

An Introduction to Open Source Tools for Data Analysis and Information Visualization

CDC, Atlanta, GA

June 14 & 15, 2016

8:30-4:30 PM EST

Michael Ginda

Sr. Data Analyst

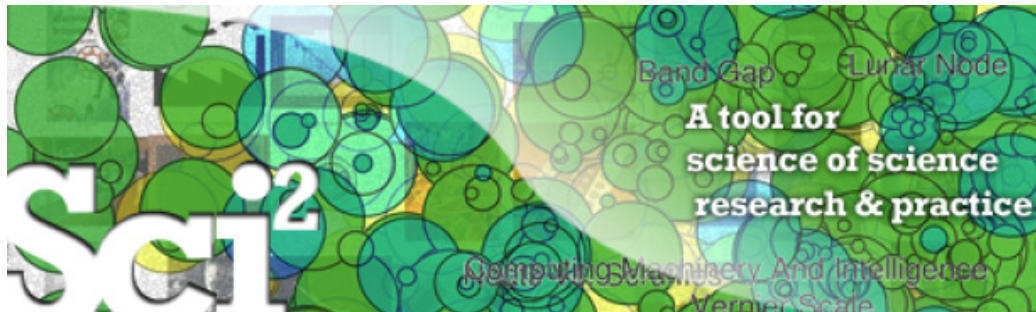
Cyberinfrastructure for Network Science Center

Indiana University,

Bloomington, Indiana, USA

mginda@Indiana.edu

<http://cns.iu.edu>



Refine ^{OPEN} 

 **Gephi**
makes graphs **handy**

8:00 Welcome and Overview of Tutorial and Attendees**8:30 Visualization Framework and Workflow Design**

- Overview of the Visualization framework
- Overview of Graphical variables, and color selection

9:15 Tool Overview and Trouble Shooting

- Sci2 Overview – scientometric analysis tool
- Open Refine – data parsing, transformation, and editing tool
- Gephi – network visualization

9:45 Geospatial Analysis

- Overview of geospatial analysis and mapping
- Geospatial Analysis: Geocoding with OpenRefine
- Geospatial Analysis: Proportional Symbol Map using CDC

11:00 Topical/Temporal Analysis: Burst Detection

- Overview of burst analysis and introductory workflow
- Burst Detection with CDC Grants

12:30 Lunch**1:30 Network Analysis**

- Introduction to Network Analysis
- ~~Network Analysis: Bimodal networks with Morbidity Data~~
- Network Analysis: Co-authorship Network with CDC publications

4:00 Wrap-up

4:30 Adjourn

- Data for the workshop
<http://cns.iu.edu/docs/data/Sci2-CDC-data.zip>
- These slides
- <http://cns.iu.edu/docs/presentation/2016-ginda-sci2tutorial-cdc.pdf>
OR
<http://cns.iu.edu/docs/presentation/2016-ginda-sci2tutorial-cdc-part1.pdf>
<http://cns.iu.edu/docs/presentation/2016-ginda-sci2tutorial-cdc-part1.pptx>
<http://cns.iu.edu/docs/presentation/2016-ginda-sci2tutorial-cdc-part2.pdf>
<http://cns.iu.edu/docs/presentation/2016-ginda-sci2tutorial-cdc-part2.pptx>
- Sci2 Tool Manual v0.5.1 Alpha, updated to match v1.0 Alpha tool release
<http://sci2.wiki.cns.iu.edu>
- Sci2 Tool v1.0 Alpha (June 13, 2012)
<http://sci2.cns.iu.edu>
- Additional Datasets
<http://sci2.wiki.cns.iu.edu/2.5+Sample+Datasets>
- Additional Plugins
<http://sci2.wiki.cns.iu.edu/3.2+Additional+Plugins>
- Some visualizations are saved as Postscript files. A free Postscript to PDF

8:00 Welcome and Overview of Tutorial and Attendees**8:30 Visualization Framework and Workflow Design**

- Overview of the Visualization framework
- Overview of Graphical variables, and color selection

9:15 Tool Overview and Trouble Shooting

- Sci2 Overview – scientometric analysis tool
- Open Refine – data parsing, transformation, and editing tool
- Gephi – network visualization

9:45 Geospatial Analysis

- Overview of geospatial analysis and mapping
- Geospatial Analysis: Geocoding with OpenRefine
- Geospatial Analysis: Proportional Symbol Map using CDC

11:00 Topical/Temporal Analysis: Burst Detection

- Overview of burst analysis and introductory workflow
- Burst Detection with CDC Grants

12:30 Lunch**1:30 Network Analysis**

- Introduction to Network Analysis
- ~~Network Analysis: Bimodal networks with Morbidity Data~~
- Network Analysis: Co-authorship Network with CDC publications

4:00 Wrap-up

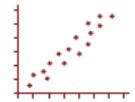

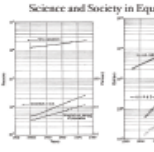
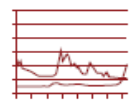


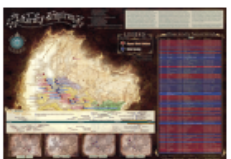




4:30 Adjourn

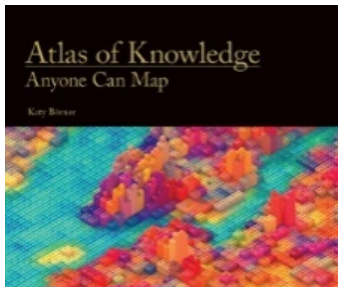
Basic Task Types

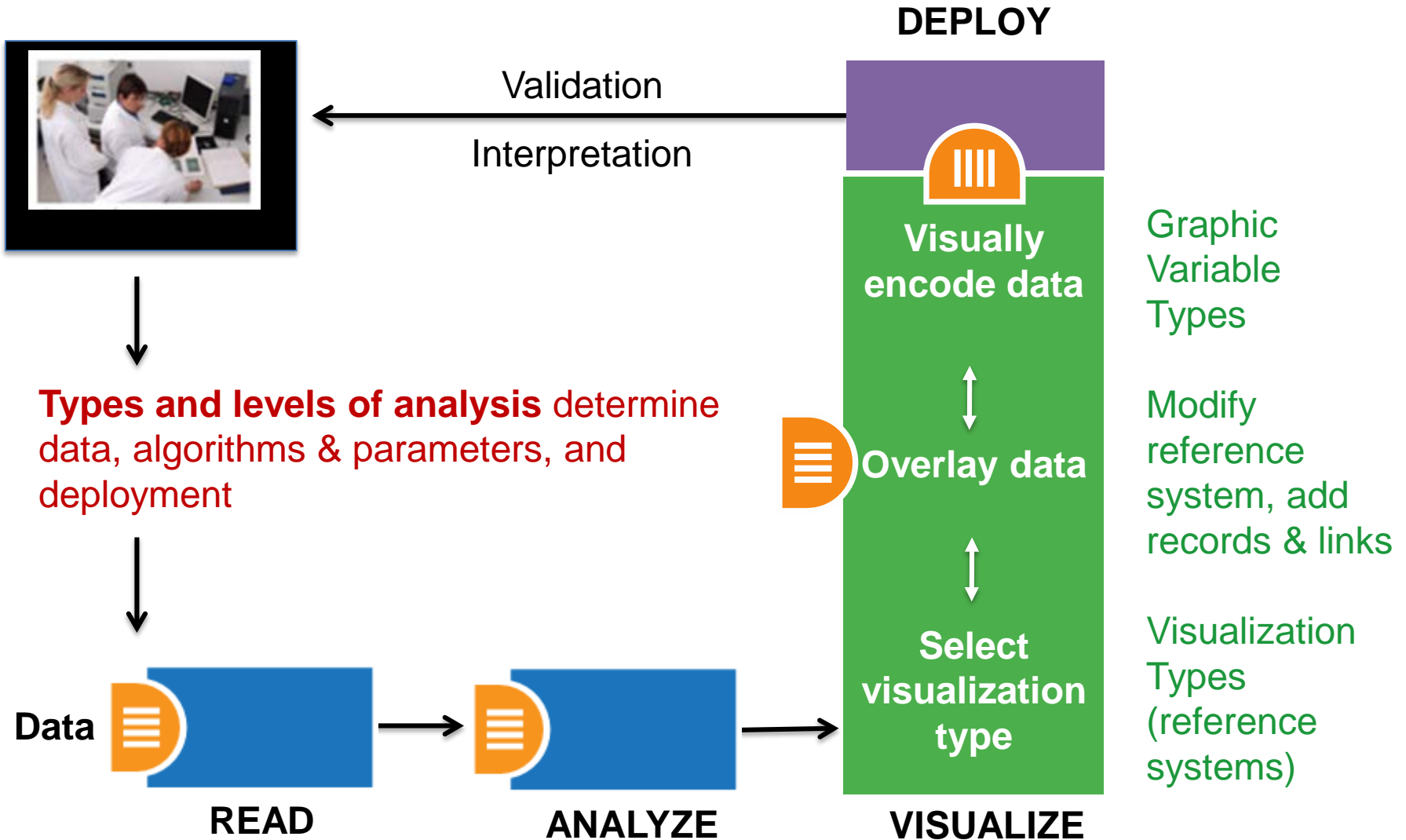
Bertin, 1967	Wehrend & Lewis, 1996	Few, 2004	Yau, 2011	Rendgen & Wiedemann, 2012	Frankel, 2012	Tool: Many Eyes	Tool: Chart Chooser	Börner, 2014
selection	categorize			category				categorize/ cluster
order	rank	ranking					table	order/rank/ sort
	distribution	distribution					distribution	distributions (also outliers, gaps)
	compare	nominal comparison & deviation	differences		compare and contrast	compare data values	comparison	comparisons
		time series	patterns over time	time	process and time	track rises and falls over time	trend	trends (process and time)
		geospatial	spatial relations	location		generate maps		geospatial
quantity		part-to- whole	proportions		form and structure	see parts of whole, analyze text	composition	compositions (also of text)
association	correlate	correlation	relationships	hierarchy		relations between data points	relationship	correlations/ relationships

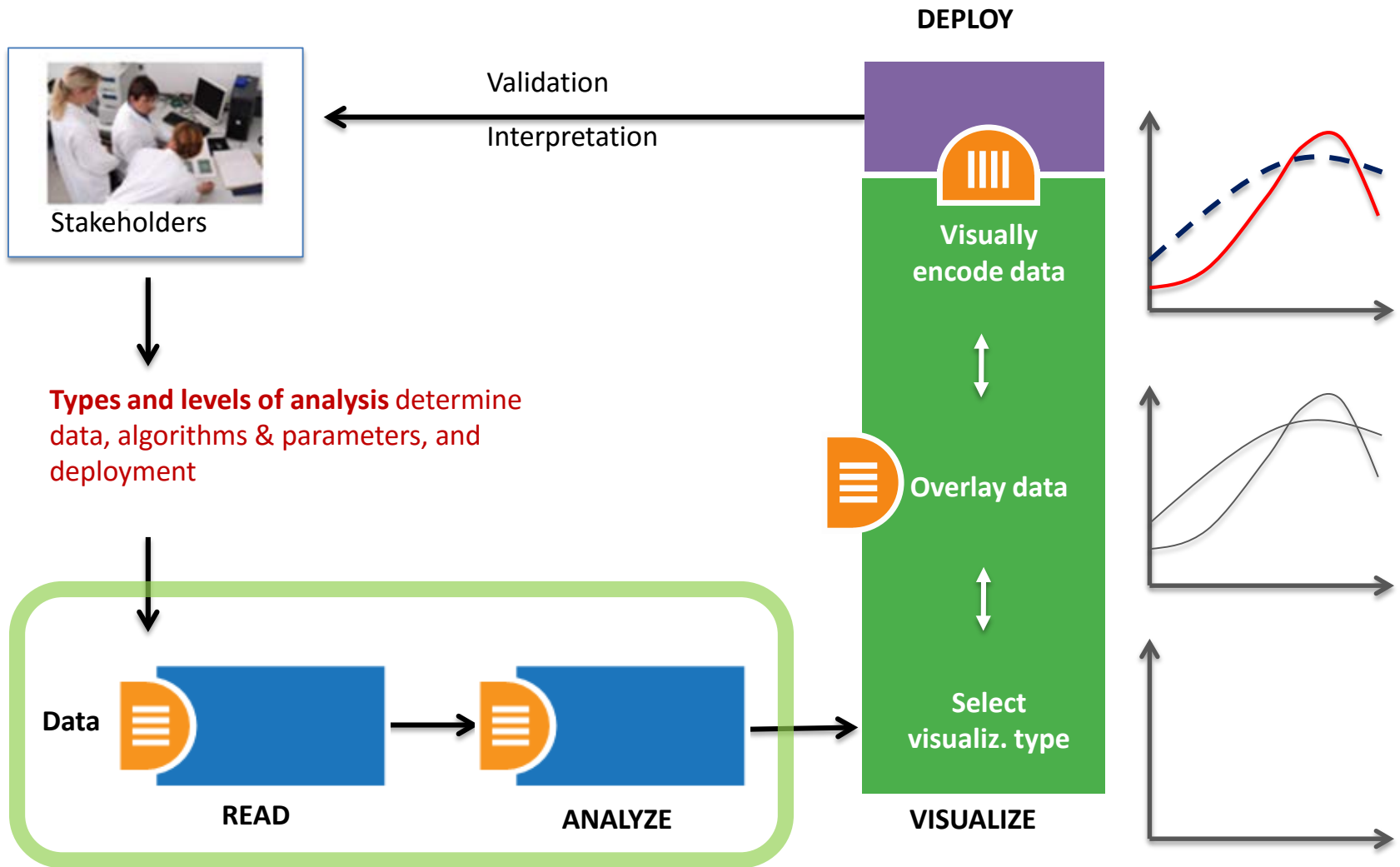
LEVELS

TYPES

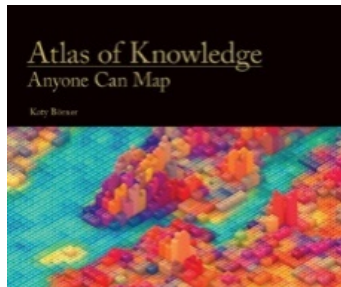
	MICRO: Individual Level about 1–1,000 records page 6	MESO: Local Level about 1,001–100,000 records page 8	MACRO: Global Level more than 100,000 records page 10
			
Statistical Analysis page 44	 Knowledge Cartography page 135	 Productivity of Russian life sciences research teams page 105	 Science and Society in Equilibrium Number of scientists versus population and R&D costs versus GNP. page 103
WHEN: Temporal Analysis page 48	 Visualizing decision-making processes page 95	 Key events in the development of the video tape recorder page 85	 Increased travel and communication speeds page 83
WHERE: Geospatial Analysis page 52	 Cell phone usage in Milan, Italy page 109	 Victorian poetry in Europe page 137	 Ecological footprint of countries page 99
WHAT: Topical Analysis page 56	 Evolving patent holdings of Apple Computer, Inc. and Jerome Lemelson page 89	 Evolving journal networks in nanotechnology page 139	 Product space showing co-export patterns of countries page 93
WITH WHOM: Network Analysis page 60	 World Finance Corporation network page 87	 Electronic and new media art networks page 133	 Map of Scientific Collaborators from 2005–2006 World-wide scholarly collaboration networks page 157



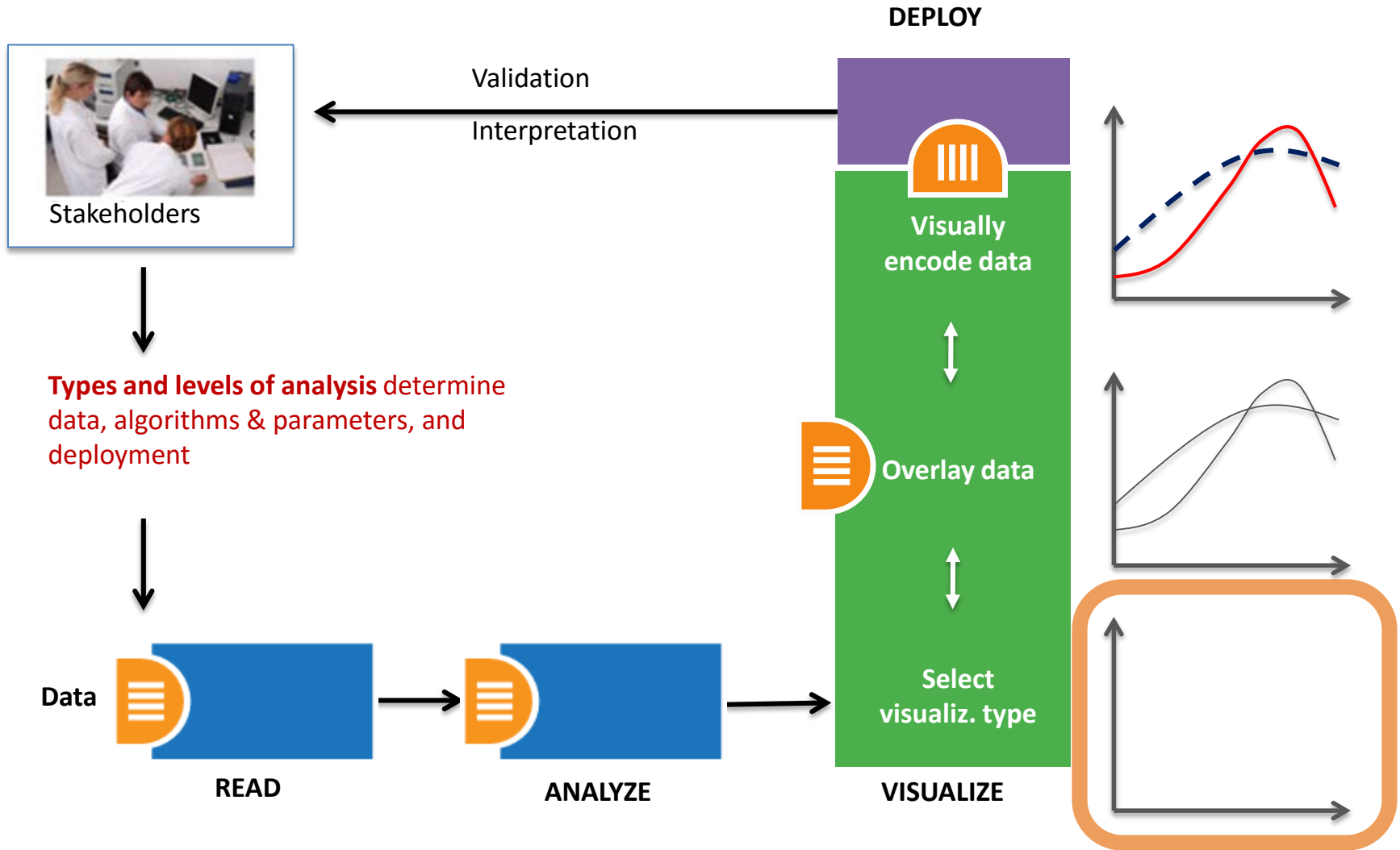




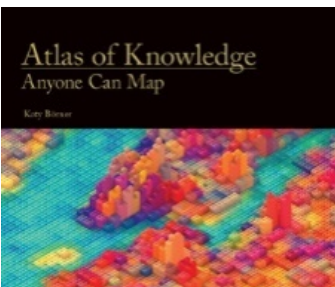
Insight Need Types page 26	Data Scale Types page 28	Visualization Types page 30	Graphic Symbol Types page 32	Graphic Variable Types page 34	Interaction Types page 26
<ul style="list-style-type: none"> • categorize/cluster • order/rank/sort • distributions (also outliers, gaps) • comparisons • trends (process and time) • geospatial • compositions (also of text) • correlations/relationships 	<ul style="list-style-type: none"> • nominal • ordinal • interval • ratio 	<ul style="list-style-type: none"> • table • chart • graph • map • network layout 	<ul style="list-style-type: none"> • geometric symbols <ul style="list-style-type: none"> point line area surface volume • linguistic symbols <ul style="list-style-type: none"> text numerals punctuation marks • pictorial symbols <ul style="list-style-type: none"> images icons statistical glyphs 	<ul style="list-style-type: none"> • spatial <ul style="list-style-type: none"> position • retinal <ul style="list-style-type: none"> form color optics motion 	<ul style="list-style-type: none"> • overview • zoom • search and locate • filter • details-on-demand • history • extract • link and brush • projection • distortion



