weight	Length	Start	End
manag	1	3.554311	1	2002	2002
rpa	1	4.079587	1	2000	2000
rpa	2	7.911986	1	2000	2000
action	1	3.430809	7	1998	2004
finger	1	3.954979	3	1999	2001
finger	2	7.274332	3	1999	2001
polycentr	1	3.681505	2	2010	
redox	1	3.954979	3	1999	2001
redox	2	7.274332	3	1999	2001
effect	1	3.465786	1	2009	2009
end	1	4.817891	2	2003	2004
web	1	3.319353	1	2004	2004
inform	1	4.417534	3	2002	2004
glycat	1	4.817891	2	2003	2004
regul	1	3.60641	3	1999	2001
intellig	1	4.172818	3	2002	2004
assist	1	3.841699	1	2003	2003
distribut	1	3.866117	1	2005	2005

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

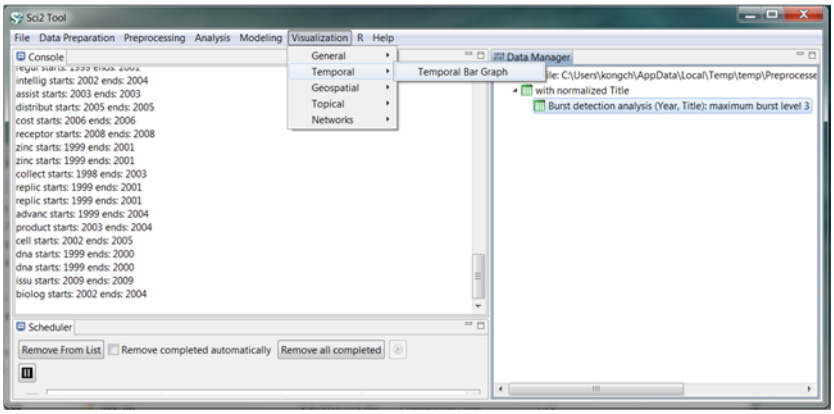
Burst Detection

Save the edited table to desktop as burst.csv and load it into Sci2.



Burst Detection

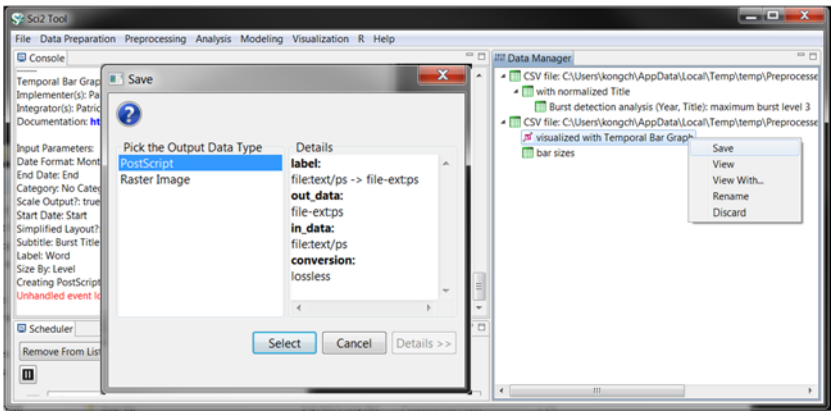
Visualize with Temporal Bar Graph



CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

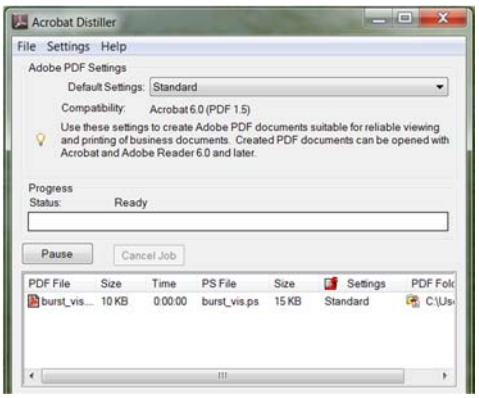
Burst Detection

Save the visualization result as a PostScript file

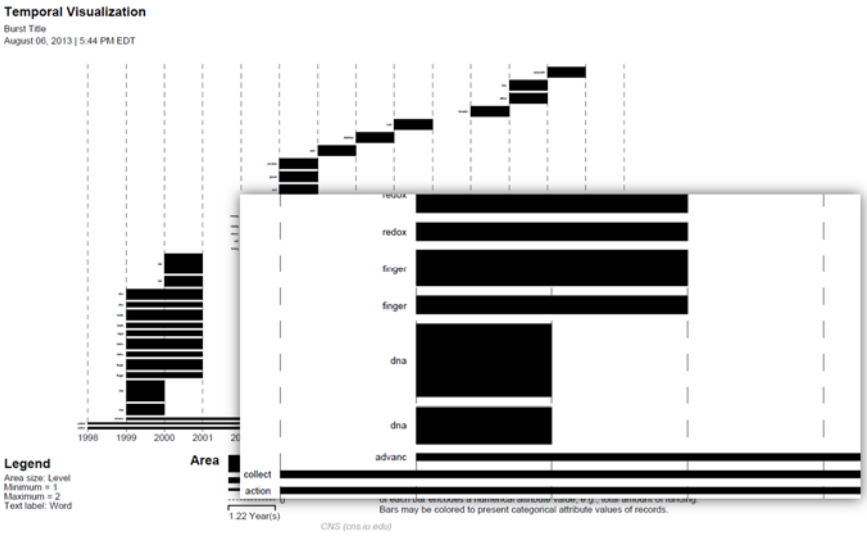


Burst Detection

Convert the saved PostScript file into PDF using Acrobat Distiller or PostScript viewers



Burst Detection



Questions?