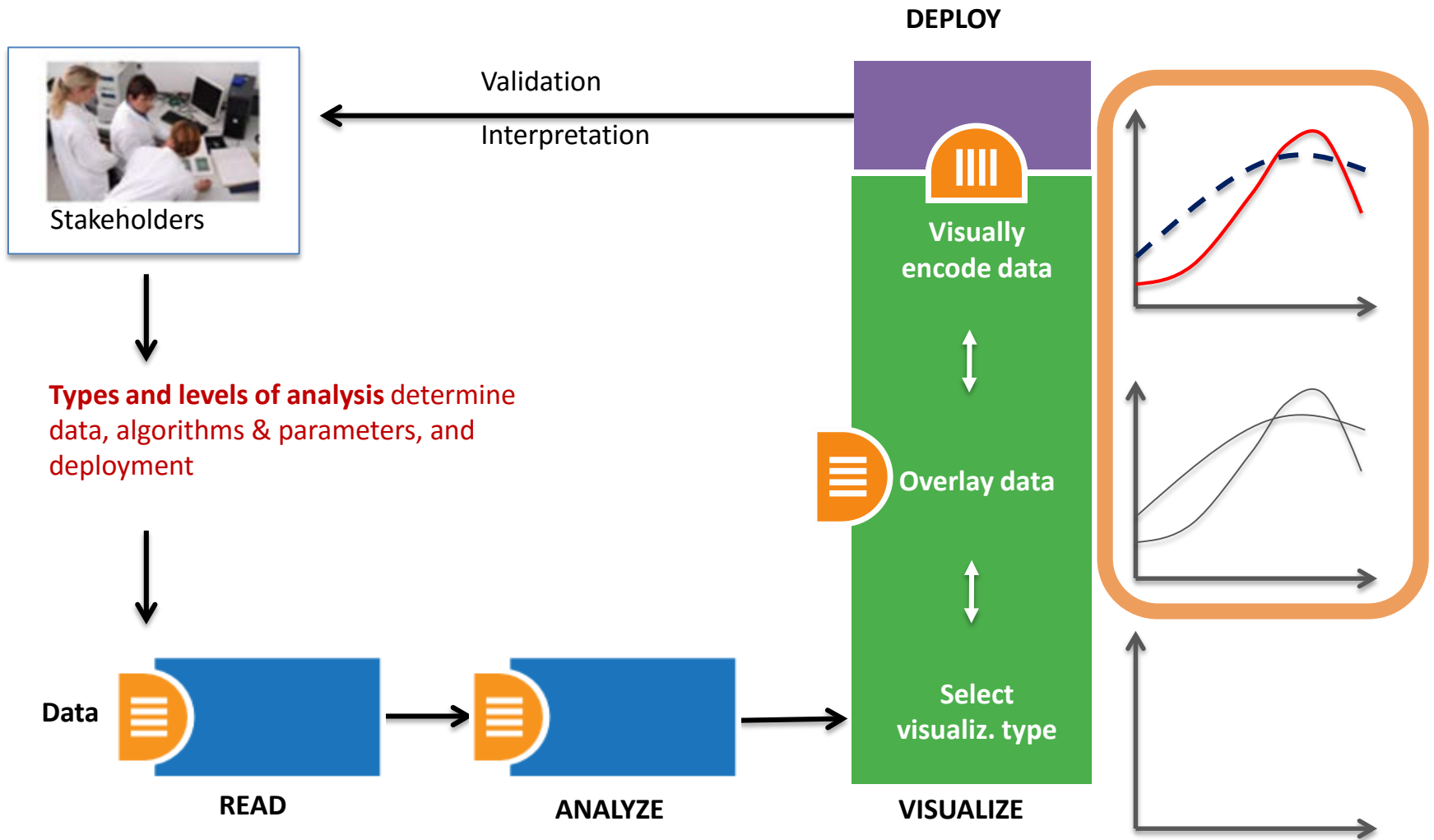
See page 24



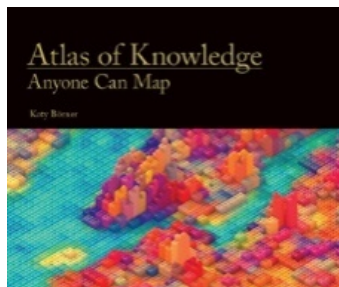
Insight Need Types page 26	Data Scale Types page 28	Visualization Types page 30	Graphic Symbol Types page 32	Graphic Variable Types page 34	Interaction Types page 26
<ul style="list-style-type: none"> • categorize/cluster • order/rank/sort • distributions (also outliers, gaps) • comparisons • trends (process and time) • geospatial • compositions (also of text) • correlations/relationships 	<ul style="list-style-type: none"> • nominal • ordinal • interval • ratio 	<ul style="list-style-type: none"> • table • chart • graph • map • network layout 	<ul style="list-style-type: none"> • geometric symbols <ul style="list-style-type: none"> point line area surface volume • linguistic symbols <ul style="list-style-type: none"> text numerals punctuation marks • pictorial symbols <ul style="list-style-type: none"> images icons statistical glyphs 	<ul style="list-style-type: none"> • spatial <ul style="list-style-type: none"> position • retinal <ul style="list-style-type: none"> form color optics motion 	<ul style="list-style-type: none"> • overview • zoom • search and locate • filter • details-on-demand • history • extract • link and brush • projection • distortion



See page 24



Insight Need Types page 26	Data Scale Types page 28	Visualization Types page 30	Graphic Symbol Types page 32	Graphic Variable Types page 34	Interaction Types page 26
<ul style="list-style-type: none"> • categorize/cluster • order/rank/sort • distributions (also outliers, gaps) • comparisons • trends (process and time) • geospatial • compositions (also of text) • correlations/relationships 	<ul style="list-style-type: none"> • nominal • ordinal • interval • ratio 	<ul style="list-style-type: none"> • table • chart • graph • map • network layout 	<ul style="list-style-type: none"> • geometric symbols <ul style="list-style-type: none"> point line area surface volume • linguistic symbols <ul style="list-style-type: none"> text numerals punctuation marks • pictorial symbols <ul style="list-style-type: none"> images icons statistical glyphs 	<ul style="list-style-type: none"> • spatial <ul style="list-style-type: none"> position • retinal <ul style="list-style-type: none"> form color optics motion 	<ul style="list-style-type: none"> • overview • zoom • search and locate • filter • details-on-demand • history • extract • link and brush • projection • distortion



See page 24

Questions?

8:00 Welcome and Overview of Tutorial and Attendees**8:30 Visualization Framework and Workflow Design**

- Overview of the Visualization framework
- Overview of Graphical variables, and color selection

9:15 Tool Overview and Trouble Shooting

- Sci2 Overview – scientometric analysis tool
- Open Refine – data parsing, transformation, and editing tool
- Gephi – network visualization

9:45 Geospatial Analysis

- Overview of geospatial analysis and mapping
- Geospatial Analysis: Geocoding with OpenRefine
- Geospatial Analysis: Proportional Symbol Map using CDC

11:00 Topical/Temporal Analysis: Burst Detection

- Overview of burst analysis and introductory workflow
- Burst Detection with CDC Grants

12:30 Lunch**1:30 Network Analysis**

- Introduction to Network Analysis
- ~~Network Analysis: Bimodal networks with Morbidity Data~~
- Network Analysis: Co-authorship Network with CDC publications

4:00 Wrap-up

4:30 Adjourn

Position: x, y; possibly z

Form:

- Size
- Shape
- Orientation/Rotation

Color:

• Value (Lightness)



• Hue (Tint)



• Saturation (Intensity)



Texture:

- Pattern, Rotation, Coarseness, Size, Density Gradient

Optics:

- Crispness, Transparency, Shading

Graphic Variable Types Versus Graphic Symbol Types

		Geometric Symbols			Linguistic Symbols		Pictorial Symbols	
		point	line	area	surface	volume	Text, Numerals, Punctuation Marks	Images, Icons, Statistical Glyphs
Symbol	x							
	y							
	z							
Form	size	NA, Not Applicable						
	shape	NA						
	orientation	NA						
	curvature	NA						
	angle	NA						
	closure	NA						
	color							
	saturation							
Texture	spacing							
	regularity							
	pattern							
	orientation	NA						
	coarseness							
	blur							
	transparency							
	blending							
	microscopic length	Point in foreground - background	Line in foreground - background	Area in foreground - background	Surface in foreground - background	Volume in foreground - background	Text in foreground - background	Icons in foreground - background
	speed							
Motion	velocity							
	acceleration	Blinking point slow - fast	Blinking line slow - fast	Blinking area slow - fast	Blinking surface slow - fast	Blinking volume slow - fast	Blinking text slow - fast	Blinking icons slow - fast

Graphic Variable Types Versus Graphic Symbol Types

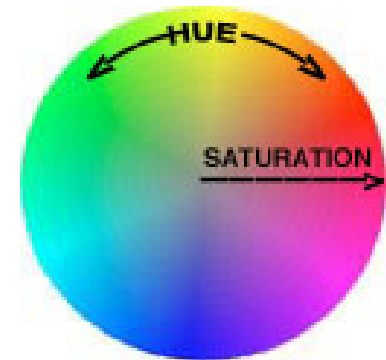
			Geometric Symbols					
			Point		Line		Area	
Spatial	x	quantitative						
	y	quantitative						
	z	quantitative						
Retinal	Form	Size	quantitative	NA (Not Applicable)				
		Shape	qualitative	NA				
		Rotation	quantitative	NA				
		Curvature	quantitative	NA				
		Angle	quantitative	NA				
		Closure	quantitative	NA				
	Color	Value	quantitative					
Hue		qualitative						
Saturation		quantitative						

Color may be used to

- convey importance or attract attention to specific symbols
- Label, categorize, compare
- imitate reality (e.g., blue lakes in maps)
- generate emotions—orange and red are perceived as warm and active while blue, purple are cold and passive.

Do NOT use color

- for displaying the layout of objects in space
- how they are moving, or
- what their shapes are.



Simultaneous contrast with surrounding or background colors can dramatically alter color appearance, making one color look like another or two similar colors look very different.

Color

- **Value**

(Lightness, shade, tone, percent value, density, intensity, luminance, brightness) equals amount of light coming from a source or being reflected from an object. Ratio between the maximum and the minimum brightness values is also called contrast ratio.

- **Hue**

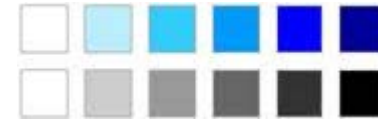
(Tint) related to the wavelength of the stimulus. Categorical and should never be used to encode magnitude. Need to select sequence carefully, e.g., yellow through orange to red.

- **Saturation**

(Intensity) is related to how much white content is in the stimulus. Monochromatic hues are very highly saturated. Higher saturated (purer) colors appear in the foreground while low saturation (dull) colors fade into background.

- **Sequential schemes**

Qualitative



(Single hue) best for ordered data that progress from low to high. Use light colors for low data values to dark colors for high data values. Example: heat maps or isomaps.

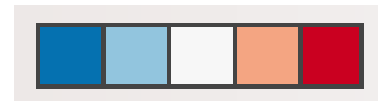
- **Binary Schemes**

Quantitative

(Two colors) use color opponents to show divergence such as black/white; red/green; yellow/blue.

- **Diverging schemes**

Quantitative



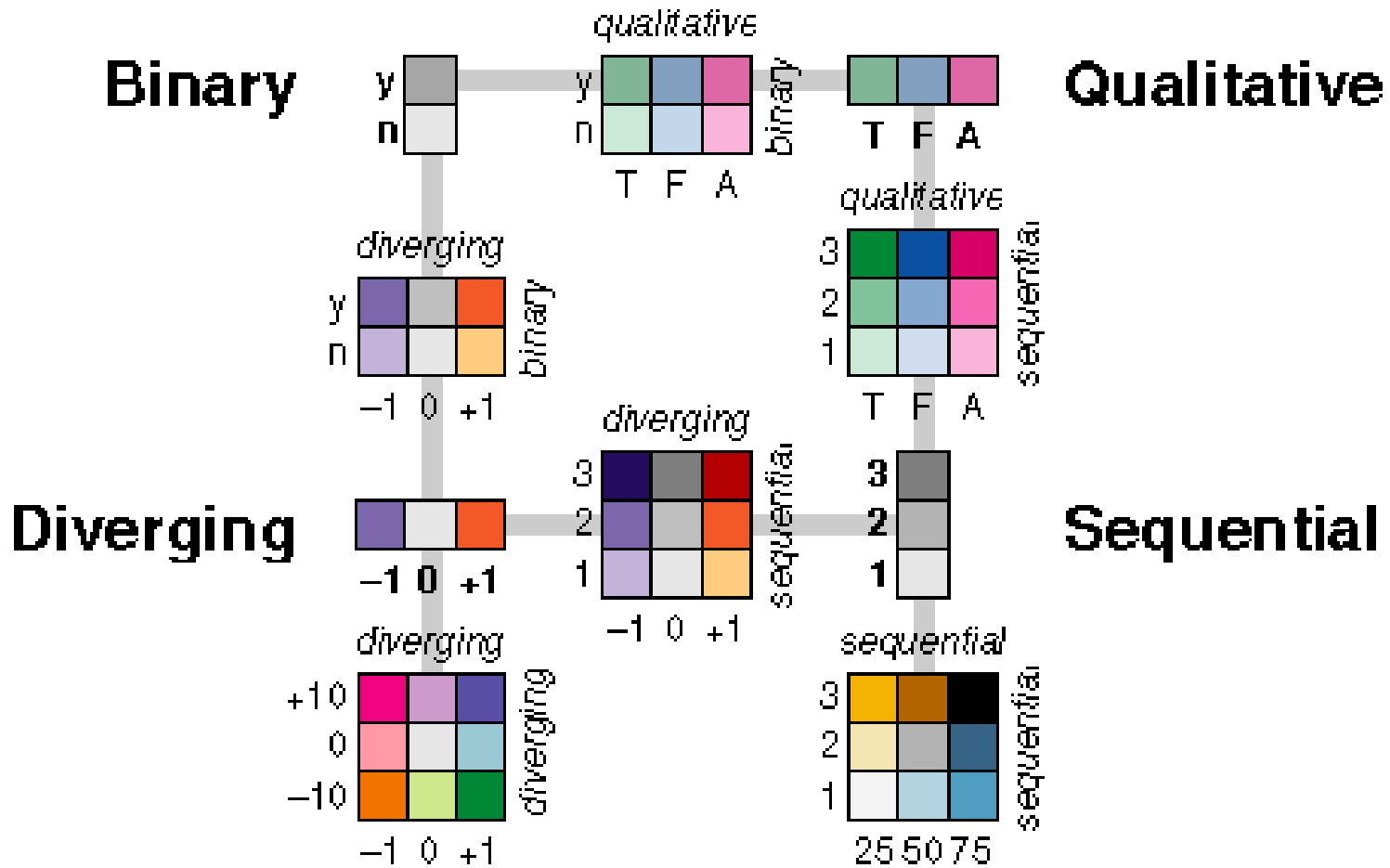
(Bi-polar) put equal emphasis on mid-range critical values and extremes at both ends of the data range. The critical class or break in the middle of the legend is emphasized with light colors and low and high extremes are emphasized with dark colors that have contrasting hues

- **Qualitative schemes**

Qualitative



(Full spectral) do not imply magnitude differences between legend classes, and hues are used to create the primary visual differences between classes. Qualitative schemes are best suited to representing nominal or categorical data.



Questions?

8:00 Welcome and Overview of Tutorial and Attendees**8:30 Visualization Framework and Workflow Design**

- Overview of the Visualization framework
- Overview of Graphical variables, and color selection

9:15 Tool Overview and Trouble Shooting

- Sci2 Overview – scientometric analysis tool
- Open Refine – data parsing, transformation, and editing tool
- Gephi – network visualization

9:45 Geospatial Analysis

- Overview of geospatial analysis and mapping
- Geospatial Analysis: Geocoding with OpenRefine
- Geospatial Analysis: Proportional Symbol Map using CDC

11:00 Topical/Temporal Analysis: Burst Detection

- Overview of burst analysis and introductory workflow
- Burst Detection with CDC Grants

12:30 Lunch**1:30 Network Analysis**

- Introduction to Network Analysis
- ~~Network Analysis: Bimodal networks with Morbidity Data~~
- Network Analysis: Co-authorship Network with CDC publications

4:00 Wrap-up

4:30 Adjourn

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.

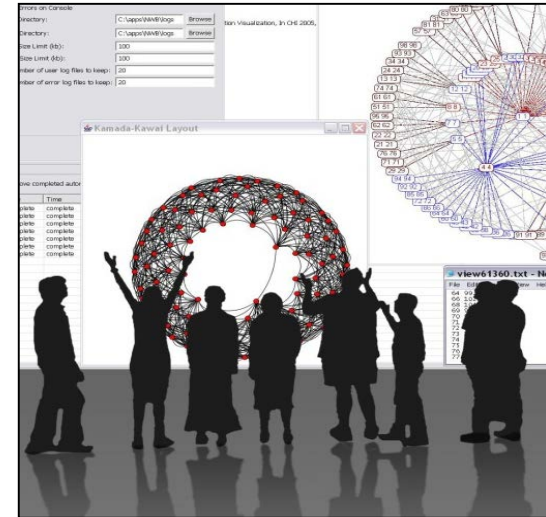
Rather than making things larger or smaller, macroscopes let us observe what is too great, slow, or complex for us to comprehend or sometimes even notice.



Microscopes



Telescopes



Macroscopes

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

8:00 Welcome and Overview of Tutorial and Attendees**8:30 Visualization Framework and Workflow Design**

- Overview of the Visualization framework
- Overview of Graphical variables, and color selection

9:15 Tool Overview and Trouble Shooting

- **Sci2 Overview – scientometric analysis tool**
- Open Refine – data parsing, transformation, and editing tool
- Gephi – network visualization

9:45 Geospatial Analysis

- Overview of geospatial analysis and mapping
- Geospatial Analysis: Geocoding with OpenRefine
- Geospatial Analysis: Proportional Symbol Map using CDC

11:00 Topical/Temporal Analysis: Burst Detection

- Overview of burst analysis and introductory workflow
- Burst Detection with CDC Grants

12:30 Lunch**1:30 Network Analysis**

- Introduction to Network Analysis
- Network Analysis: Bimodal networks with Morbidity Data
- Network Analysis: Co-authorship Network with CDC publications

4:00 Wrap-up

4:30 Adjourn



OSGi & Cyberinfrastructure Shell (CIShell)

- CIShell (<http://cishell.org>) is an open source software specification for the integration and utilization of datasets, algorithms, and tools
- It extends the Open Services Gateway Initiative (OSGi) (<http://osgi.org>), a standardized, modularized service platform
- CIShell provides “sockets” into which algorithms, tools, and datasets can be plugged using a wizard-driven process

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)

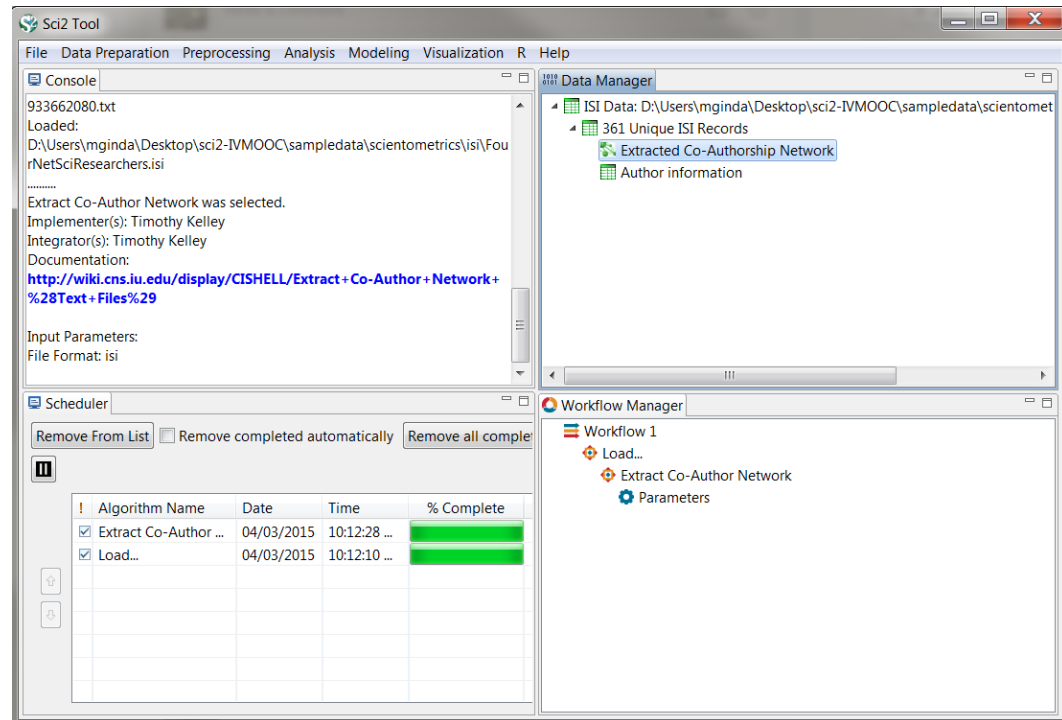
Image Formats

- JPEG (*.jpg)
- PDF (*.pdf)
- PostScript (*.ps)

Formats are documented at <http://sci2.wiki.cns.iu.edu/display/SCI2TUTORIAL/2.3+Data+Formats>.

Use

- **Menu** to read data, run algorithms.
- **Console** to see work log, references to seminal works.
- **Data Manager** to select, view, save loaded, simulated, or derived datasets.
- **Scheduler** to see status of algorithm execution.
- **Workflow Manager** to keep track of the algorithms and parameters run in an analysis



All workflows are recorded into a log file (see /sci2/logs/...). If errors occur, they are saved in a error log to ease bug reporting.

All algorithms are documented online; workflows are given in tutorials, see Sci2 Manual at <http://sci2.wiki.cns.iu.edu>

Problem: Sci2 will not run on my computer...

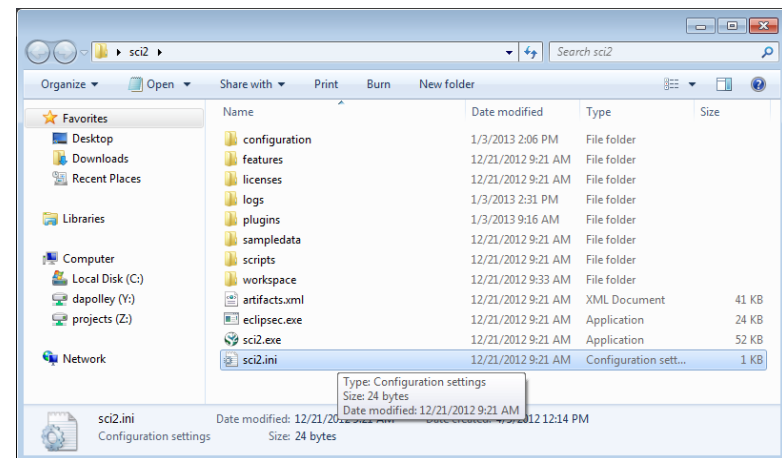
Solution: Sci2 runs on Java 1.6 (32 bit) or newer. If you are having trouble installing Sci2, you may want to install the latest 32 bit version of Java. You can run 32 bit and 64 bit versions of Java simultaneously.

After you've installed 32-bit Java, you will need to indicate the path that Sci2 uses to correct version of Java.

Target the **javaw.exe** file – which is likely located in the directory, **C:\Program Files (x86)\Java\jre7\bin**, assuming you have installed Java in the default place.

Last, you'll need to open directory where Sci2 is installed, and open **sci2.ini** using an editor, like Notepad++, and delete the contents, and replace them with this:

```
-vm
C:\Program Files
(x86)\Java\jre7\bin\javaw.exe
-vmargs
-Xms15m
-Xmx350m
```



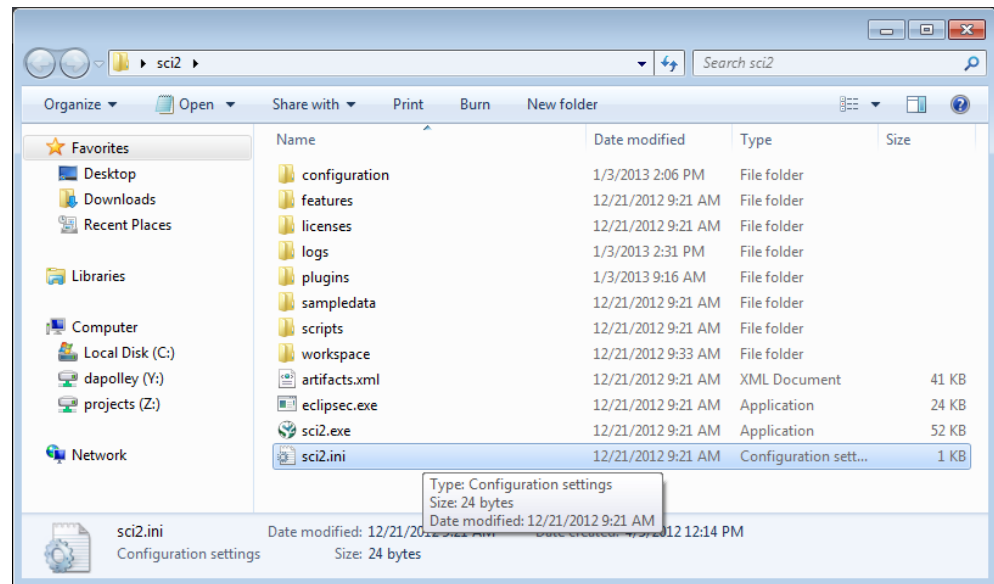
Problem: Sci2 is taking a long time to complete a process...

Solution: Due to the constraints of the Java virtual machine, the amount of memory available to a Java application must be determined before the application starts. The current default allotment of 350 Megabytes is a balance between providing enough memory for most uses of the tool, while not causing the Sci2 Tool to crash on machines with too little memory.

To add more virtual memory for Sci2 to use, open the file "**sci.ini**" in your Sci2 directory and edit the lines:

-vmargs	→	-vmargs
-Xms15m		-Xms256m
-Xmx350m		-Xmx512m

The recommended maximum virtual memory for Sci2 is 1024m, and up to 1536m.



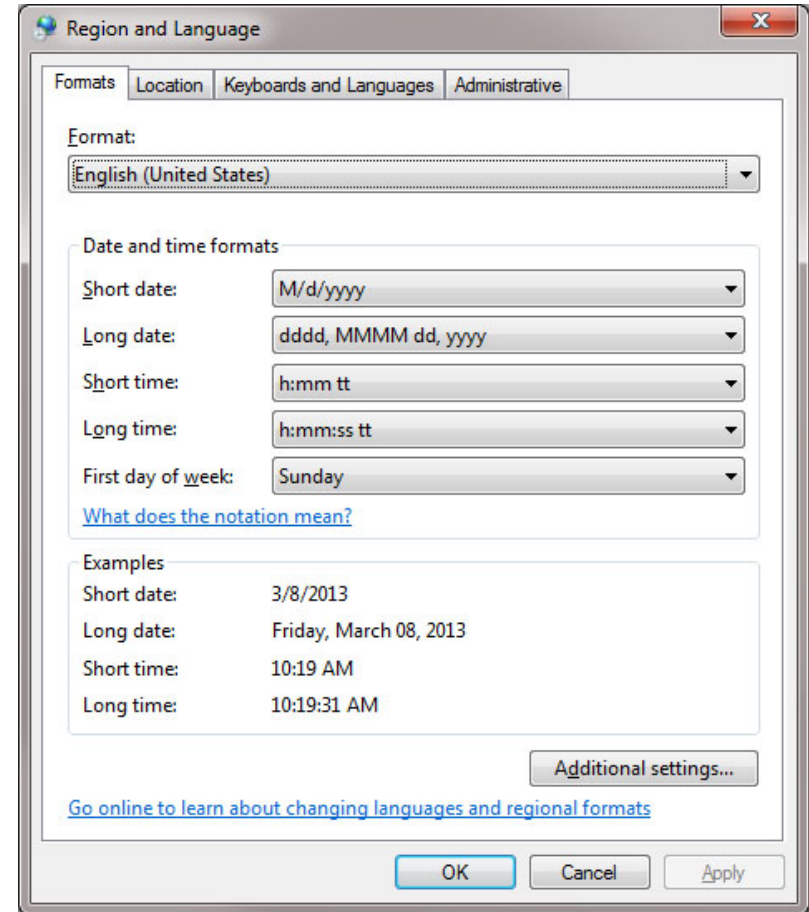
Problem: I'm having trouble loading data into Sci2 and having trouble running the temporal bar graph algorithm...

Solution: Your system is likely set up to a locale other than the United States.

If you are working on a Windows machine and you want to change the locale, go to the Start Menu and select Control Panel and then select Region and Language, then change the format to English (United States)

Mac and Linux users should refer to the Changing System Locale Guide is available here:

<http://wiki.cns.iu.edu/display/SCI2TUTORIAL/Changing+System+Locale>



8:00 Welcome and Overview of Tutorial and Attendees**8:30 Visualization Framework and Workflow Design**

- Overview of the Visualization framework
- Overview of Graphical variables, and color selection

9:15 Tool Overview and Trouble Shooting

- Sci2 Overview – scientometric analysis tool
- Open Refine – data parsing, transformation, and editing tool
- Gephi – network visualization

9:45 Geospatial Analysis

- Overview of geospatial analysis and mapping
- Geospatial Analysis: Geocoding with OpenRefine
- Geospatial Analysis: Proportional Symbol Map using CDC

11:00 Topical/Temporal Analysis: Burst Detection

- Overview of burst analysis and introductory workflow
- Burst Detection with CDC Grants

12:30 Lunch**1:30 Network Analysis**

- Introduction to Network Analysis
- ~~Network Analysis: Bimodal networks with Morbidity Data~~
- Network Analysis: Co-authorship Network with CDC publications

4:00 Wrap-up

4:30 Adjourn



The screenshot shows the OpenRefine website homepage. At the top, there's a navigation bar with social media links for GitHub and Twitter. Below that is a large banner with the text "OPEN Refine" and a blue diamond icon, followed by the tagline "A free, open source, powerful tool for working with messy data".

The main content area features a "#Welcome!" section with a brief introduction to OpenRefine (formerly Google Refine) and a note about its rebranding. Below this is a section titled "Using OpenRefine - The Book" which includes a list of topics covered in the book:

1. Import data in various formats
2. Explore datasets in a matter of seconds
3. Apply basic and advanced cell transformations
4. Deal with cells that contain multiple values
5. Create instantaneous links between datasets
6. Filter and partition your data easily with regular expressions
7. Use named-entity extraction on full-text fields to automatically identify topics
8. Perform advanced data operations with the General Refine Expression Language

Below the list is a section titled "Introduction to OpenRefine" with a sub-section "1. Explore Data". It includes a video player showing a "Google Refine 2.0 - Introduction (1 of 3) (video version 2)".

On the left side of the page, there is a sidebar with navigation links: Home, Download, Documentation, Community, Post archive, OpenRefine News: Spring 2016, OpenRefine News: December 2015, OpenRefine News: November 2015, OpenRefine News: October 2015, OpenRefine News: September 2015, OpenRefine News: August 2015, OpenRefine News: July 2015, OpenRefine News: June 2015, Mapping OpenRefine Ecosystem, 2014 survey results, A Governance Model for OpenRefine, and Using OpenRefine: a

OpenRefine (formerly GoogleRefine) tool for working with messy data: cleaning it; transforming it from one format into another; and extending it with web services and external data.

Active development and user community that have [created online tutorials](#) and [documentation](#) to support new users.

Imports wide variety of data formats, and processes a wide variety of data types (string, Boolean, arrays, dates, math) that lets a user flexibly filtering/faceting, edit and transform their data with regular expressions (GREL language) and JYTHON.

8:00 Welcome and Overview of Tutorial and Attendees**8:30 Visualization Framework and Workflow Design**

- Overview of the Visualization framework
- Overview of Graphical variables, and color selection

9:15 Tool Overview and Trouble Shooting

- Sci2 Overview – scientometric analysis tool
- Open Refine – data parsing, transformation, and editing tool
- **Gephi – network visualization**

9:45 Geospatial Analysis

- Overview of geospatial analysis and mapping
- Geospatial Analysis: Geocoding with OpenRefine
- Geospatial Analysis: Proportional Symbol Map using CDC

11:00 Topical/Temporal Analysis: Burst Detection

- Overview of burst analysis and introductory workflow
- Burst Detection with CDC Grants

12:30 Lunch**1:30 Network Analysis**

- Introduction to Network Analysis
- Network Analysis: Bimodal networks with Morbidity Data
- Network Analysis: Co-authorship Network with CDC publications

4:00 Wrap-up

4:30 Adjourn

Gephi is an open source network analysis and visualization tool developed by the [Gephi Consortium](#).

Gephi allows users to perform exploratory analysis of various types of networks. Gephi imports a wide variety of network formats: GDF (GUESS), GraphML (NodeXL), GML, NET (Pajek), GEXF. Gephi also has a wide user community that offers support and actively develops customizable plugins for layouts, metrics, data sources, manipulation tools, rendering presets and more.

A [full list of features](#) is found on the Gephi.org website. [Documentation](#) & [tutorials](#) for the tool, [system requirements](#) and [GitHub repository](#) also are useful resources.

Questions?

8:00 Welcome and Overview of Tutorial and Attendees**8:30 Visualization Framework and Workflow Design**

- Overview of the Visualization framework
- Overview of Graphical variables, and color selection

9:15 Tool Overview and Trouble Shooting

- Sci2 Overview – scientometric analysis tool
- Open Refine – data parsing, transformation, and editing tool
- Gephi – network visualization

9:45 Geospatial Analysis

- Overview of geospatial analysis and mapping
- Geospatial Analysis: Geocoding with OpenRefine
- Geospatial Analysis: Proportional Symbol Map using CDC

11:00 Topical/Temporal Analysis: Burst Detection

- Overview of burst analysis and introductory workflow
- Burst Detection with CDC Grants

12:30 Lunch**1:30 Network Analysis**

- Introduction to Network Analysis
- ~~Network Analysis: Bimodal networks with Morbidity Data~~
- Network Analysis: Co-authorship Network with CDC publications

4:00 Wrap-up

4:30 Adjourn

- **Geospatial analysis** is a set of methods to represent data as they relate to locations, such as geographic coordinates and boundaries. Geospatial analysis involves can be complex in how data is collected, transformed and encoded into maps and other visualizations.
- Sci2's Geospatial Analysis capabilities are strongest in creating prototype or publication ready proportional symbol maps and choropleth maps for the US state level and world map projections.
- Additionally, Sci2 allows for the creation of network overlays for US and World map projections.

Terminology

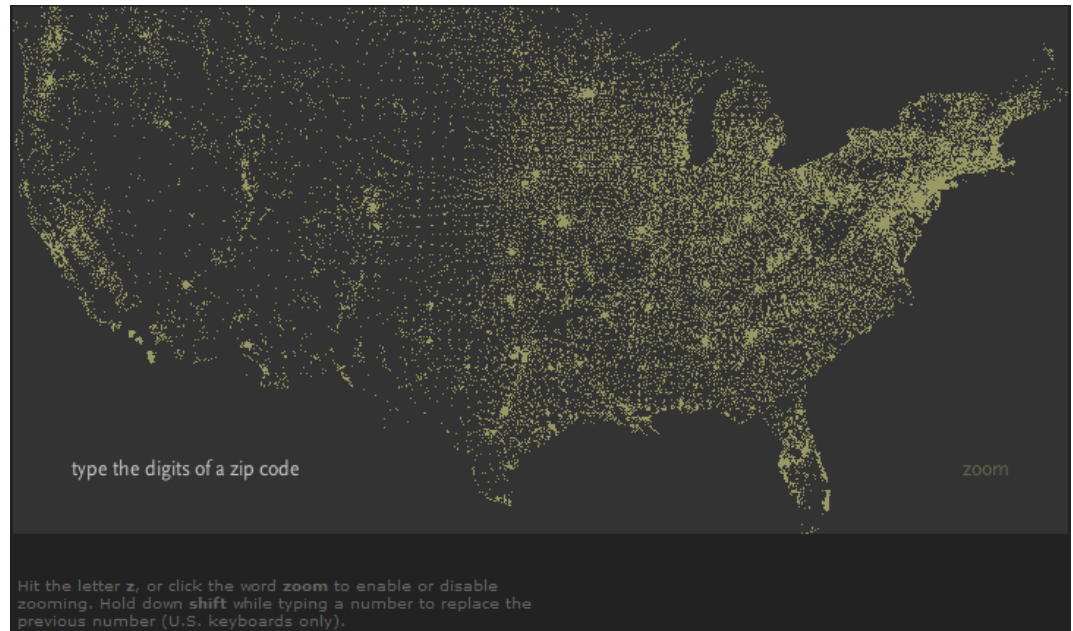
- **Geocode:** Location of a record (e.g., address, census tract, postal code, geographic coordinates).
- **Geographic coordinates:** Locations on the surface of the Earth expressed in degrees of latitude and longitude.
- **Geodesic:** The shortest distance between two points on the surface of a spheroid.
- **Great Circle:** Shortest distance between two points on Earth—i.e., a circular line which runs around the Earth at its fattest point.
- **Gazetteers:** Lists of geographic places and their coordinates, along with other information such as area, population, and cultural statistics used to geocode—see [Bing Geocoder in Sci2 Wiki](#).

Representation of Geospatial Data

- Addresses
- US Zip codes, see <http://benfry.com/zipdecode>
- US Census blocks
- US Congressional districts
- US States
- Countries



- Latitude/Longitude

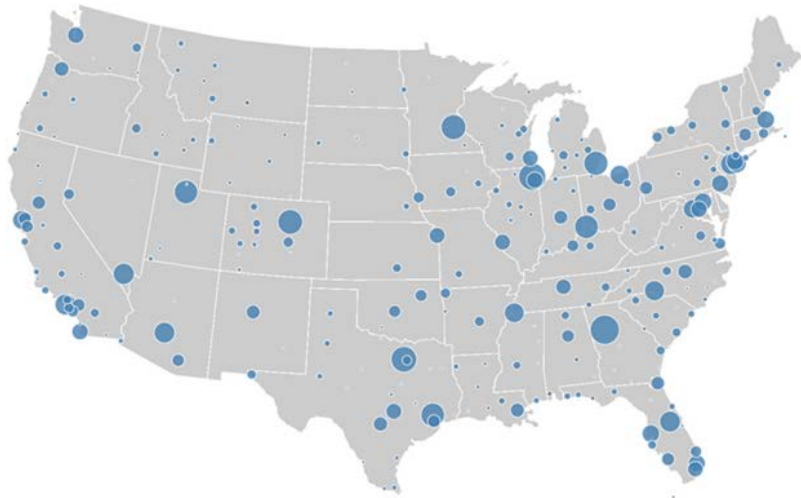


Map Types

Proportional symbol map

Represents data variables by symbols that are sized, colored, etc. according to their amount. Data is (or can be) aggregated at points within areas.