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Map of Science

Question: Are our publications playing to the advertised strengths of our research group? What are our real strengths? Where do we need to focus our efforts to improve our presence?

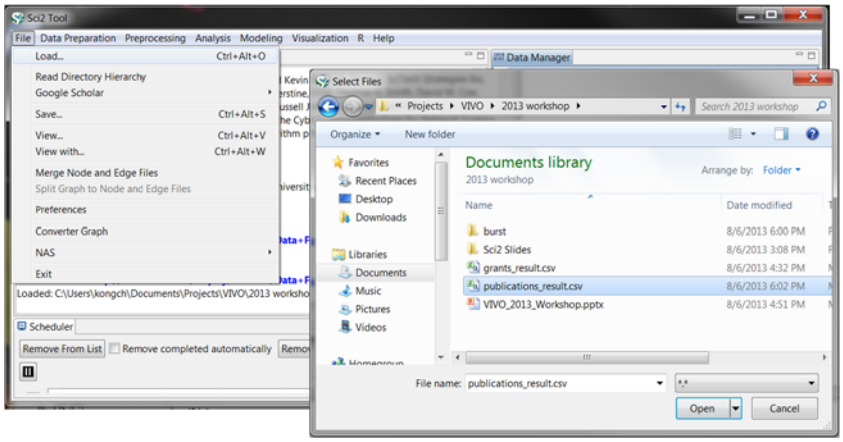
Visualization: A Map of Science that maps research efforts into science domains using published journal distribution.

Data: Journal names for each publication.

We have all the required data in publication_result.csv in the download packet.

Map of Science

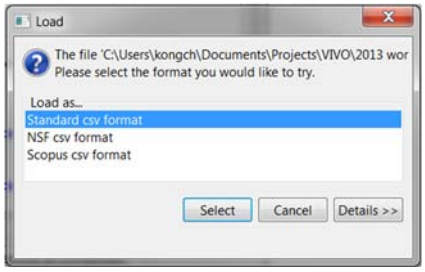
Load publication_result.csv file into Sci2



CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

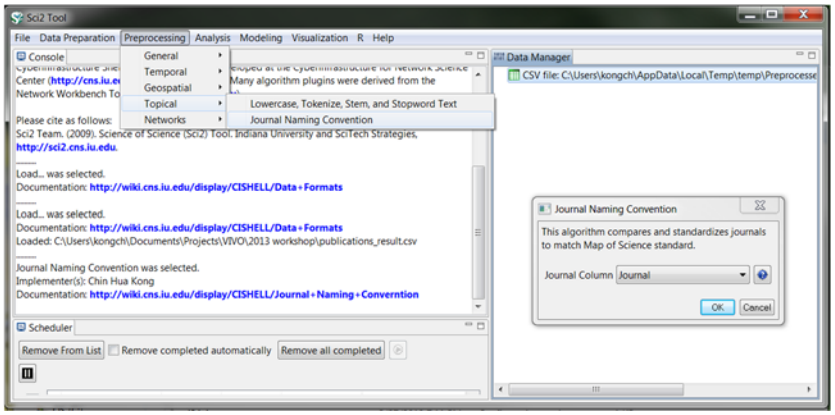
Map of Science

Load the CSV file as **Standard csv format**. For other supported data formats in Sci2, see http://sci2.wiki.cns.iu.edu/display/SCI2TUTORIAL/2.3+Data+Formats



Map of Science

Clean up journals name to Map of Science standard *by using* the **Journal Naming Convention** plugin.



CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Map of Science

Journal names are often written and abbreviated in multiple ways.

New England Journal of Medicine
The New England Journal of Medicine
New England Journal of Medicine, The
NEJM: New England Journal of Medicine
NEJM – New England Journal of Medicine

The Journal Naming Convention plugin tries to normalize these names and link them to lookup tables within Sci2 to create the best matching possible.

Map of Science

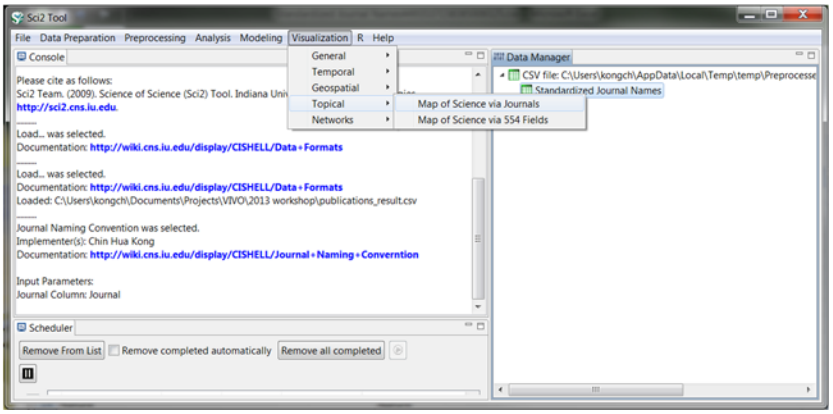
Journal Name	Standardized Journal Name
Proceedings of the National Academy of Sciences of the United S	Proceedings Of The National Academy Of Sciences Of T
Quarterly Journal of Experimental Psychology	QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY
Networks and Heterogeneous Media	NETWORKS AND HETEROGENEOUS MEDIA
Proceedings of the National Academy of Sciences of the United S	Proceedings Of The National Academy Of Sciences Of T
Scientific Programming	Scientific Programming
International Journal of Bifurcation and Chaos	International Journal Of Bifurcation And Chaos
Networks and Spatial Economics	NETWORKS & SPATIAL ECONOMICS
Advances in Complex Systems	Advances In Complex Systems
Journal of insect science (Online)	
BMC Genomics	Bmc Genomics
Computational Materials Science	Computational Materials Science
Games and Culture	Games and Culture
Nature	Nature
Journal of the Learning Sciences	Journal Of The Learning Sciences
Journal of Affective Disorders	Journal Of Affective Disorders
Physical Review E - Statistical, Nonlinear, and Soft Matter Physics	Physical Review E - Statistical, Nonlinear, And Soft Matt
International Journal of Medical Informatics	International Journal Of Medical Informatics
Breast Cancer Research and Treatment	Breast Cancer Research And Treatment
Alcoholism: Clinical and Experimental Research	Alcoholism, Clinical And Experimental Research