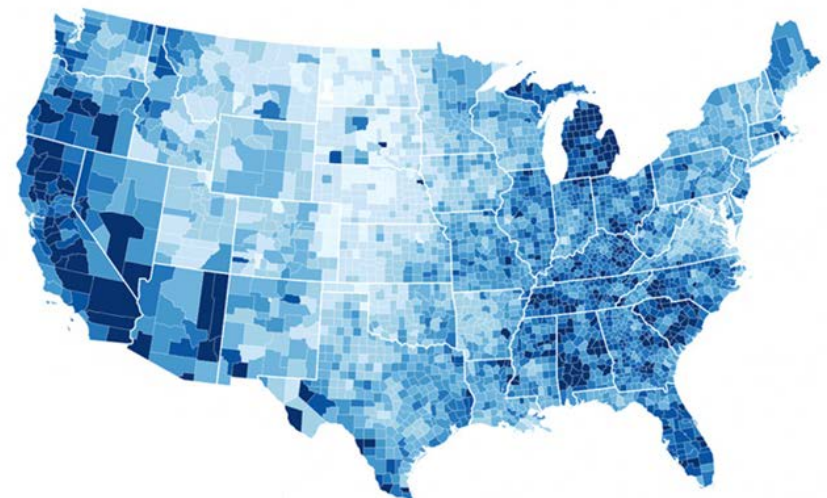
Do NOT use for densities, ratios, or scales, which should be rendered as choropleth map.



Choropleth map

Represents data variables such as densities, ratios, or rates by proportionally colored or patterned areas.

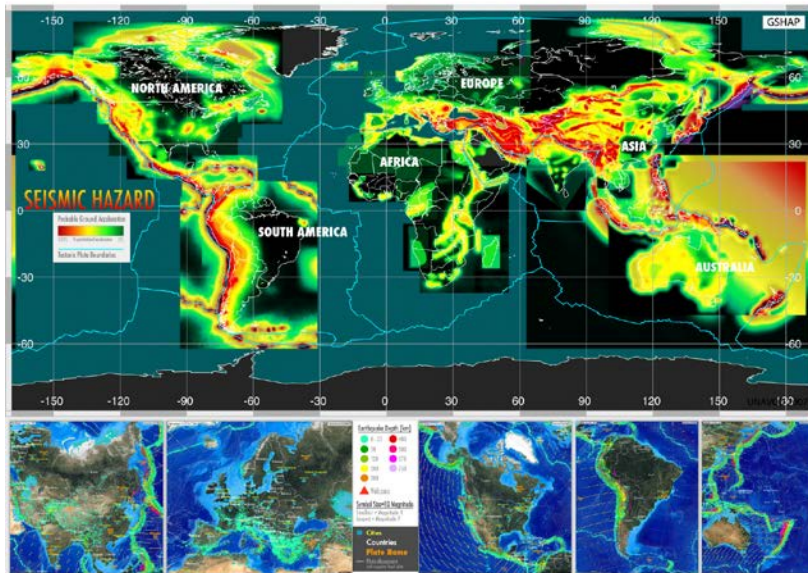
Each artificial collection unit is called a *chronogram* and has a distinctive color or shading.



Map Types

Heat (isopleth) maps

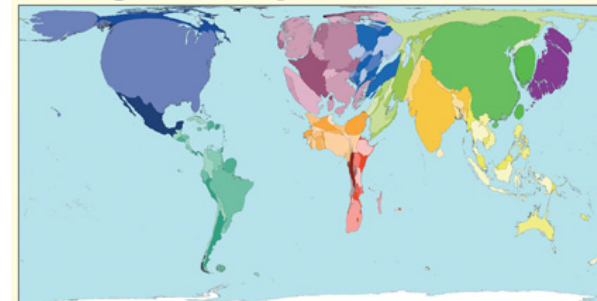
represent continuous data variable values by colors. While choropleth maps color predefined regions, heat maps might show color-based contour lines that connect points of equal value or value-by-area maps.



Cartograms

are not drawn to scale. Instead, they distort geographical areas in proportion to data values. Familiarity with regions is necessary. Mostly used for world, continental, and country maps.

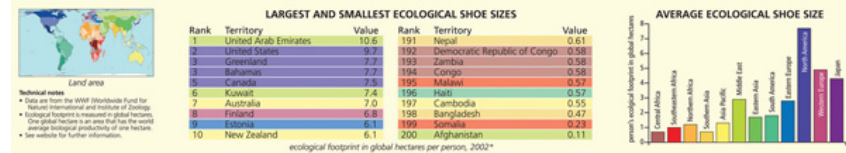
Ecological Footprint



The ecological footprint is a measure of the area needed to support a population's lifestyle. This includes the consumption of food, fuel, wood, and fibres. Pollution, such as carbon dioxide emissions, is also counted as part of the footprint.

The United States, China and India have the largest ecological footprints. Without knowing population size we cannot understand what this means about individuals' ecological demands. Large populations live in China and India, in both territories resource use is below the world average. The per person footprint in the United States is almost five times the world average, and almost ten times what would be sustainable.

Territory size shows the proportion of the worldwide ecological footprint which is made there.

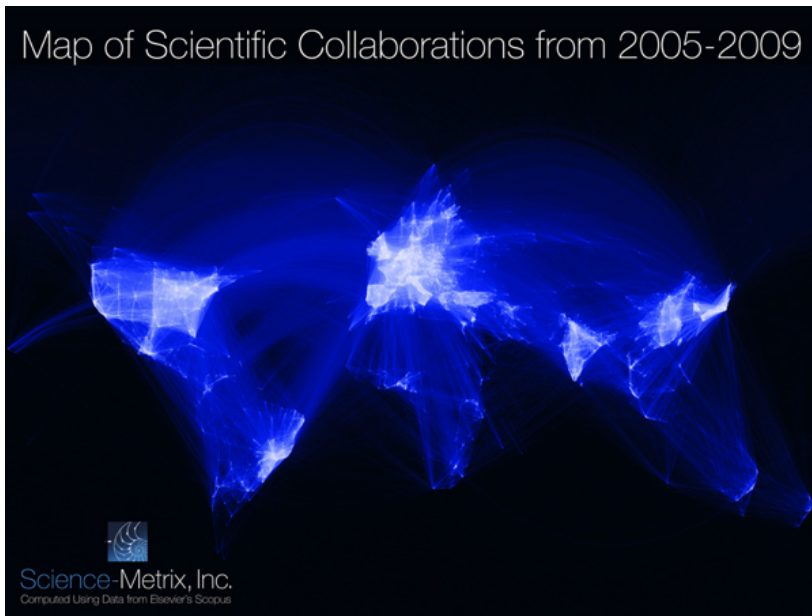


"People consume resources and ecological services from all over the world, so their footprint is the sum of these areas, wherever they may be on the planet."
 The Living Planet Report, 2006
 www.worldmapper.org © Copyright 2006 SAO Group (Sheffield) and Mark Newman (University of Michigan) Map 322

Map Types

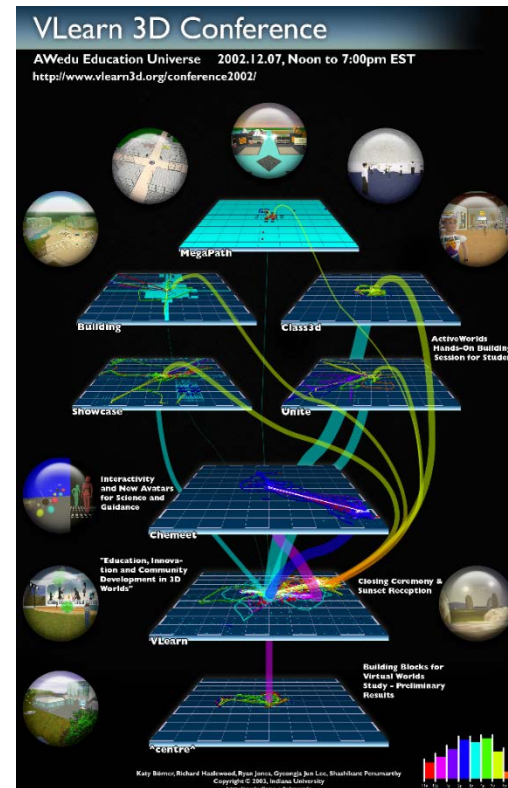
Flow maps

show the paths that (in)tangible objects take to get from one geospatial place to another. Variables such as capacity or maximum speed are encoded proportionally by line width or color.



Space-time cubes

Display entities, locations, and events over time.



Questions?

8:00 Welcome and Overview of Tutorial and Attendees**8:30 Visualization Framework and Workflow Design**

- Overview of the Visualization framework
- Overview of Graphical variables, and color selection

9:15 Tool Overview and Trouble Shooting

- Sci2 Overview – scientometric analysis tool
- Open Refine – data parsing, transformation, and editing tool
- Gephi – network visualization

9:45 Geospatial Analysis

- Overview of geospatial analysis and mapping
- **Geospatial Analysis: Geocoding with OpenRefine**
- Geospatial Analysis: Proportional Symbol Map using CDC

11:00 Topical/Temporal Analysis: Burst Detection

- Overview of burst analysis and introductory workflow
- Burst Detection with CDC Grants

12:30 Lunch**1:30 Network Analysis**

- Introduction to Network Analysis
- ~~Network Analysis: Bimodal networks with Morbidity Data~~
- Network Analysis: Co-authorship Network with CDC publications

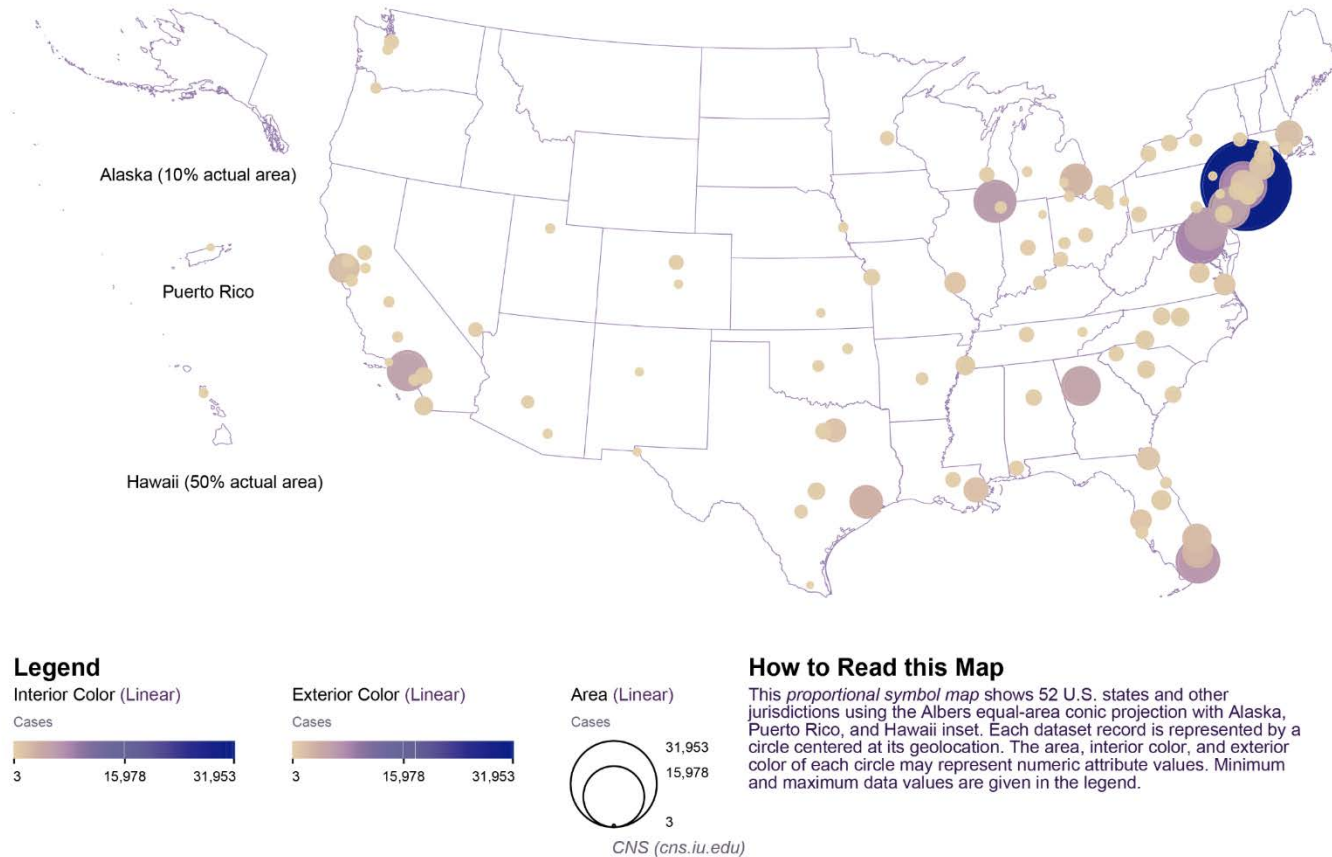
4:00 Wrap-up

4:30 Adjourn

Proportional symbol map

Represents data variables by symbols that are sized, colored, etc. according to their amount. Data is (or can be) aggregated at points within areas.

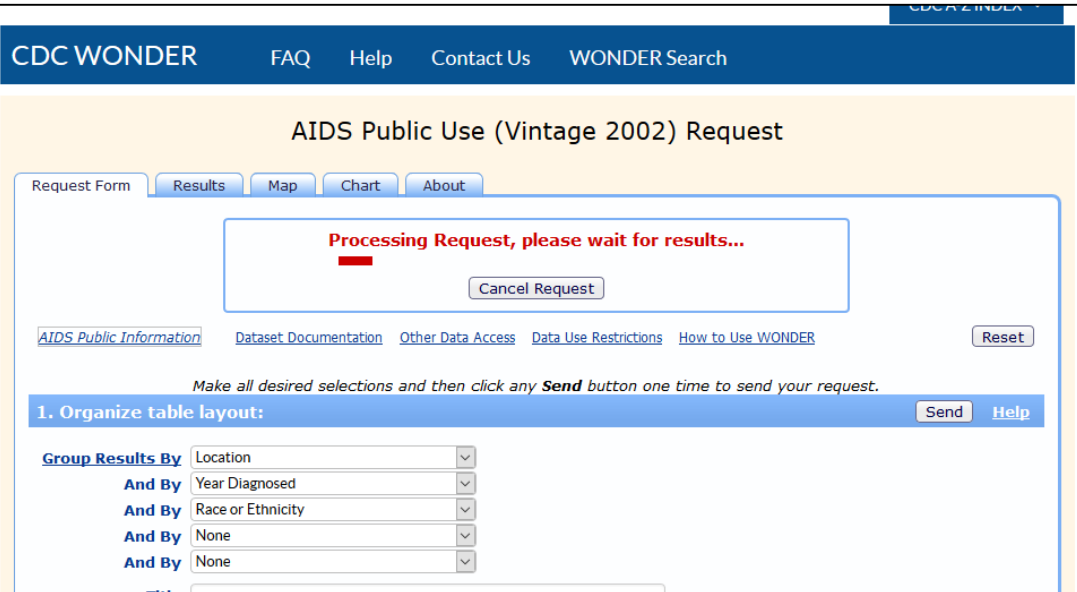
Do NOT use for densities, ratios, or scales, which should be rendered as choropleth map.



The AIDS Public Information Dataset U.S. Surveillance

“Current HIV/AIDS data and statistics are available from the National Center for HIV/AIDS, Viral Hepatitis, STD, and TB Prevention (NCHHSTP) in the [NCHHSTP Atlas](#).”

This workflow uses the **1981 – 2002 Archive case reports**: By date, place, demographics, case definition, risk factors and vital status.



CDC WONDER FAQ Help Contact Us WONDER Search

AIDS Public Use (Vintage 2002) Request

Request Form Results Map Chart About

Processing Request, please wait for results...

Cancel Request

[AIDS Public Information](#) [Dataset Documentation](#) [Other Data Access](#) [Data Use Restrictions](#) [How to Use WONDER](#) Reset

Make all desired selections and then click any **Send** button one time to send your request.

1. Organize table layout: Send Help

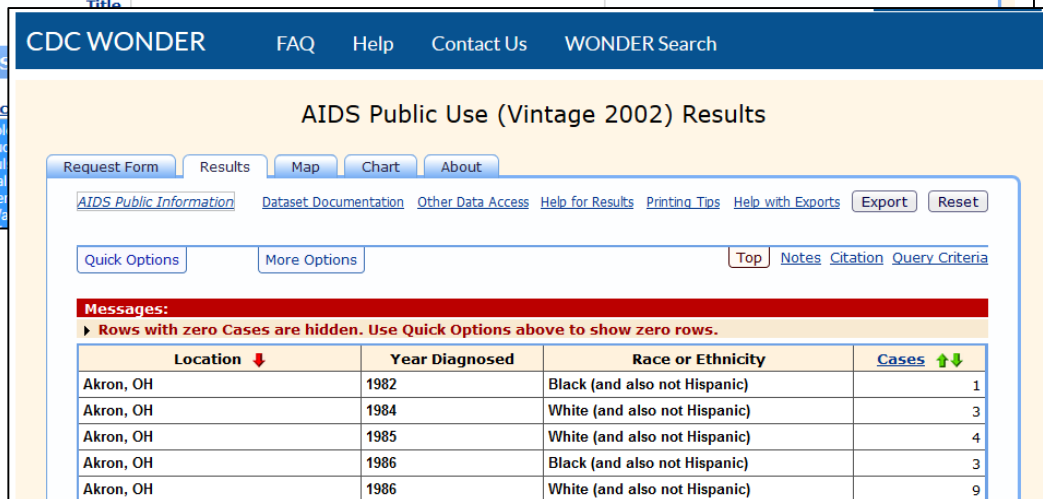
Group Results By: Location

And By: Year Diagnosed

And By: Race or Ethnicity

And By: None

And By: None



CDC WONDER FAQ Help Contact Us WONDER Search

AIDS Public Use (Vintage 2002) Results

Request Form Results Map Chart About

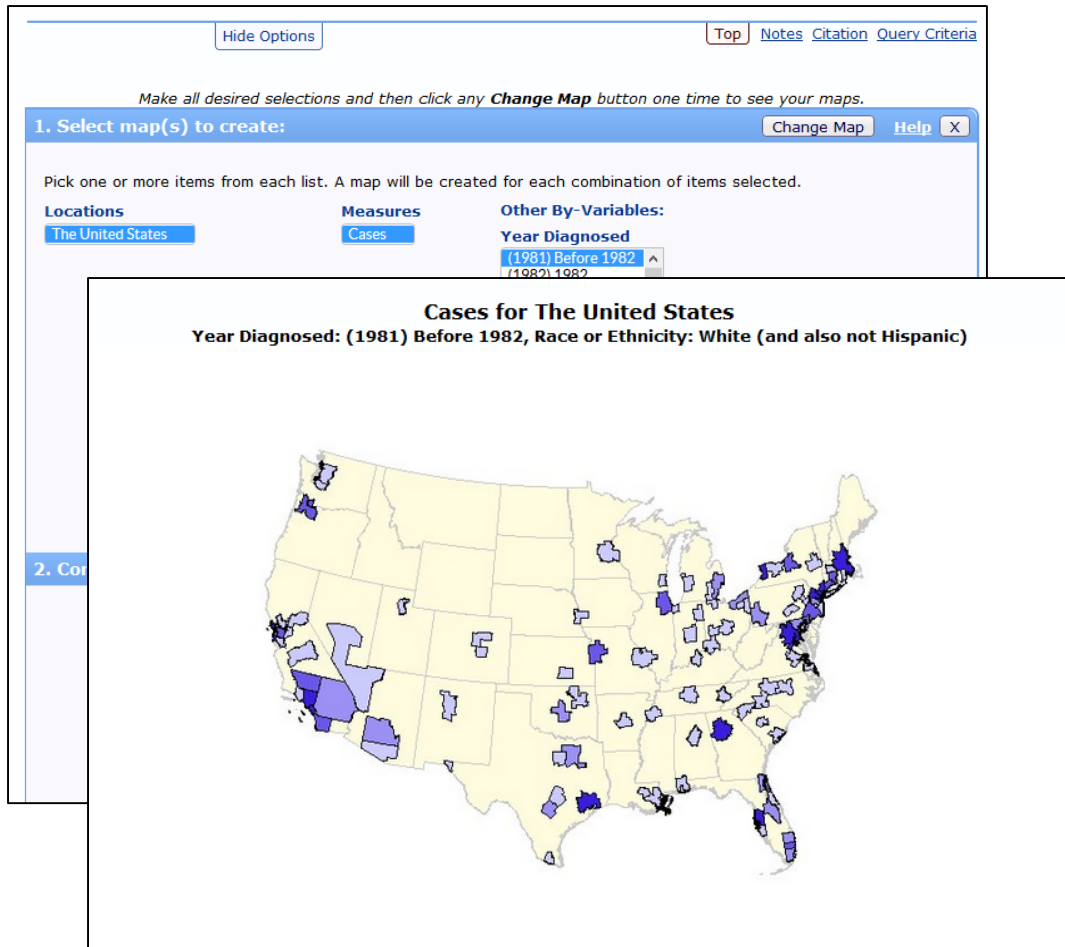
[AIDS Public Information](#) [Dataset Documentation](#) [Other Data Access](#) [Help for Results](#) [Printing Tips](#) [Help with Exports](#) Export Reset

Quick Options More Options Top Notes Citation Query Criteria

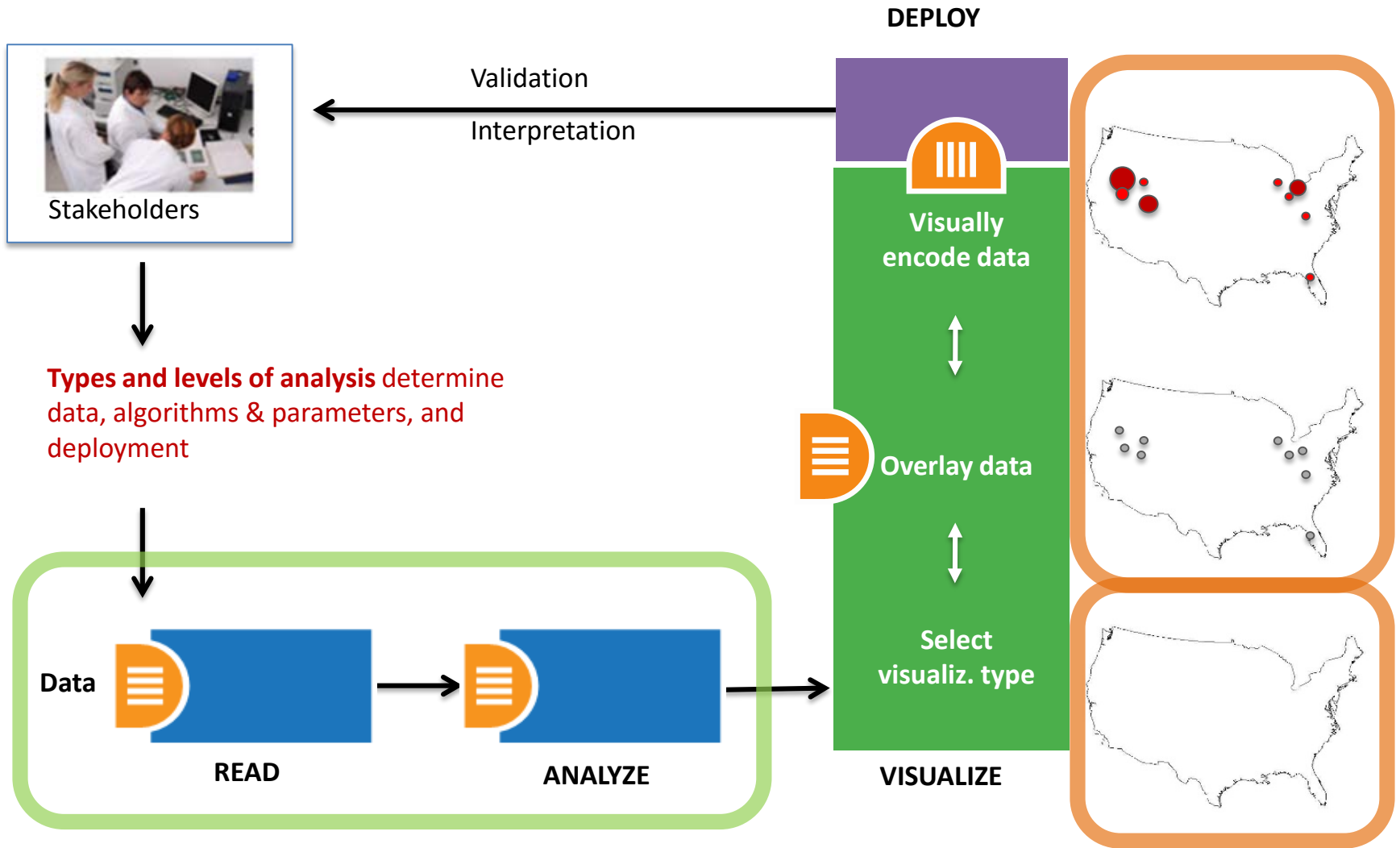
Messages:

► Rows with zero Cases are hidden. Use Quick Options above to show zero rows.

Location ↓	Year Diagnosed	Race or Ethnicity	Cases ↑↓
Akron, OH	1982	Black (and also not Hispanic)	1
Akron, OH	1984	White (and also not Hispanic)	3
Akron, OH	1985	White (and also not Hispanic)	4
Akron, OH	1986	Black (and also not Hispanic)	3
Akron, OH	1986	White (and also not Hispanic)	9



- Mapping on this data base is available, that provides many methods to subset queried data across variables provided.
- Limited to JPEG raster image.
- Works only with data sets available on CDC Wonder.



Google refine A power tool for working with messy data.

« Start Over Configure Parsing Options Project name: CDC AIDS Diagnosis CityLocation 8196 Create Project »

Notes	Location	Location Code	Year Diagnosed	Year Diagnosed Code	Race or Ethnicity	Race or Ethnicity Code	Cases
1.	Akron, OH	80	1982	1982	Black (and also not Hispanic)	2054-5	1
2.	Akron, OH	80	1984	1984	White (and also not Hispanic)	2106-3	3
3.	Akron, OH	80	1985	1985	White (and also not Hispanic)	2106-3	4
4.	Akron, OH	80	1986	1986	Black (and also not Hispanic)	2054-5	3
5.	Akron, OH	80	1986	1986	White (and also not Hispanic)	2106-3	9
6.	Akron, OH	80	1987	1987	Black (and also not Hispanic)	2054-5	6
7.	Akron, OH	80	1987	1987	White (and also not Hispanic)	2106-3	14
8.	Akron, OH	80	1988	1988	Black (and also not Hispanic)	2054-5	11
9.	Akron, OH	80	1988	1988	Hispanic	2135-2	1
10.	Akron, OH	80	1988	1988	White (and also not Hispanic)	2106-3	15
11.	Akron, OH	80	1989	1989	Black (and also not Hispanic)	2054-5	7
12.	Akron, OH	80	1989	1989	White (and also not Hispanic)	2106-3	19
13.	Akron, OH	80	1990	1990	Black (and also not Hispanic)	2054-5	6
14.	Akron, OH	80	1990	1990	White (and also not Hispanic)	2106-3	23
15.	Akron, OH	80	1991	1991	Black (and also not Hispanic)	2054-5	11
16.	Akron, OH	80	1991	1991	White (and also not Hispanic)	2106-3	38
17.	Akron, OH	80	1992	1992	Black (and also not Hispanic)	2054-5	17
18.	Akron, OH	80	1992	1992	White (and also not Hispanic)	2106-3	41
19.	Akron, OH	80	1993	1993	Black (and also not Hispanic)	2054-5	28
20.	Akron, OH	80	1993	1993	White (and also not Hispanic)	2106-3	37
21.	Akron, OH	80	1994	1994	Black (and also not Hispanic)	2054-5	15
22.	Akron, OH	80	1994	1994	Hispanic	2135-2	1
23.	Akron, OH	80	1994	1994	White (and also not Hispanic)	2106-3	31
24.	Akron, OH	80	1995	1995	Black (and also not Hispanic)	2054-5	27
25.	Akron, OH	80	1995	1995	Hispanic	2135-2	1
26.	Akron, OH	80	1995	1995	White (and also not Hispanic)	2106-3	32
27.	Akron, OH	80	1996	1996	Black (and also not Hispanic)	2054-5	17

Parse data as

Character encoding Update Preview

CSV / TSV / separator-based files

Line-based text files

Fixed-width field text files

PC-Axis text files

JSON files

RDF/JSON files

XML files

Open Document Format spreadsheets (ods)

RDF/XML files

Columns are separated by

commas (CSV)

tabs (TSV)

custom 't'

Escape special characters with \

Ignore first 0 line(s) at beginning of file

Parse next 1 line(s) as column headers

Discard initial 0 row(s) of data

Load at most 0 row(s) of data

Parse cell text into numbers, dates, ...

Quotation marks are used to enclose cells containing column separators

Store blank rows

Store blank cells as nulls

Store file source (file names, URLs) in each row

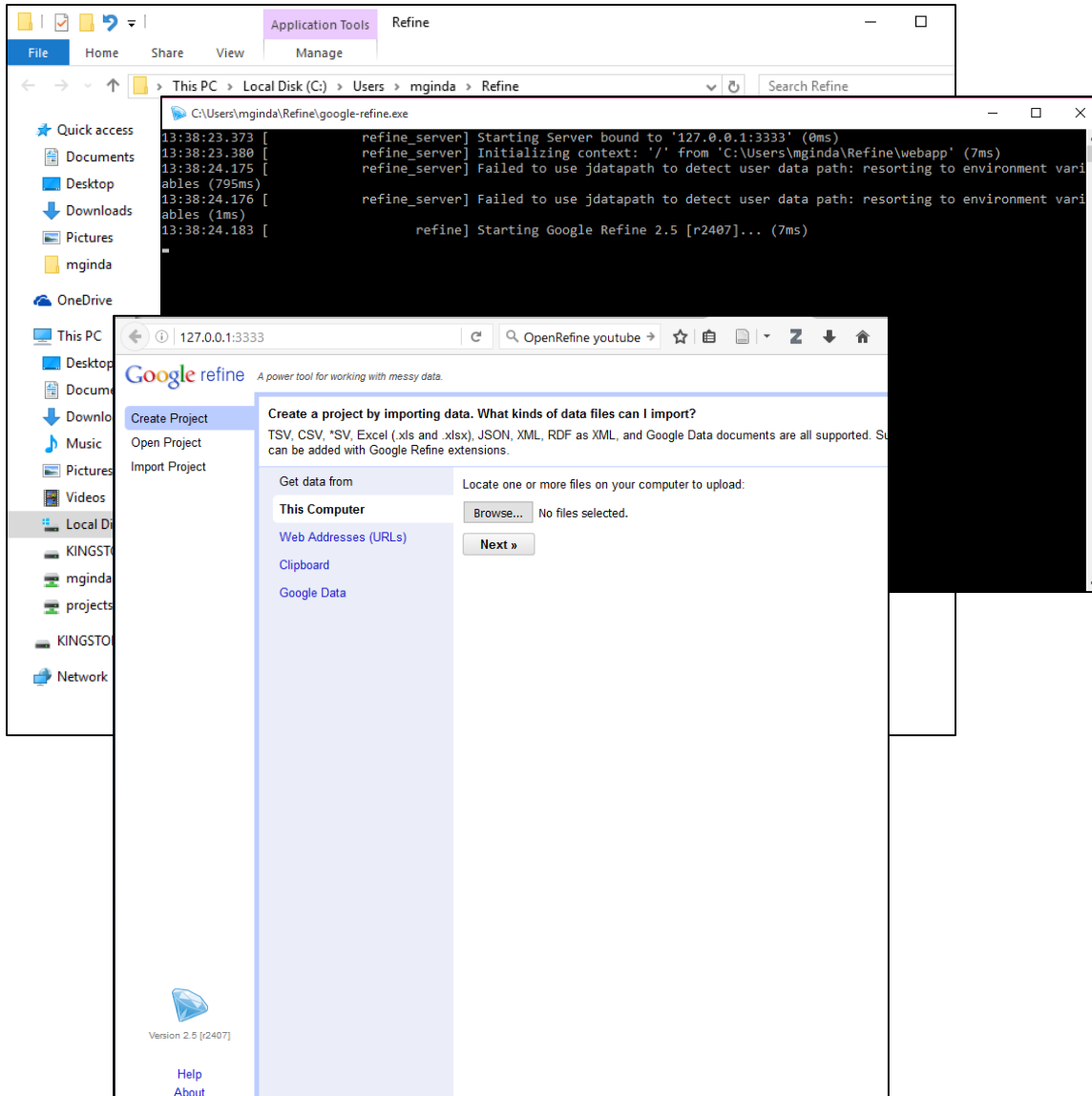
Version 2.5 (2407)

Help About

For this portion of the workshop, we will use OpenRefine to review the data set, filter and facet data, and geocode the locations.

The process of geolocation uses Google's Geocoding API to identify latitude and longitude positions for address, city or state, country or place names.

Other open geocoding APIs may be used in place of the Google geocoder.

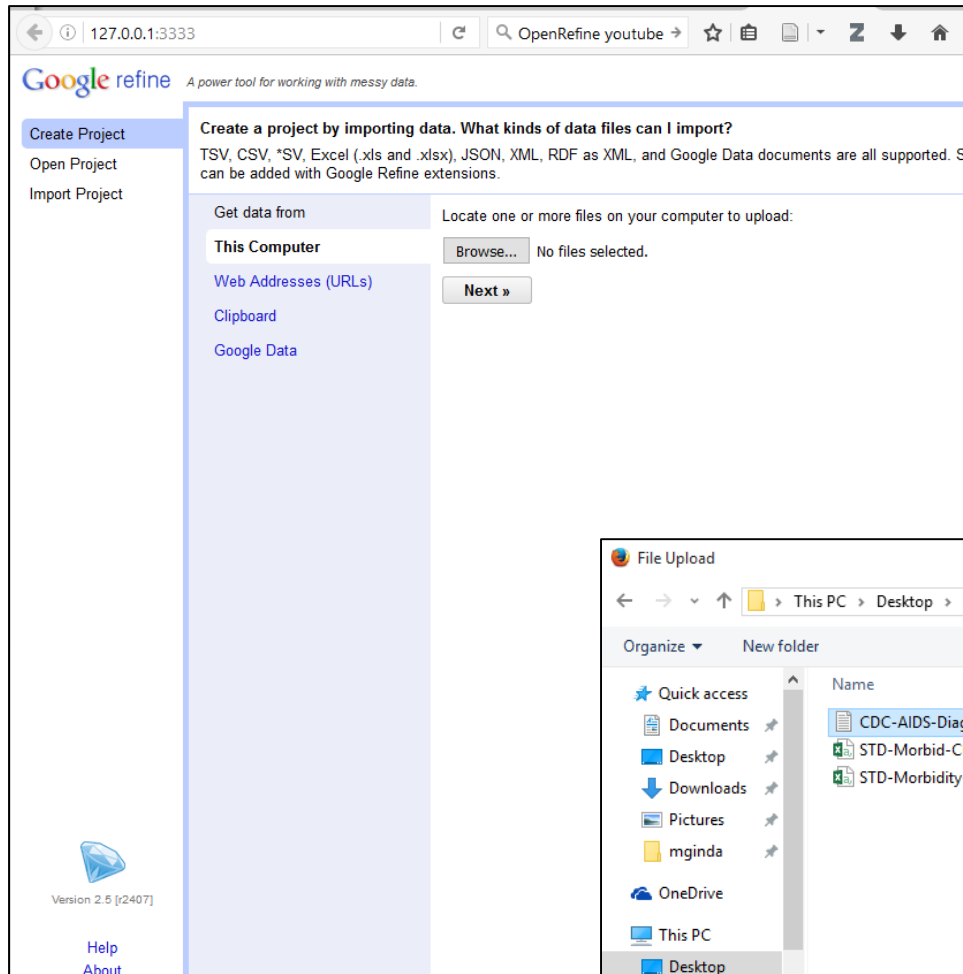


First we need to start OpenRefine. Navigate to the Desktop folder Refine.

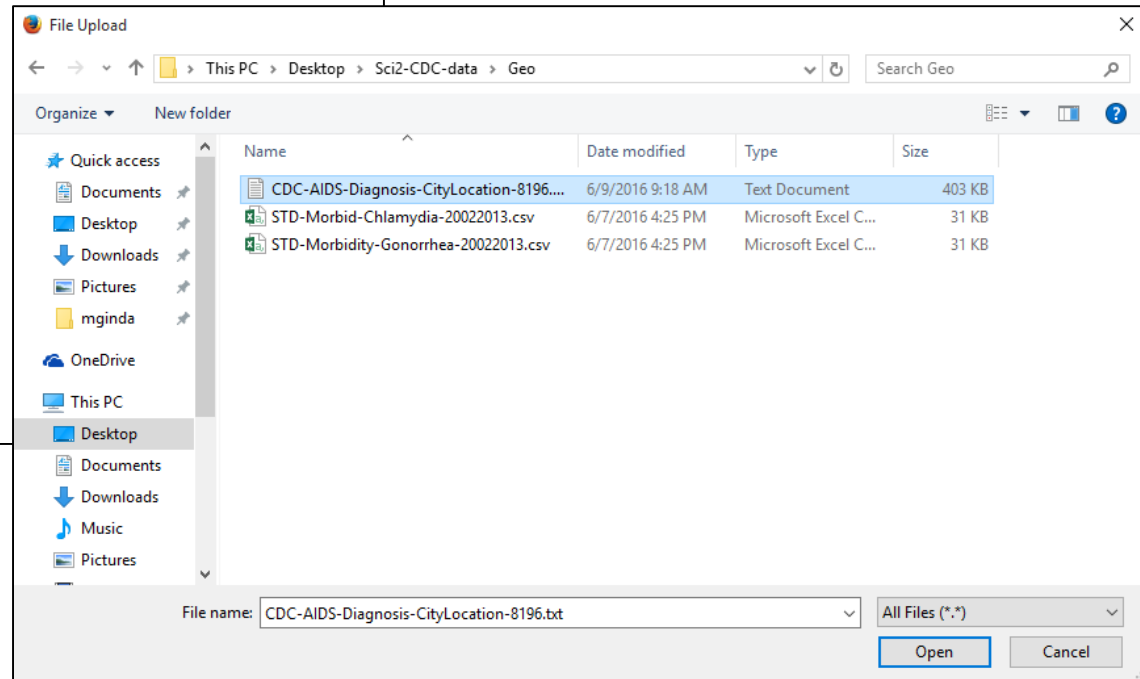
In the directory, select the executable file “google-refine.exe”.

A console window will appear, this may be ignored for now, but can indicate error messages and process log.