Sample result

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

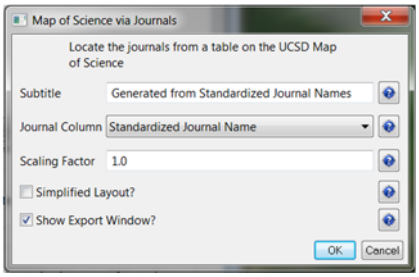
Map of Science

Visualize the result with **Map of Science via Journals** plugin.

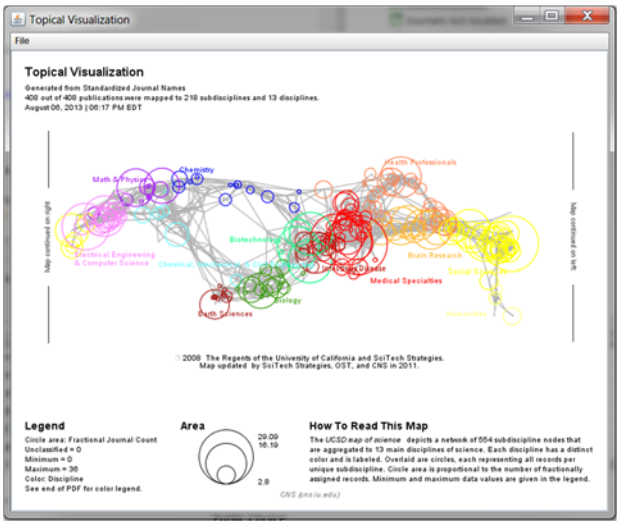


Map of Science

For parameters details, please refer to
http://sci2.wiki.cns.iu.edu/display/CISHELL/Map+of+Science+via+Journals

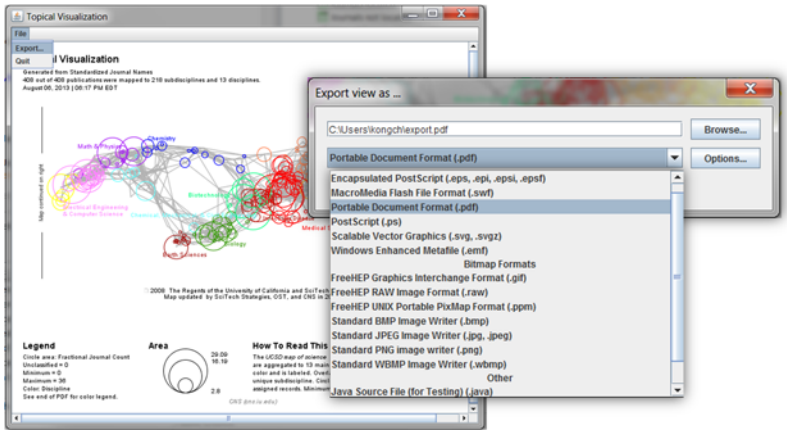


Map of Science



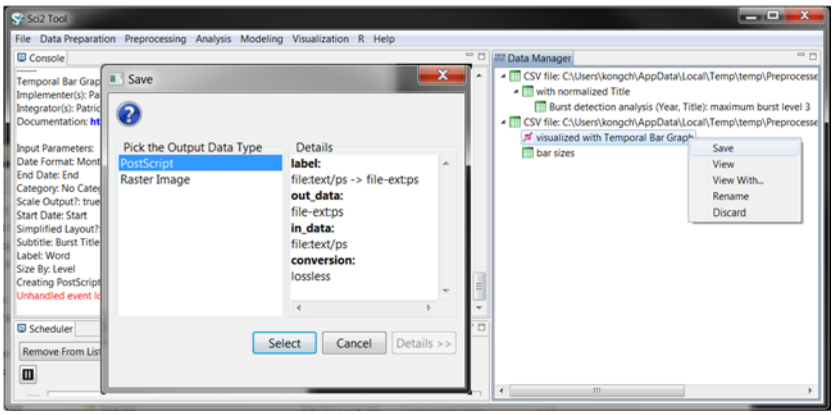
Map of Science

Save the visualization result into different formats



Map of Science

Save the visualization result as a PostScript file

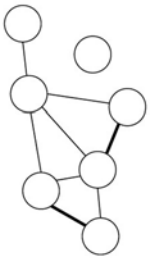


Questions?

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Introduction to Networks

Undirected Networks



Nodes:

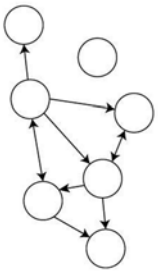
Edges:

Node Degree:
Number of edges connected to nodes

Isolates:
Nodes that are not connected to the rest of the network

Edge Weight:
Demonstrates relative importance of relationships

Directed Networks



Edge Direction:
Directional relationship is represented by arrows

In-Degree:
Number of incoming edges

Out-Degree:
Number of outgoing edges

Network

Question: How well are our paper authors collaborating? Is the entire department working together or are there sub-groups that only interact with each other? Who are the centers of the collaboration network?

Visualization: Network Graph using GUESS and Gephi.