A new browser window will appear at the default domain 127.0.0.1:3333/3334



Once OpenRefine has started. You will want to create a project. First, you need to select your data source, which in this case is a sample data file on this computer,
“CDC-AIDS-Diagnosis-CityLocation-8196.txt”
 Once the data file is uploaded, hit next.



Once the data is loaded, it has to be parsed by OpenRefine. In this case, the data is in TSV format, and has column headers for the data.

The base setting for a TSV file are going to be all that is required. In the right hand corner, select **“Create Project”**

There are a number of parsing options beyond TSV/CSV.

Google refine A power tool for working with messy data.

Create Project « Start Over Configure Parsing Options Project name CDC AIDS Diagnosis CityLocation 8196 Create Project »


Notes	Location	Location Code	Year Diagnosed	Year Diagnosed Code	Race or Ethnicity	Race or Ethnicity Code	Cases
1.	Akron, OH	80	1982	1982	Black (and also not Hispanic)	2054-5	1
2.	Akron, OH	80	1984	1984	White (and also not Hispanic)	2106-3	3
3.	Akron, OH	80	1985	1985	White (and also not Hispanic)	2106-3	4
4.	Akron, OH	80	1986	1986	Black (and also not Hispanic)	2054-5	3
5.	Akron, OH	80	1986	1986	White (and also not Hispanic)	2106-3	9
6.	Akron, OH	80	1987	1987	Black (and also not Hispanic)	2054-5	6
7.	Akron, OH	80	1987	1987	White (and also not Hispanic)	2106-3	14
8.	Akron, OH	80	1988	1988	Black (and also not Hispanic)	2054-5	11
9.	Akron, OH	80	1988	1988	Hispanic	2135-2	1
10.	Akron, OH	80	1988	1988	White (and also not Hispanic)	2106-3	15
11.	Akron, OH	80	1989	1989	Black (and also not Hispanic)	2054-5	7
12.	Akron, OH	80	1989	1989	White (and also not Hispanic)	2106-3	19
13.	Akron, OH	80	1990	1990	Black (and also not Hispanic)	2054-5	6
14.	Akron, OH	80	1990	1990	White (and also not Hispanic)	2106-3	23
15.	Akron, OH	80	1991	1991	Black (and also not Hispanic)	2054-5	11
16.	Akron, OH	80	1991	1991	White (and also not Hispanic)	2106-3	38
17.	Akron, OH	80	1992	1992	Black (and also not Hispanic)	2054-5	17
18.	Akron, OH	80	1992	1992	White (and also not Hispanic)	2106-3	41
19.	Akron, OH	80	1993	1993	Black (and also not Hispanic)	2054-5	28
20.	Akron, OH	80	1993	1993	White (and also not Hispanic)	2106-3	37
21.	Akron, OH	80	1994	1994	Black (and also not Hispanic)	2054-5	15
22.	Akron, OH	80	1994	1994	Hispanic	2135-2	1
23.	Akron, OH	80	1994	1994	White (and also not Hispanic)	2106-3	31
24.	Akron, OH	80	1995	1995	Black (and also not Hispanic)	2054-5	27
25.	Akron, OH	80	1995	1995	Hispanic	2135-2	1
26.	Akron, OH	80	1995	1995	White (and also not Hispanic)	2106-3	32
27.	Akron, OH	80	1996	1996	Black (and also not Hispanic)	2054-5	17

Google refine CDC AIDS Diagnosis CityLocation 8196 txt [Permalink](#) Open... Export ▾ Help

Facet / Filter Undo / Redo ◂◃ Extensions: Freebase ▾

5311 rows Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last »

All	Notes	Location	Location Code	Year Diagnosed	Year Diagnosed Code	Race or Ethnicity	Race or Ethnicity Code
☆	1.	Akron, OH	80	1982	1982	Black (and also not Hispanic)	2054-5
☆	2.	Akron, OH	80	1984	1984	White (and also not Hispanic)	2106-3
☆	3.	Akron, OH	80	1985	1985	White (and also not Hispanic)	2106-3
☆	4.	Akron, OH	80	1986	1986	Black (and also not Hispanic)	2054-5
☆	5.	Akron, OH	80	1986	1986	White (and also not Hispanic)	2106-3
☆	6.	Akron, OH	80	1987	1987	Black (and also not Hispanic)	2054-5
☆	7.	Akron, OH	80	1987	1987	White (and also not Hispanic)	2106-3
☆	8.	Akron, OH	80	1988	1988	Black (and also not Hispanic)	2054-5
☆	9.	Akron, OH	80	1988	1988	Hispanic	2135-2
☆	10.	Akron, OH	80	1988	1988	White (and also not Hispanic)	2106-3

Using facets and filters 

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started? [Watch these screencasts](#)

Version 2.5 [r240] [Help](#) [About](#)

Google refine CDC AIDS Diagnosis CityLocation 8196 txt Permalink

Facet / Filter Undo / Redo 0 **5311 rows**

Show as: rows records Show: 5 10 25 50 rows

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started? [Watch these screencasts](#)

Facet / Filter Undo / Redo 0

Refresh Reset All Remove All

Notes change invert reset

53 choices Sort by: name count Cluster

Suggested Citation: US Department of Health and Human Services (US DHHS), Centers for Disease Control and Prevention (CDC), -1

Title: -1

Vital Status: All -1
with less than 500,000 population, in a non-metropolitan area, or whose metropolitan area of residence is unknown, and for all -1

(blank) -5253 include

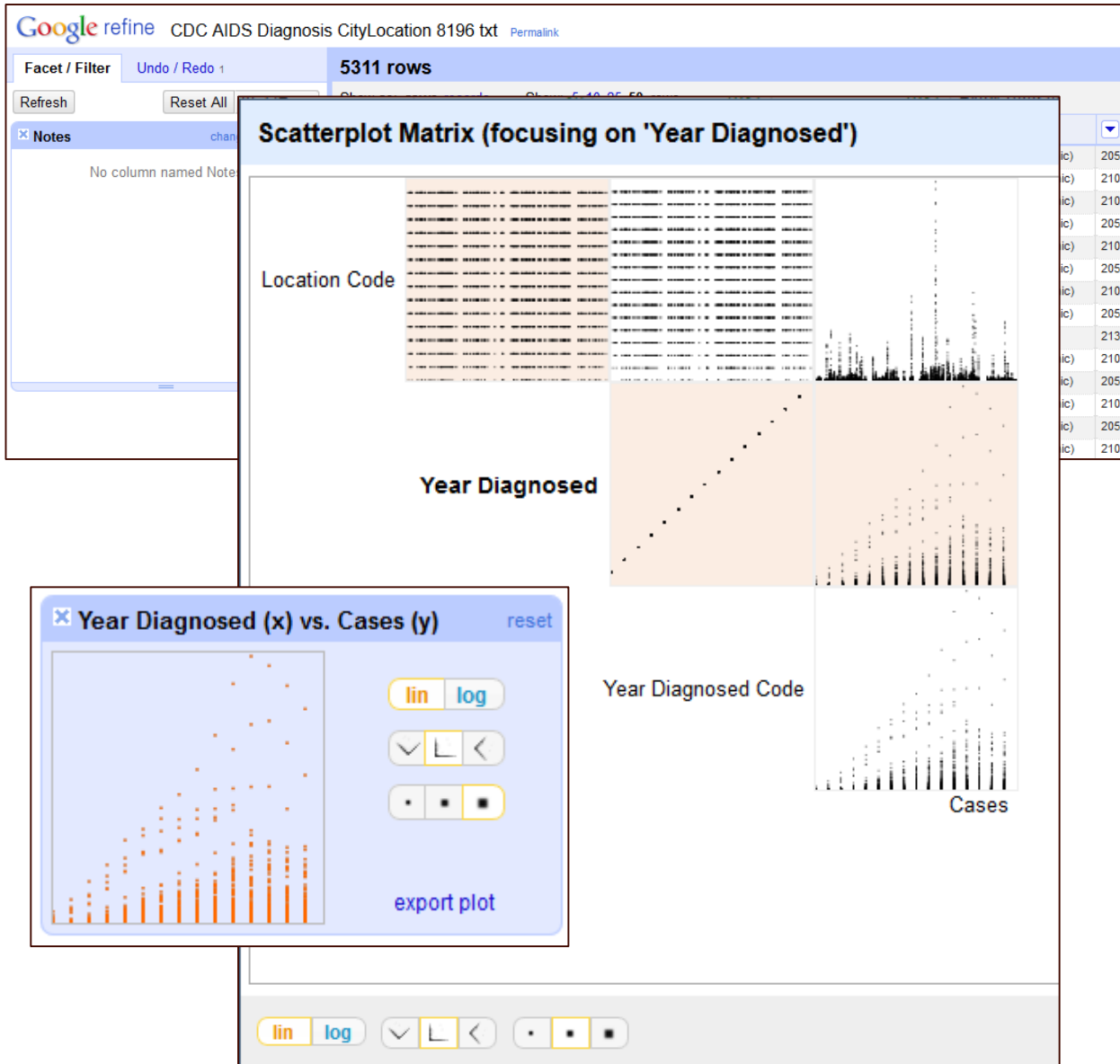
Facet by choice counts

	Notes	Location	Location Code	Year Diagnosed	Year
1.	Facet				1982
2.	Text filter				1984
3.	Edit cells				1985
4.	Edit column				1986
5.	Transpose				1986
6.	Sort				1987
7.					1987
					1988
					1988
					1988
					80
					1988

The first task is to review the fields in the data set. The “notes” column is a good place to start, as it looks empty.

From the “Notes” drop down menu, navigate to the “Text Facet” option. A box will appear in the Facet/Filter bar on the left side of the screen.

In the Notes filter, select “**(blank)**”, and then in the upper right corner of the filter select “**invert**”, which removes any row without text in the “Notes” column. This lets us review the remaining text.



Next, let's review the **“Year Diagnosed”** columns. It appears that the both copies of the column are duplicated, but to find out we can use the **“Scatter facet”** to look at the relationships between numeric variables.

The scatter plots show the plots related to **“Year diagnosed”** column are highlighted. As are other variables.

You may scale the plots, pivot on the axis, and adjust the point size.

Google refine CDC AIDS Diagnosis CityLocation 8196 txt [Permalink](#)

Facet / Filter Undo / Redo 0 **5253 matching rows** (5311 total)

Refresh Reset All Remove All Show as: rows records Show: 5 10 25 50 rows

Notes change invert reset 53 choices Sort by: name count Cluster

Suggested Citation: US Department of Health and Human Services (US DHHS), Centers for Disease Control and Prevention (CDC), 1

Title: Vital with in a metr and f (blar Facet

All	Notes	Location	Location Code	Year Diagnosed	Year Diagnosed	Race or Ethnicity	Race o
1.	Facet		80	1982		Black (and also not Hispanic)	2054-5
2.	Text filter		80	1984		White (and also not Hispanic)	2106-3
3.			80	1985		White (and also not Hispanic)	2106-3
4.	Edit cells		80	1986		Black (and also not Hispanic)	2054-5
5.	Edit column						
6.			80	1987			
7.			80	1987			
8.			80	1988			
9.			80	1988			
10.			80	1988			
11.			80	1989	1989	Black (and also not Hispanic)	2054-5
12.			80	1989	1989	White (and also not Hispanic)	2106-3
13.			80	1990	1990	Black (and also not Hispanic)	2054-5
14.			80	1990	1990	White (and also not Hispanic)	2106-3
15.			80	1991	1991	Black (and also not Hispanic)	2054-5
16.			80	1991	1991	White (and also not Hispanic)	2106-3
17.			80	1992	1992	Black (and also not Hispanic)	2054-5
18.			80	1992	1992	White (and also not Hispanic)	2106-3

5253 matching rows (5311 total)

Show as: rows records Show: 5 10 25 50 rows

All	Location	Location Code	Year Diagnosed	Year Diagnosed	Race or Ethnicity	Race o	
1.	Akron, OH		80	1982		Black (and also not Hispanic)	2054-5
2.	Akron, OH		80	1984		White (and also not Hispanic)	2106-3
3.	Akron, OH		80	1985		White (and also not Hispanic)	2106-3
4.	Akron, OH		80	1986		Black (and also not Hispanic)	2054-5
5.	Akron, OH		80	1986			
6.	Akron, OH		80	1987			
7.	Akron, OH		80	1987			
8.	Akron, OH		80	1988			
9.	Akron, OH		80	1988			
10.	Akron, OH		80	1988			
11.	Akron, OH		80	1989	1989	Black (and also not Hispanic)	2054-5
12.	Akron, OH		80	1989	1989	White (and also not Hispanic)	2106-3
13.	Akron, OH		80	1990	1990	Black (and also not Hispanic)	2054-5
14.	Akron, OH		80	1990	1990	White (and also not Hispanic)	2106-3
15.	Akron, OH		80	1991	1991	Black (and also not Hispanic)	2054-5
16.	Akron, OH		80	1991	1991	White (and also not Hispanic)	2106-3
17.	Akron, OH		80	1992	1992	Black (and also not Hispanic)	2054-5
18.	Akron, OH		80	1992	1992	White (and also not Hispanic)	2106-3

Facet

- Facet
- Text filter
- Edit cells
- Edit column**
 - Split into several columns...
 - Transpose
 - Sort...
 - View
 - Reconcile
- Add column based on this column...
- Add column by fetching URLs...
- Add columns from Freebase ...
- Rename this column
- Remove this column**
- Move column to beginning
- Move column to end
- Move column left
- Move column right

After reviewing the “Notes” and “Years diagnosed” columns, we need to remove it from the data.

From the drop down menu, select **Edit Columns > Remove this column.**

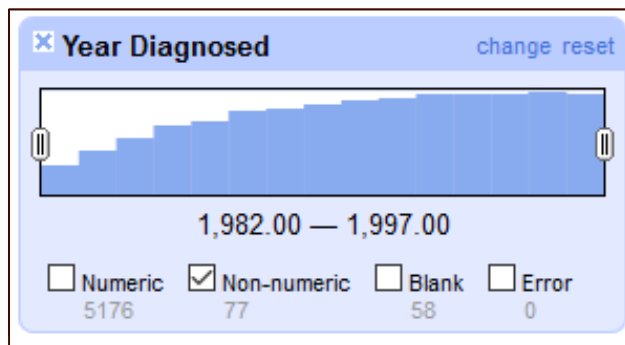
5311 rows

Show as: rows records Show: 5 10 25 50 rows

All	Location	Location Code	Year Diagnosed	Year Diagnosed	Race or Ethnicity
☆	1. Akron, OH	80	Facet	Text facet	not Hispanic)
☆	2. Akron, OH	80	Text filter	Numeric facet	not Hispanic)
☆	3. Akron, OH	80	Edit cells	Timeline facet	not Hispanic)
☆	4. Akron, OH	80	Edit column	Scatterplot facet	not Hispanic)
☆	5. Akron, OH	80	Transpose	Custom text facet...	not Hispanic)
☆	6. Akron, OH	80	Sort...	Custom numeric facet...	not Hispanic)
☆	7. Akron, OH	80	View	Customized facets	not Hispanic)
☆	8. Akron, OH	80	Reconcile	1988	White (and also not Hispanic)
☆	9. Akron, OH	80		1989	1989 Black (and also not Hispanic)
☆	10. Akron, OH	80		1989	1989 White (and also not Hispanic)
☆	11. Akron, OH	80			
☆	12. Akron, OH	80			

With large data sets it helps to know the type of data in all fields. The “Year Diagnosed” column is a good case to review because it is a categorical field that has been parsed by OpenRefine as a numeric.

From the drop down menu, select **Facet > Numeric Facet**. A new box appears in the left side that shows the distribution of records based on the numeric values, as well as check boxes for errors, blanks and non-numeric values.



0 rows

code	Year Diagnosed	Year D
520	Before 1982	edit
520	Before 1982	Edit this cell
680	Before 1982	
720	Before 1982	
720	Before 1982	

Uncheck all but “non-numeric”.

77 matching rows (5311 total)

Show as: **rows** records Show: 5 10 25 50 rows

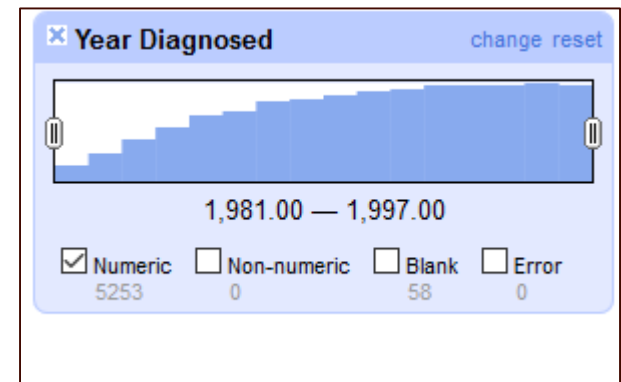
All	Location	Location Code	Year Diagnosed	Year Diagnosed	Race or Ethnicity	Race or Ethnicit	Cases
☆	193. Atlanta, GA	520	Before 1982				3
☆	194. Atlanta, GA	520	Before 1982				6
☆	300. Bakersfield, CA	680	Before 1982				2
☆	349. Baltimore, MD	720	Before 1982				1
☆	350. Baltimore, MD	720	Before 1982				1
☆	443. Bergen-Passaic, NJ	875	Before 1982				1
☆	444. Bergen-Passaic, NJ	875	Before 1982	1981	Black (and also not Hispanic)	2054-5	1
☆	445. Bergen-Passaic, NJ	875	Before 1982	1981	White (and also not Hispanic)	2106-3	4

Data type: **number** ▼
 1981|
 Apply Apply to All Identical Cells Cancel
Enter Ctrl-Enter Esc

To update the value “Before 1982”, take your mouse and move over the a cell in the “**Year Diagnosed**” column.

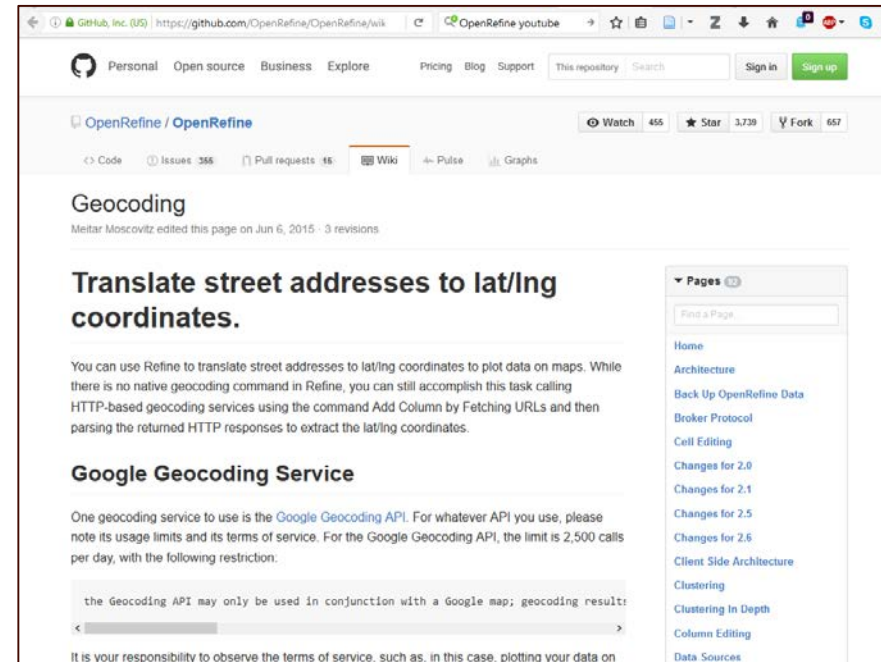
Select the data type “number”, and update the text to “1981”. Afterwards, select “Apply to All Identical Cells”, which should leave the table blank.

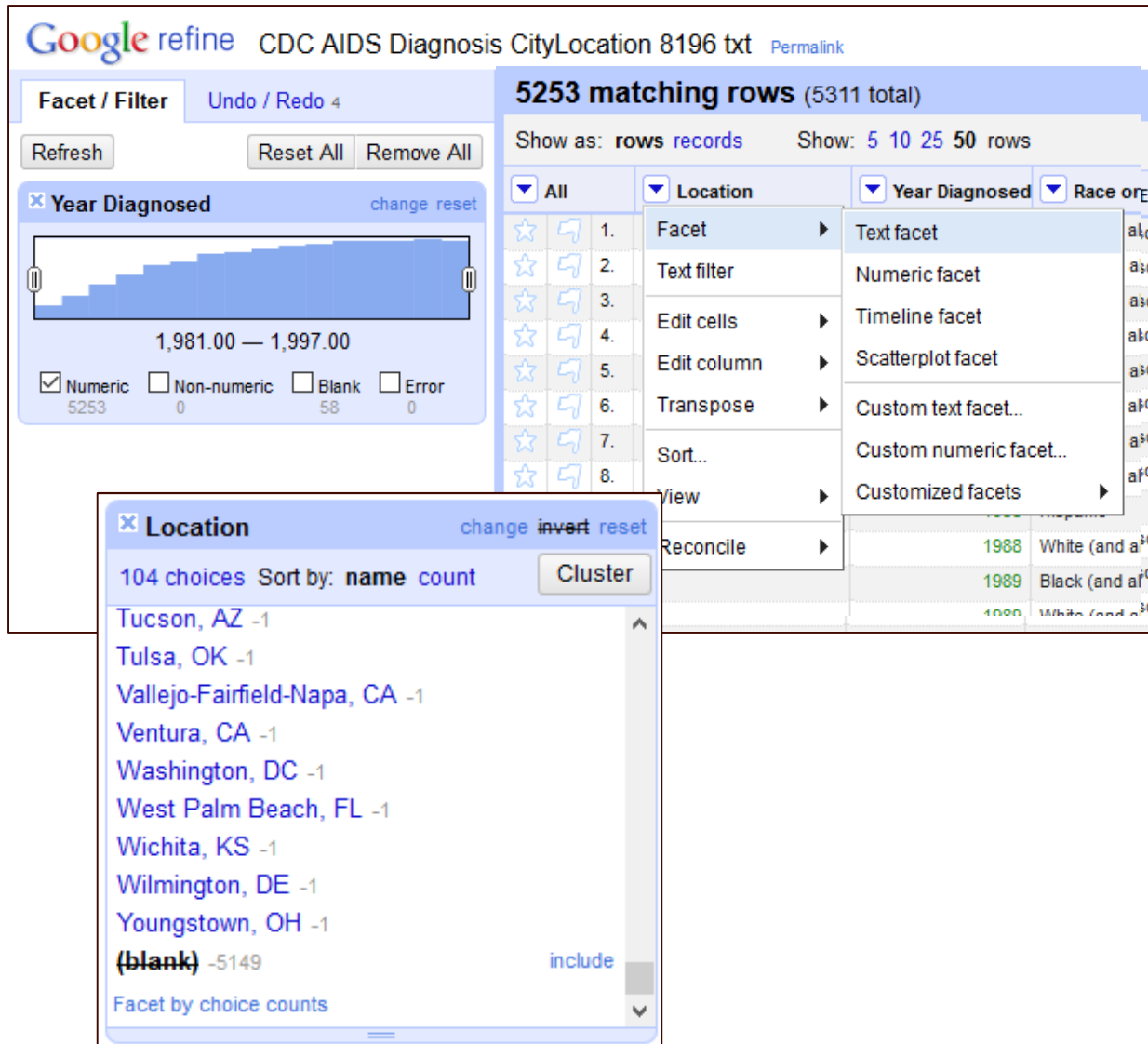
To see values, in the “Year Diagnosed” numeric filter, move the check from “Non-numeric” to “Numeric.” Note: “Blank” rows are leftover from the notes field. Keep this filter open.



After updating the year values, we will start the process of geocoding the data using Google's Geocoding API. The process takes three main steps: **fetching** the data to the geocoding API, **parsing** the data, and **cleaning** up the data for use.

This tutorial follows the documentation created by Moscovitz, M., & Morris, T. (2015). Geocoding · OpenRefine/OpenRefine Wiki · GitHub. Retrieved May 16, 2016, from <https://github.com/OpenRefine/OpenRefine/wiki/Geocoding>





Google refine CDC AIDS Diagnosis CityLocation 8196 txt [Permalink](#)

Facet / Filter Undo / Redo 4

Refresh Reset All Remove All

5253 matching rows (5311 total)

Show as: rows records Show: 5 10 25 50 rows

Year Diagnosed 1,981.00 — 1,997.00

Numeric 5253 Non-numeric 0 Blank 58 Error 0

Location 104 choices Sort by: name count

Tucson, AZ -1
Tulsa, OK -1
Vallejo-Fairfield-Napa, CA -1
Ventura, CA -1
Washington, DC -1
West Palm Beach, FL -1
Wichita, KS -1
Wilmington, DE -1
Youngstown, OH -1
(blank) -5149

Facet by choice counts

1. Facet
2. Text filter
3. Edit cells
4. Edit column
5. Transpose
6. Sort...
7. view
8. Reconcile

Text facet
Numeric facet
Timeline facet
Scatterplot facet
Custom text facet...
Custom numeric facet...
Customized facets

1988 White (and a...
1989 Black (and a...
1990 White (and a...)

Users are limited to 2500 records per day. To avoid the data fetch limits, we must reduce the location records to the minimal amount of geolocations.

First we need to limit the number of geolocations. First in the drop down menu for “**Locations**” select **Edit column > Blank down**. Then apply a **Text facet**, and select and invert all “(blank)” values.

This creates a final list of unique geolocations.

Add column by fetching URLs based on column Location

New column name Throttle delay milliseconds

On error set to blank store error

Formulate the URLs to fetch:

Expression Language No syntax error.

Preview History Starred Help

row	value	
		"http://maps.google.com/maps/api/geocode/json?sensor=false&address=" + escape(value, "url")
1.	Akron, OH	http://maps.google.com/maps/api/geocode/json?sensor=false&address=Akron%2C+OH
29.	Albany-Schenectady, NY	http://maps.google.com/maps/api/geocode/json?sensor=false&address=Albany-Schenectady%2C+NY
75.	Albuquerque, N.M.	http://maps.google.com/maps/api/geocode/json?sensor=false&address=Albuquerque%2C+N.M.
121.	Allentown, PA	http://maps.google.com/maps/api/geocode/json?sensor=false&

Create column Geocode at index 1 by fetching URLs based on column Location using expression
 grel:"http://maps.google.com/maps/api/geocode/json?sensor=false&address=" + escape(value, "url")
 14% complete Cancel

Fetching

Set up Locations column to fetch geocode data by selecting in the drop down menu **Edit column > Add column by fetching url**, which opens a new window.

Next name the new column **"Geocode"** and the Throttle Delay parameter to **"200"** milliseconds.

Then add the GREL expression:

```
"http://maps.google.com/maps/api/geocode/json?sensor=false&address=" + escape(value, "url")
```

Add column based on column Geocode

New column name

On error set to blank store error copy value from original column

Expression Language Google Refine Expression Language (GREL) ▾

No syntax error.

Preview History Starred Help

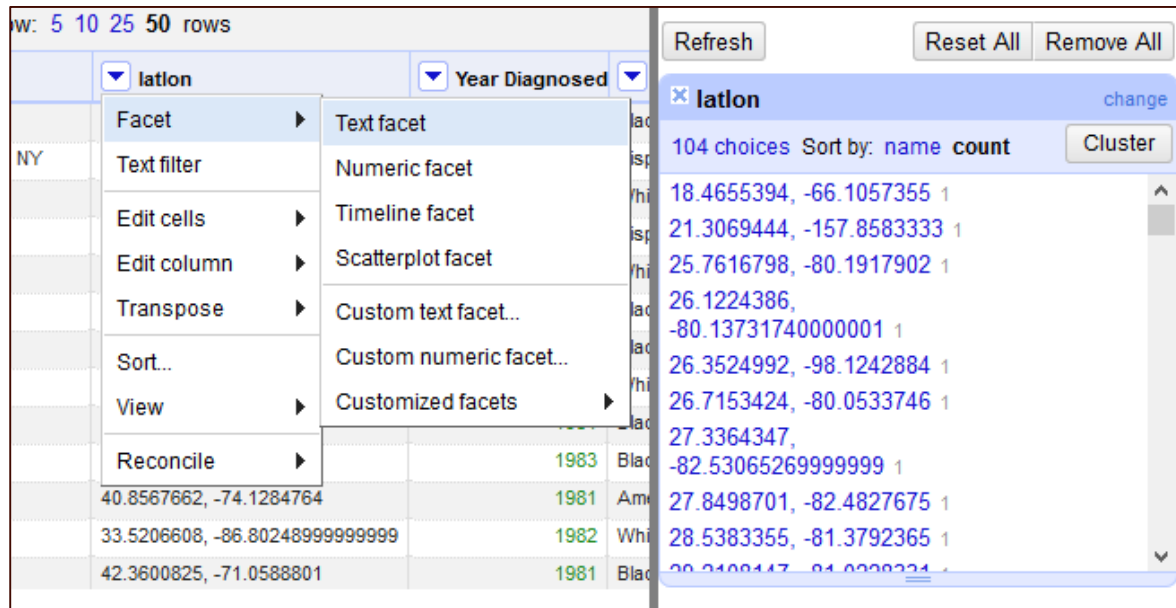
row	value	with(value.parseJson().results[0].geometry.location, pair, pair.lat +\", \" + pair.lng)
1.	{ "results": [{ "address_components": [{ "long_name": "Akron", "short_name": "Akron", "types": ["locality", "political"] }, { "long_name": "Summit County", "short_name": "Summit County", "types": ["administrative_area_level_2", "political"] }, { "long_name": "Ohio", "short_name": "OH", "types": [

Parsing

The Google Geocoding API returns data as in a JSON format, stored as text in the cell for each row. This requires that the data be parsed using the column's dropdown menu **Edit Column > Add Column based on This Column...**

The new column name is **"latlon"**, and the GREL expression for parsing is:

```
with(value.parseJson(
).results[0].geometry
.location, pair,
pair.lat +\", \" +
pair.lng)
```



Open Refine interface showing a data table with a context menu open over the 'latlon' column. The menu options include Facet, Text filter, Edit cells, Edit column, Transpose, Sort..., View, and Reconcile. The 'Edit column' option is selected, and a sub-menu is visible with options like Text facet, Numeric facet, Timeline facet, Scatterplot facet, Custom text facet..., Custom numeric facet..., and Customized facets. The 'Text facet' option is highlighted. To the right, a facet panel for 'latlon' is shown with 104 choices, sorted by name count, and a 'Cluster' button.

After parsing, we need to remove the **Geocode** column from the drop down menu select **Edit column > Remove column**

Next, we need to review the results to determine if there are any duplicate pairs of latitude and longitude values, blank values, or error codes. A **Text Facet** on the “**latlon**” column will allow us to see if this is the case.

If there are duplicates values in the filter, a secondary cleaning process that adjust/jitters values may be appropriate, or you may want to go back to your original data and clean the location names using Open Refine’s [Clustering and Editing algorithms](#) for text data.

104 matching rows (5311 total)

Show as: rows records Show: 5 10 25 50 rows

All	Location	latlon 1	latlon 2	Year Diagnosed	Race or Et
1.	Akron, OH	Facet	.5190052999999	1982	Black (and also
29.	Albany-Schenectady, NY	Text filter	-73.7562317	1983	Hispanic
75.	Albuquerque, N.M.	Edit cells	-106.6055534	1983	White (and also
121.	Allentown, PA	Edit column	-75.4901833	1982	Hispanic
161.	Ann Arbor, MI	Split into several columns...			nd also
193.	Atlanta, GA	Transpose			nd also
251.	Austin, TX	Add column based on this column...			nd also
300.	Bakersfield, CA	Add column by fetching URLs...			nd also
349.	Baltimore, MD	Add columns from Freebase ...			nd also
408.	Baton Rouge, LA	Rename this column			nd also
443.	Bergen-Passaic, NJ	Remove this column	40.8567662		n Indian
505.	Birmingham, AL	Move column to beginning	33.5206608	-8	nd also
536.	Boston, MA	Move column to end	42.3600825		nd also
608.	Buffalo, NY	Move column left	42.8864467999999		nd also
663.	Charleston, SC	Move column right	32.7764749	-7	nd also
699.	Charlotte, NC		35.2270869		nd also
747.	Chicago, IL		41.8781136		nd also

Enter new column name

OK Cancel

After determining if there are any duplicate locations, we need to split the **latlon** column into two, and then rename them

First, from the **latlon** column's drop down menu, select **Edit columns > Split into several columns...**

None of the parameters in the pop up box need to be updated. Select "OK", and fields "latlon1" and "latlon2" are created, while the initial column is deleted.

Next, for **latlon1** select **Edit columns > Rename this column** and enter the value "**Latitude**". Repeat for **latlon2** with the value "**Longitude**".

Location change

104 choices Sort by: name count Cluster

Tucson, AZ 1
Tulsa, OK 1
Vallejo-Fairfield-Napa, CA 1
Ventura, CA 1
Washington, DC 1
West Palm Beach, FL 1
Wichita, KS 1
Wilmington, DE 1
Youngstown, OH 1
(blank) 5149

Facet by choice counts

Facet / Filter Undo / Redo 11

Refresh Reset All Remove All

latlon change

No column named latlon

Year Diagnosed change reset

1,981.00 — 1,997.00

Numeric 5253 Non-numeric 0 Blank 58 Error 0

Location change

104 choices Sort by: name count Cluster

Tucson, AZ 1
Tulsa, OK 1
Vallejo-Fairfield-Napa, CA 1
Ventura, CA 1
Washington, DC 1
West Palm Beach, FL 1
Wichita, KS 1
Wilmington, DE 1
Youngstown, OH 1
(blank) 5149

Facet by choice counts

5253 matching rows (5311 total)

Show as: rows records Show: 5 10 25 50 rows

	All	Location	latitude	longitude	Year Diagnosed		
1.	Akron, OH	41.0814447	-81.51900529999999	1984	White (and also not Hispanic)	2106-3	3
2.				1985	White (and also not Hispanic)	2106-3	4
3.				1986	Black (and also not Hispanic)	2054-5	3
4.				1986	White (and also not Hispanic)	2106-3	9
5.				1987	Black (and also not Hispanic)	2054-5	6
6.				1987	White (and also not Hispanic)	2106-3	14
7.				1988	Black (and also not Hispanic)	2054-5	11
8.				1988	Hispanic	2135-2	1
9.				1988	White (and also not Hispanic)	2106-3	15
10.				1989	Black (and also not Hispanic)	2054-5	7
11.				1989	White (and also not Hispanic)	2106-3	19
12.				1990	Black (and also not Hispanic)	2054-5	6
13.				1990	White (and also not Hispanic)	2106-3	23
14.				1991	Black (and also not Hispanic)	2054-5	11
15.				1991	White (and also not Hispanic)	2106-3	38
16.				1992	Black (and also not Hispanic)	2054-5	17
17.				1992	White (and also not Hispanic)	2106-3	41
18.				1993	Black (and also not Hispanic)	2054-5	28
19.				1993	White (and also not Hispanic)	2106-3	37
20.				1994	Black (and also not Hispanic)	2054-5	15
21.				1994	Hispanic	2135-2	1
22.				1994	White (and also not Hispanic)	2106-3	31
23.				1995	Black (and also not Hispanic)	2054-5	27
24.				1995	Hispanic	2135-2	1
25.				1995	White (and also not Hispanic)	2106-3	32
26.				1996	Black (and also not Hispanic)	2054-5	17
27.				1996	White (and also not Hispanic)	2106-3	28
28.				1983	Hispanic	2135-2	2
29.	Albany-Schenectady, NY	42.6525793	-73.7562317	1983	White (and also not Hispanic)	2106-3	4
30.				1984	Black (and also not Hispanic)	2054-5	2
31.				1984	Hispanic	2135-2	3
32.				1984	White (and also not Hispanic)	2106-3	14
33.				1984	White (and also not Hispanic)	2106-3	14

After transforming the latitude and longitude data, we may now stop filtering the **Location** column, and show blank values.

The final data transformations, blank values for the locations, latitude and longitude values need to be replaced for all rows in the data set.

Navigate in the column's drop down menu to **Edit cells > Fill down**. Repeat for all three columns.

Google refine CDC AIDS Diagnosis CityLocation 8196 txt [Permalink](#)

Facet / Filter Undo / Redo 11

Refresh Reset All Remove All

latlon change

No column named latlon

Year Diagnosed change reset

5253 matching rows (5311 total)

Show as: rows records Show: 5 10 25 50 rows

	All	Location	latitude	longitude	Year Diagnosed	Race or Ethnicity	Race or Ethnicit	Cases
1.	Facet		41.0814447	-81.51900529999999	1982	Black (and also not Hispanic)	2054-5	1
2.	Text filter				1984	White (and also not Hispanic)	2106-3	3
3.	Edit cells				1985	White (and also not Hispanic)	2106-3	4
4.	Edit column				1986	Black (and also not Hispanic)	2054-5	3
5.	Transpose				1986	White (and also not Hispanic)	2106-3	9
6.	Sort...				1987	Black (and also not Hispanic)	2054-5	6
7.	View				1987	White (and also not Hispanic)	2106-3	14
8.	Reconcile				1988	Black (and also not Hispanic)	2054-5	11
9.					1988	Hispanic	2135-2	1
10.					1988	White (and also not Hispanic)	2106-3	15
11.					1989	Black (and also not Hispanic)	2054-5	7
12.					1989	White (and also not Hispanic)	2106-3	19
13.					1990	Black (and also not Hispanic)	2054-5	6
14.					1990	White (and also not Hispanic)	2106-3	23
15.					1991	Black (and also not Hispanic)	2054-5	11
16.					1991	White (and also not Hispanic)	2106-3	38



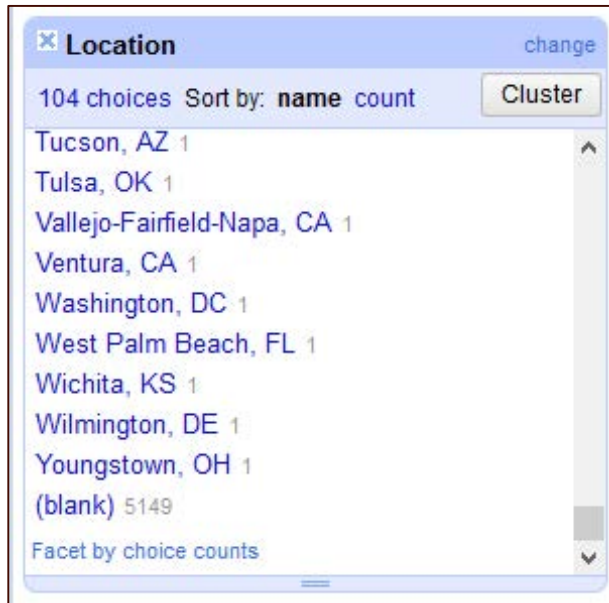
Race or Ethnicity 6 choices Sort by: name count Cluster

- American Indian /Alaskan Native 422
- Asian / Pacific Islander 551
- Black (and also not Hispanic) 1392
- Hispanic 1201
- Unknown 151
- White (and also not Hispanic) 1536

Facet by choice counts

We also may remove the **Race or Ethnicity code** column.

Finally, we want to filter the data set to leave only the values that we are interested. To help, we need to add one more text filter to the **Race or Ethnicity** field.