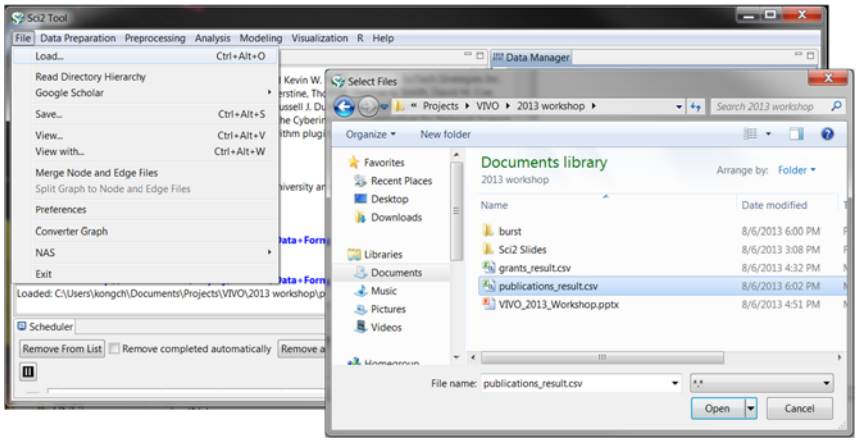
Data: Publications and authors.

We have all the required data in publication_result.csv in the download packet.

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

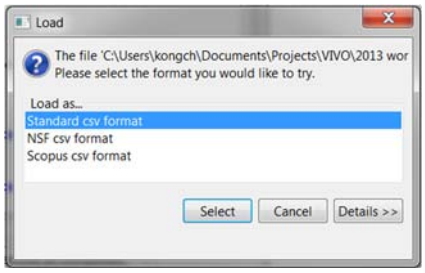
Network

Load publication_result.csv file into Sci2



Network

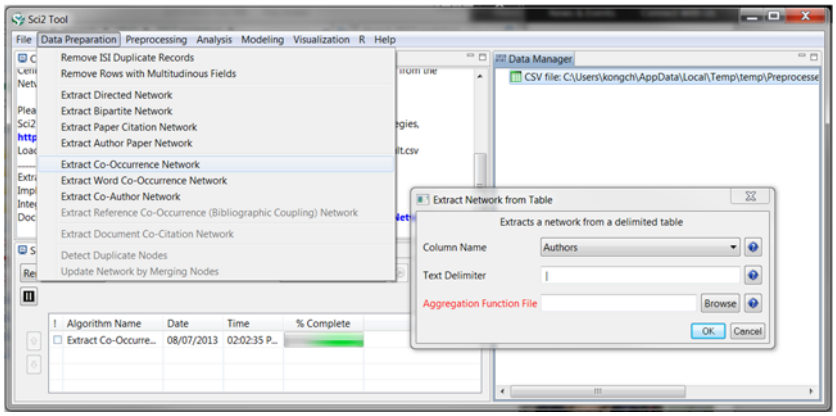
Load the CSV file as **Standard csv format**. For other supported data formats in Sci2, see http://sci2.wiki.cns.iu.edu/display/SCI2TUTORIAL/2.3+Data+Formats



CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

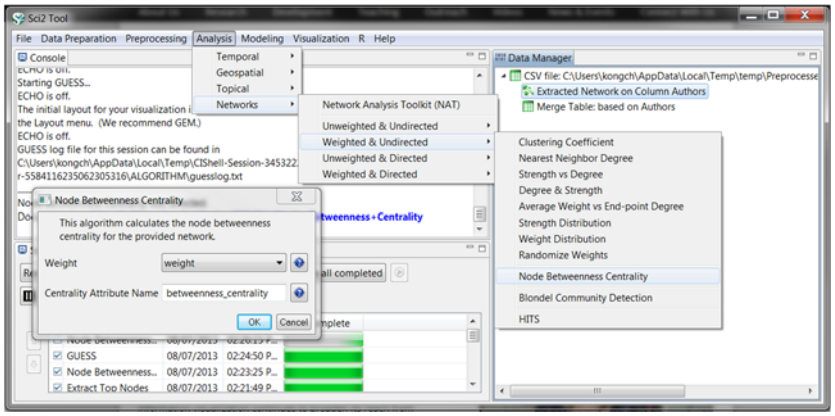
Network

Extract the Co-Author network using **Extract Co-Occurrence Network** plugin.



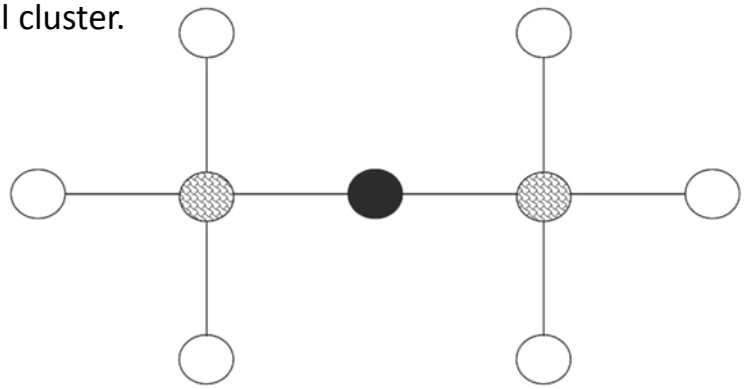
Network

Generate **Node Betweenness Centrality** that represent the connectivity scores.



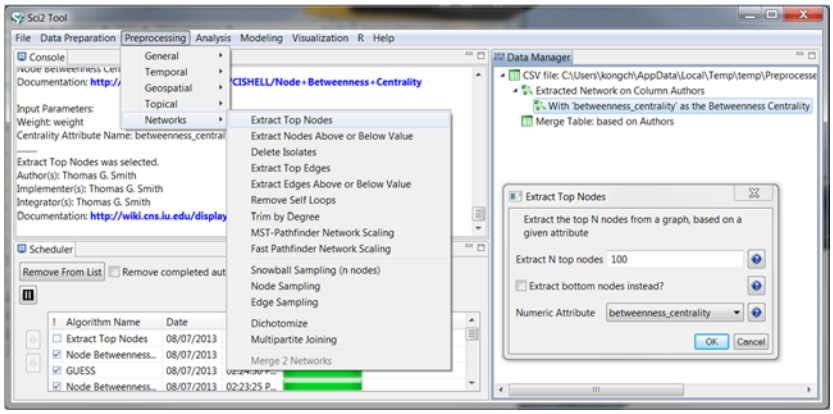
Network

The darkest node has the highest betweenness centrality score. It is the central connection of the two small cluster.



Network

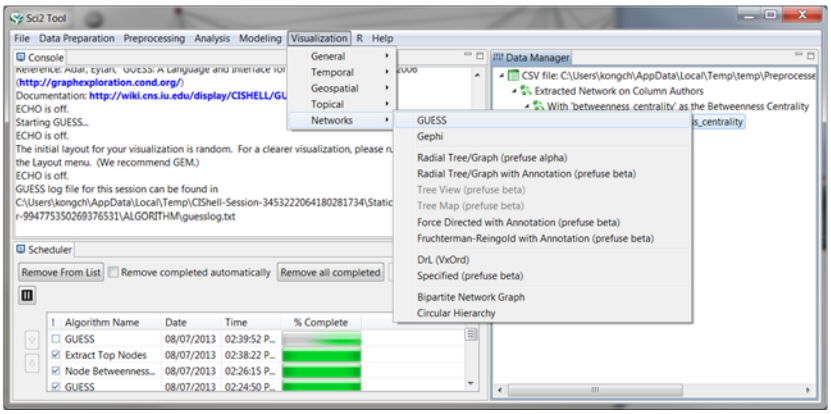
Extract Top 100 Nodes using **Extract Top Nodes** plugin.
We would like to focus on the top connected authors



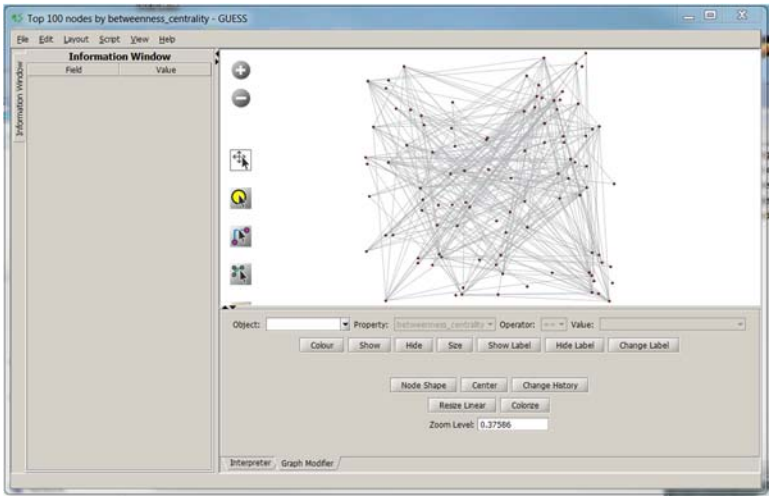
CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Network

Visualize the network with GUESS



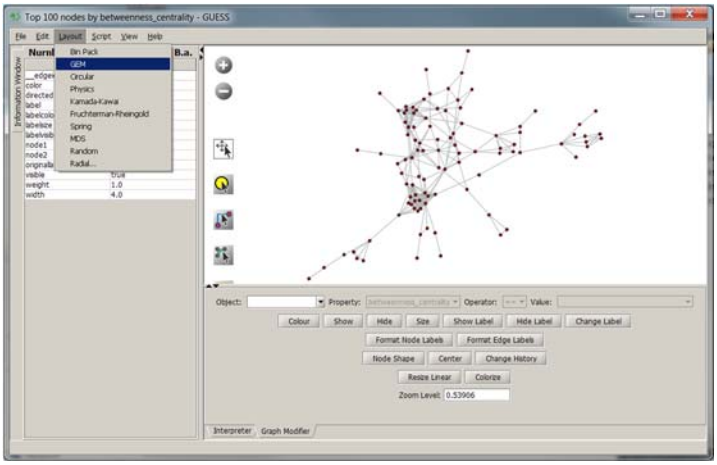
Network



CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

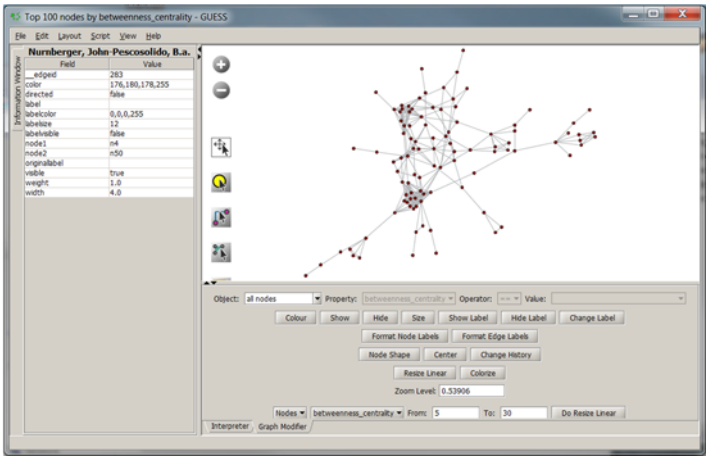
Network

Apply GEM layout



Network

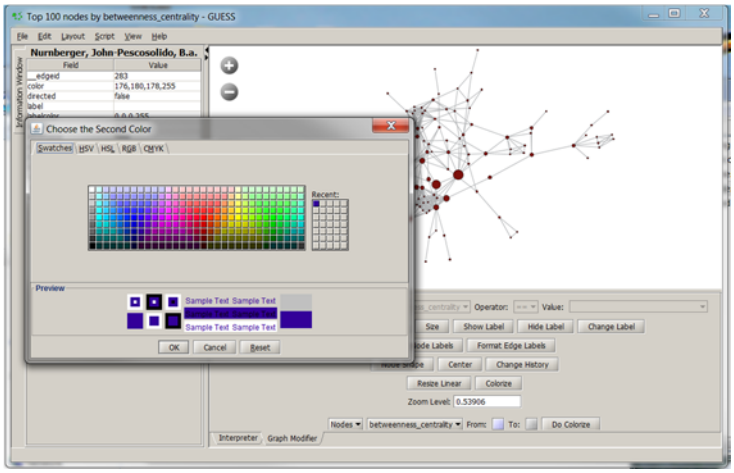
Select **Resize Linear** button to resize the nodes



CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

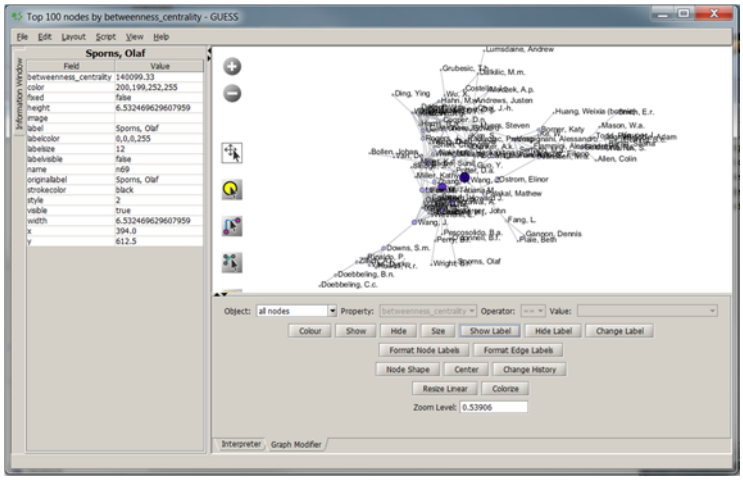
Network

Select **Colorize** button to color the nodes



Network

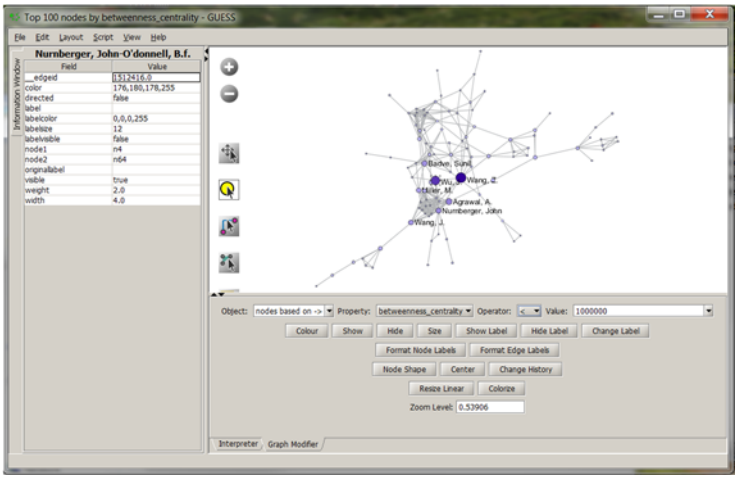
Select **Show Label**.



CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Network

Use filtering to hide labels of smaller nodes



Guess vs Gephi

Guess

- Easy to use with the Graph Modifier
- Powerful Python scripting with the Interpreter
- Doesn't require additional software

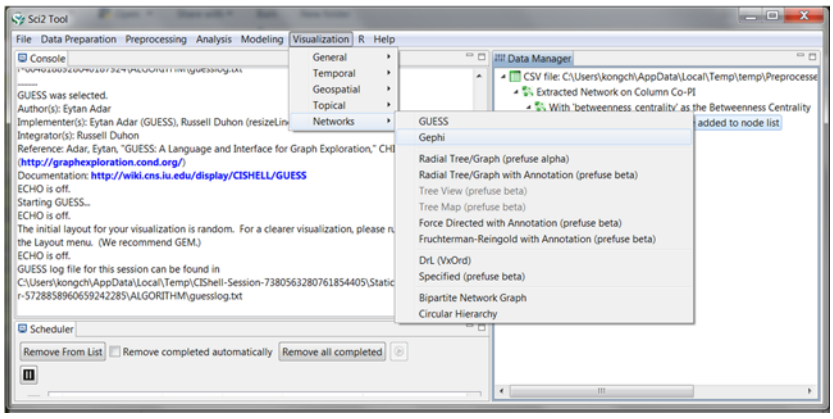
Gephi

- Works efficiently with large graph
- Supports data editing through table view
- Supports export to SVG, PDF, PNG
- Supports advance graphic functionality

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

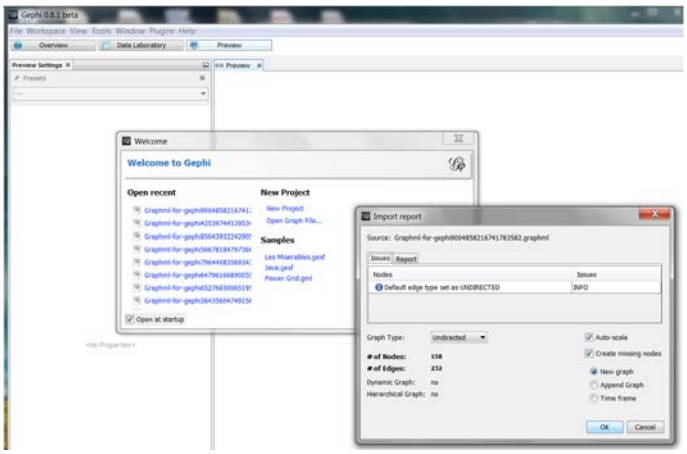
Network

Visualize the network with Gephi



Network

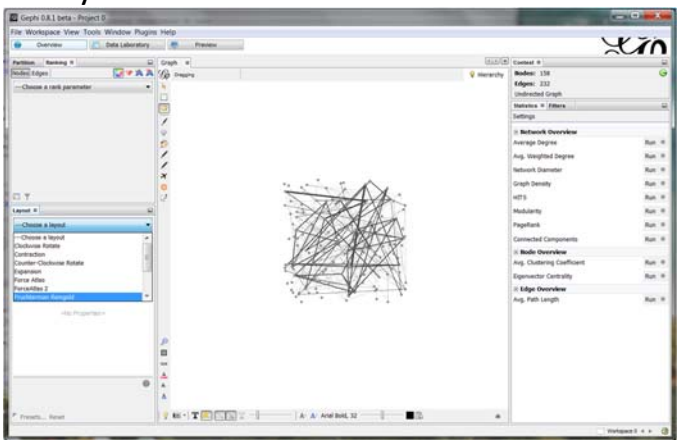
Select **OK** to load the graph into Gephi



CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

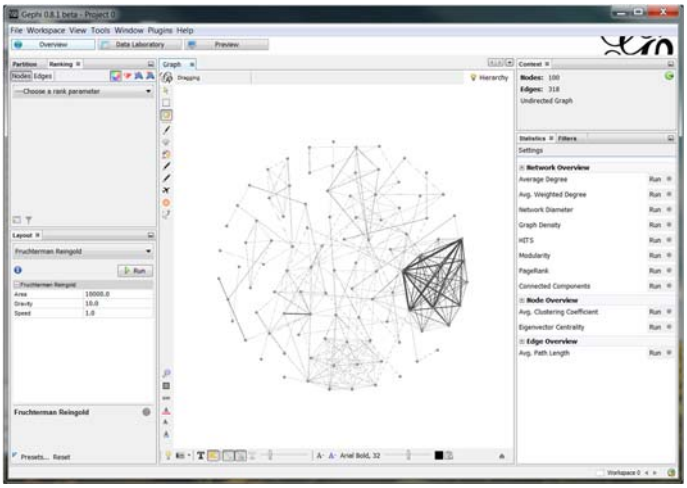
Network

Switch to Overview tab and layout with Fruchterman Reingold layout



Network

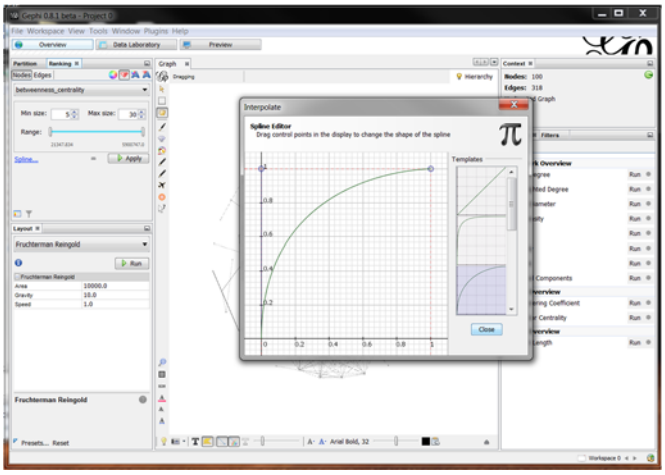
Run and stop the layout as desired



CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

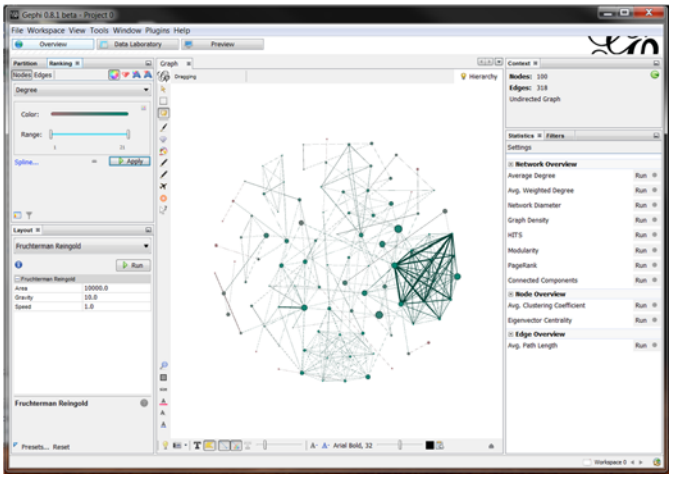
Network

Size code the nodes by betweenness / Degree



Network

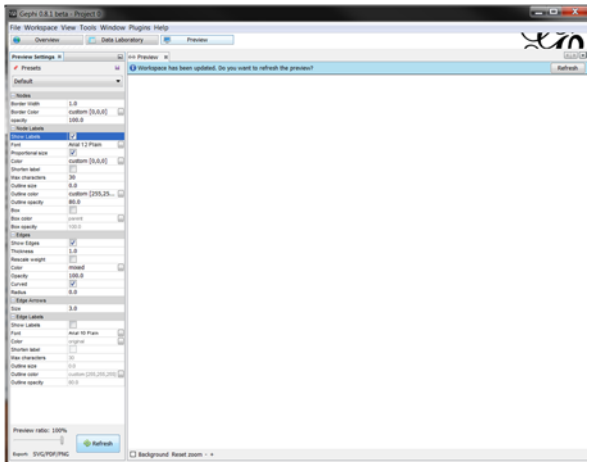
Colorize the nodes by Degree



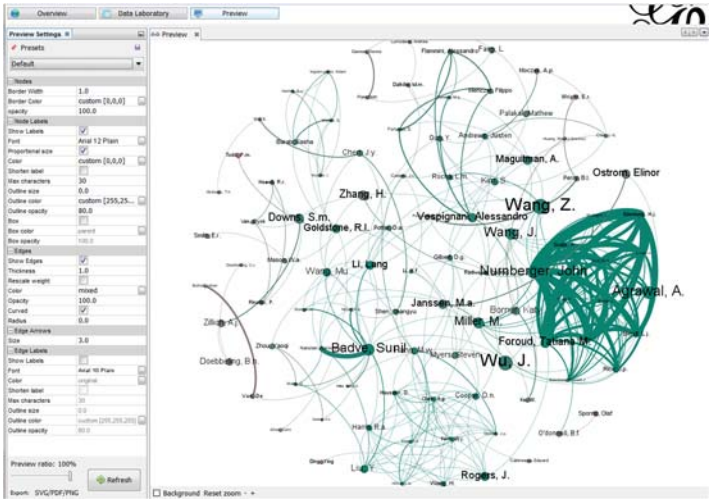
CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Network

Switch to **Preview**. Select **Show Labels** and **Refresh**



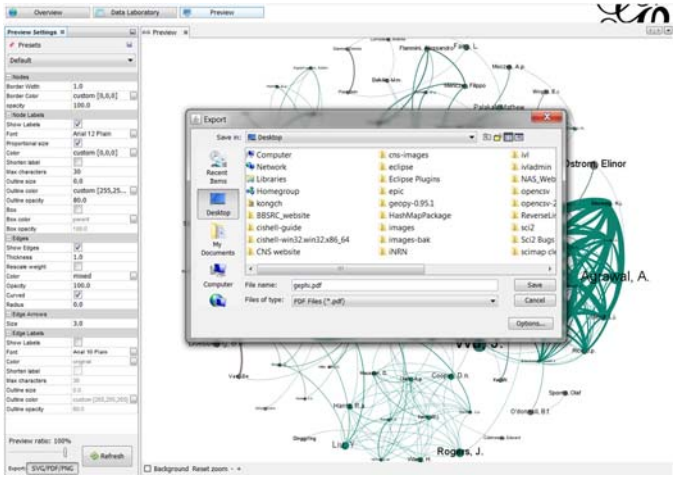
Network



Visualization result

Network

Export to SVG, PDF, or PNG



Network

Try to visualize the following networks using the grant_result.csv data:

- Co-PI Network using PI column
  http://wiki.cns.iu.edu/pages/viewpage.action?pageId=2785284
- Co-occurrence word network using title column
  http://wiki.cns.iu.edu/pages/viewpage.action?pageId=2200066#id-514StudyingFourMajorNetSciResearchersISIData-5145WordCo-OccurrenceNetwork5145WordCo-OccurrenceNetwork
- Work on bipartite network such as PI to Agency network, Grant to PI network, etc.
  http://wiki.cns.iu.edu/pages/viewpage.action?pageId=2785293

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Questions?

A new level of
Data Aggregation
Empowered by VIVO

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

http://nrn.cns.iu.edu

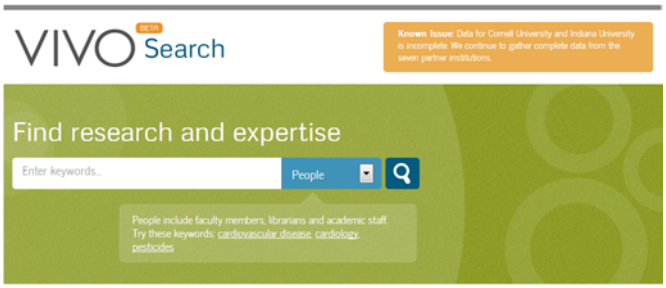
iNRN

A centralized place for sharing and accessing networks, data, knowledge, and expertise. Further explores the evolution of VIVO.

http://vivosearch.org

Vivosearch.org

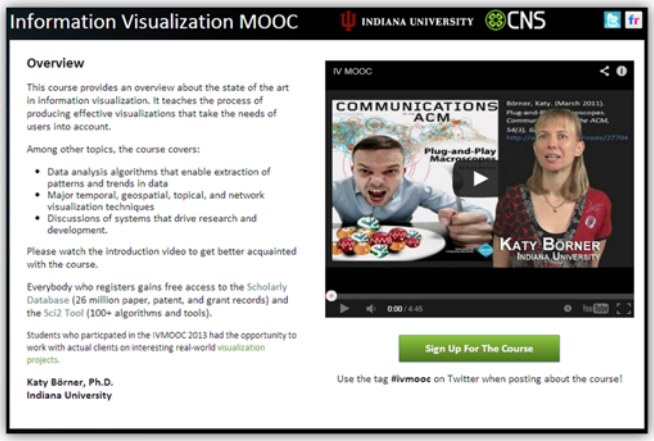
A universal search service for public VIVO instances.
Find resources across multiple instances



http://ivmooc.cns.iu.edu

Learn More

Learning new visualization techniques through IVMOOC,
a **free** online visualization course!



http://cns.iu.edu/products.html

Develop More

Looking for visualization solution, and development collaboration?

Visit us at http://cns.iu.edu/products.html

CNS Cyberinfrastructure for Network Science Center
VIVO Data and Visualizations

Thank You