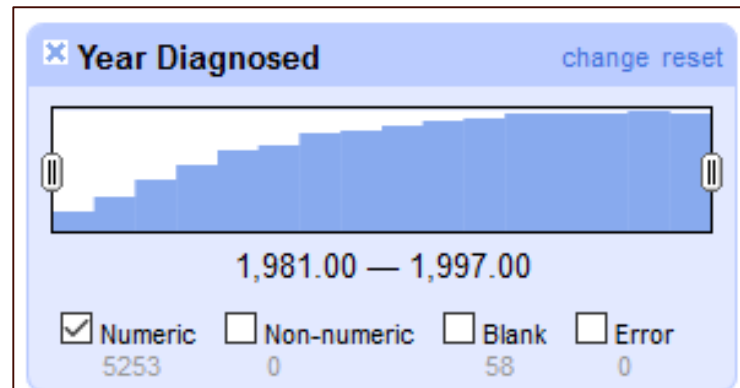


Location 104 choices Sort by: name count Cluster

- Tucson, AZ 1
- Tulsa, OK 1
- Vallejo-Fairfield-Napa, CA 1
- Ventura, CA 1
- Washington, DC 1
- West Palm Beach, FL 1
- Wichita, KS 1
- Wilmington, DE 1
- Youngstown, OH 1
- (blank) 5149

Facet by choice counts

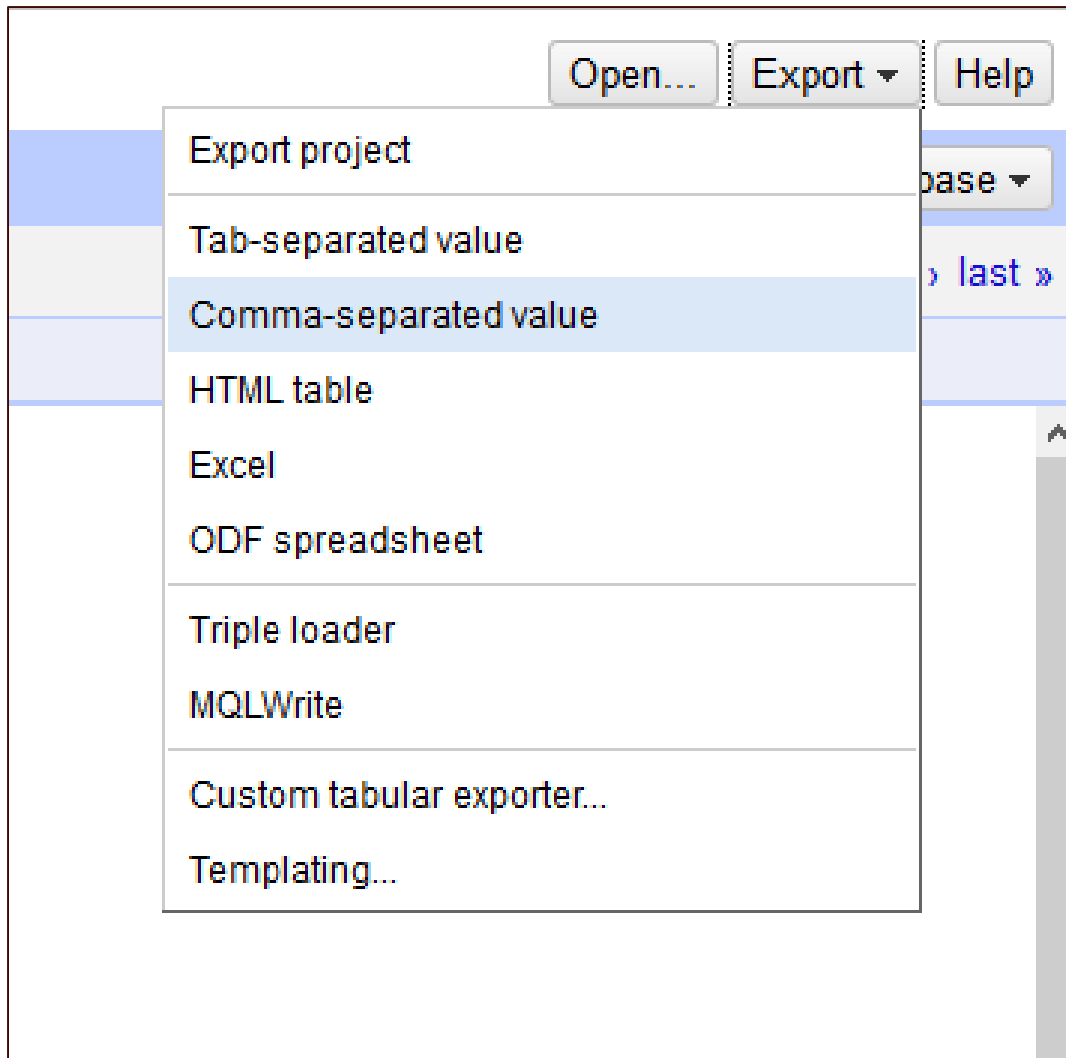
Now using these filters, select a date range and group to subset the data again.



Year Diagnosed change reset

1,981.00 — 1,997.00

Numeric 5253 Non-numeric 0 Blank 58 Error 0



Finally, with a selected data, it is time to export as a CSV formatted data table.

In the right hand corner of the browser window, select the **Export** drop down menu, and the format **Comma-separated value**.

8:00 Welcome and Overview of Tutorial and Attendees**8:30 Visualization Framework and Workflow Design**

- Overview of the Visualization framework
- Overview of Graphical variables, and color selection

9:15 Tool Overview and Trouble Shooting

- Sci2 Overview – scientometric analysis tool
- Open Refine – data parsing, transformation, and editing tool
- Gephi – network visualization

9:45 Geospatial Analysis

- Overview of geospatial analysis and mapping
- Geospatial Analysis: Geocoding with OpenRefine
- **Geospatial Analysis: Proportional Symbol Map using CDC**

11:00 Topical/Temporal Analysis: Burst Detection

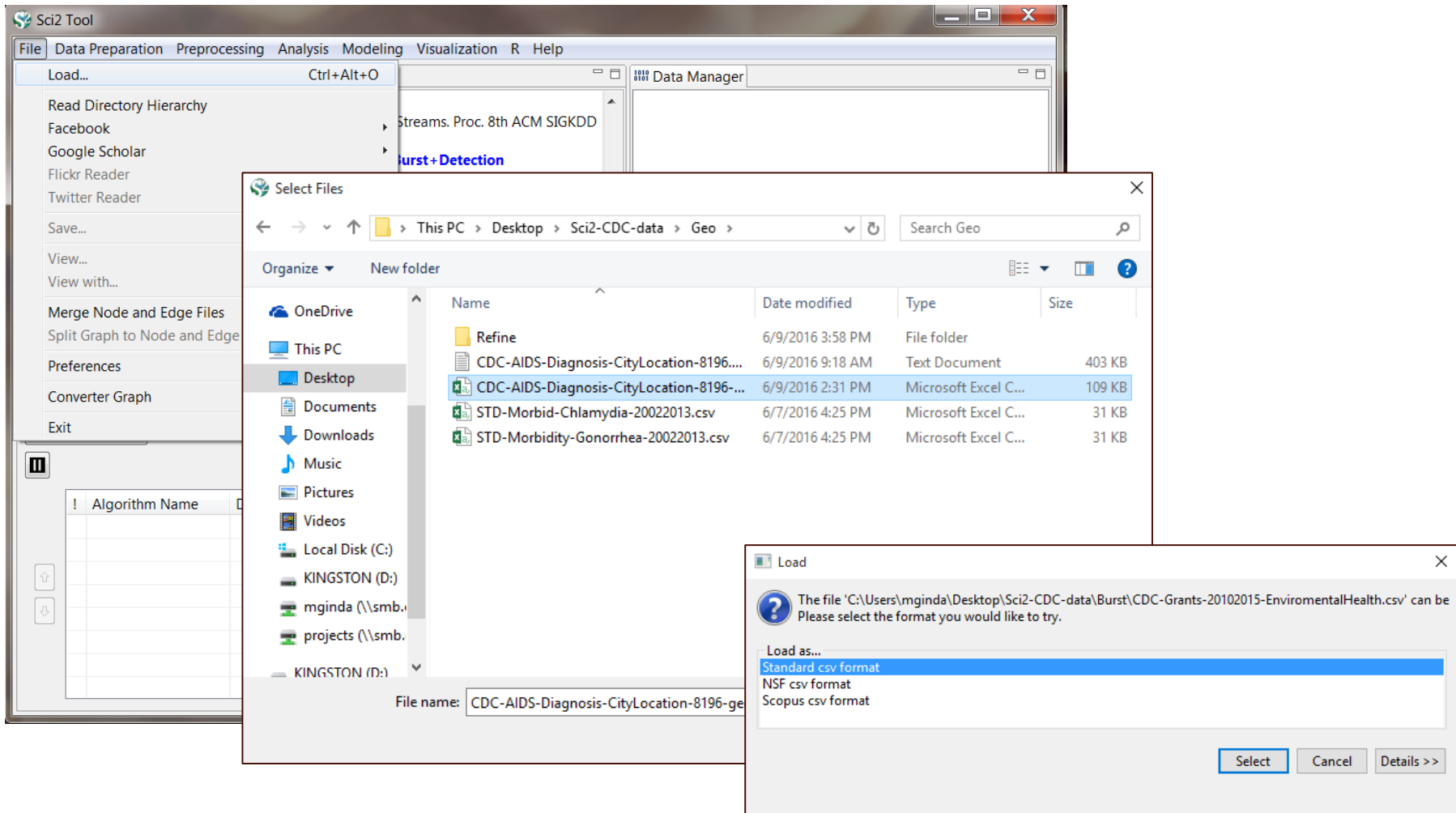
- Overview of burst analysis and introductory workflow
- Burst Detection with CDC Grants

12:30 Lunch**1:30 Network Analysis**

- Introduction to Network Analysis
- Network Analysis: Bimodal networks with Morbidity Data
- Network Analysis: Co-authorship Network with CDC publications

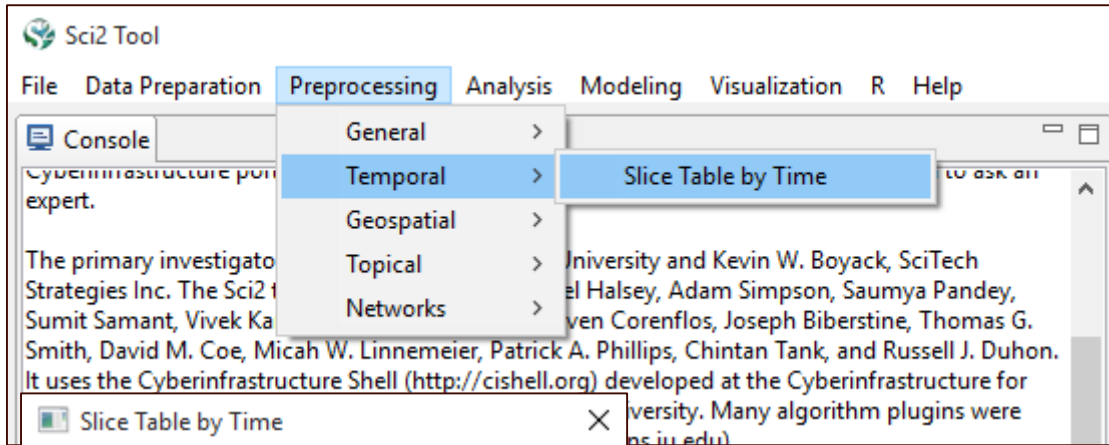
4:00 Wrap-up

4:30 Adjourn



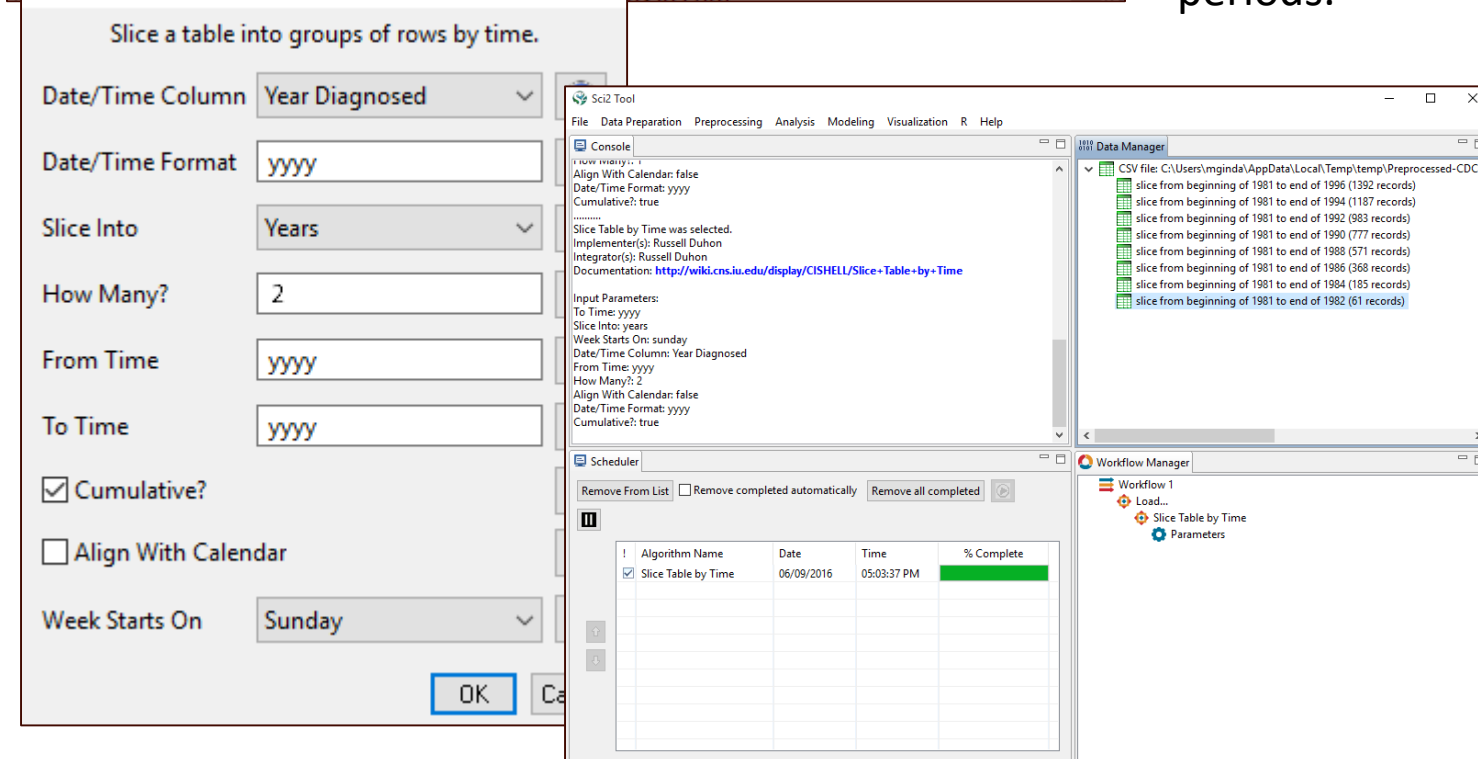
Load ***CDC-AIDS-Diagnosis-CityLocation-8196-geocoded-filtered-race-black.csv***

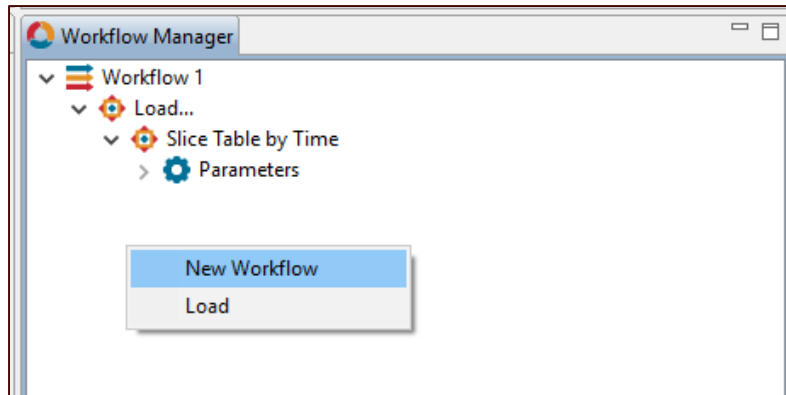
Located in CDC data directory: Sci2-CDC-data-> burst



Next, navigate **Preprocessing > Temporal > Slice Table by Time** and then enter the following parameters in the pop-up box.

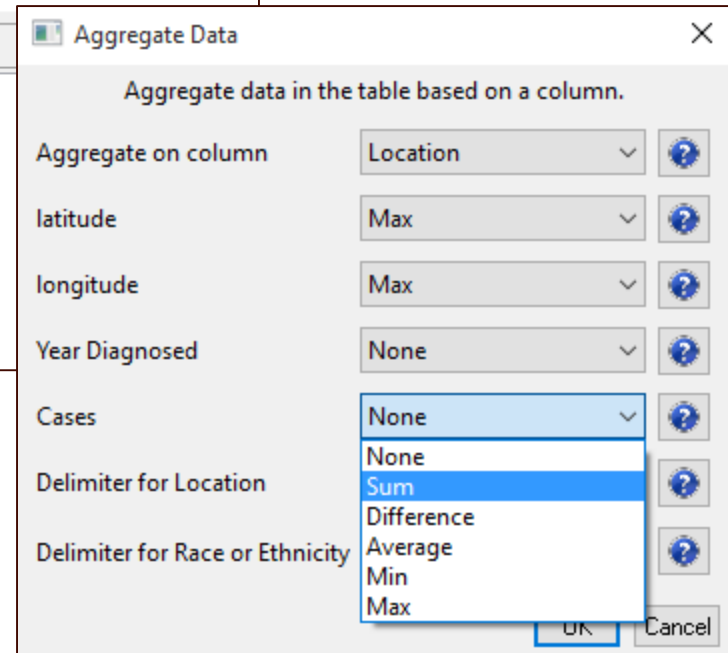
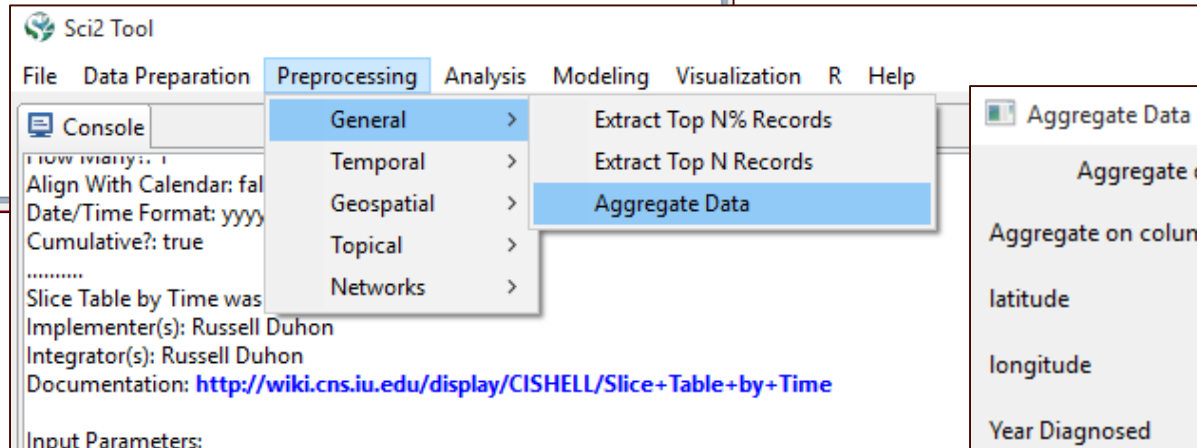
We will create a cumulative series of maps for two year periods.





Next, create a new workflow in the Workflow manager by right clicking the menu area and selecting “New Workflow”. This allows us to duplicate the next aggregation step.

Then navigate to **Preprocessing > General > Aggregate Data**.

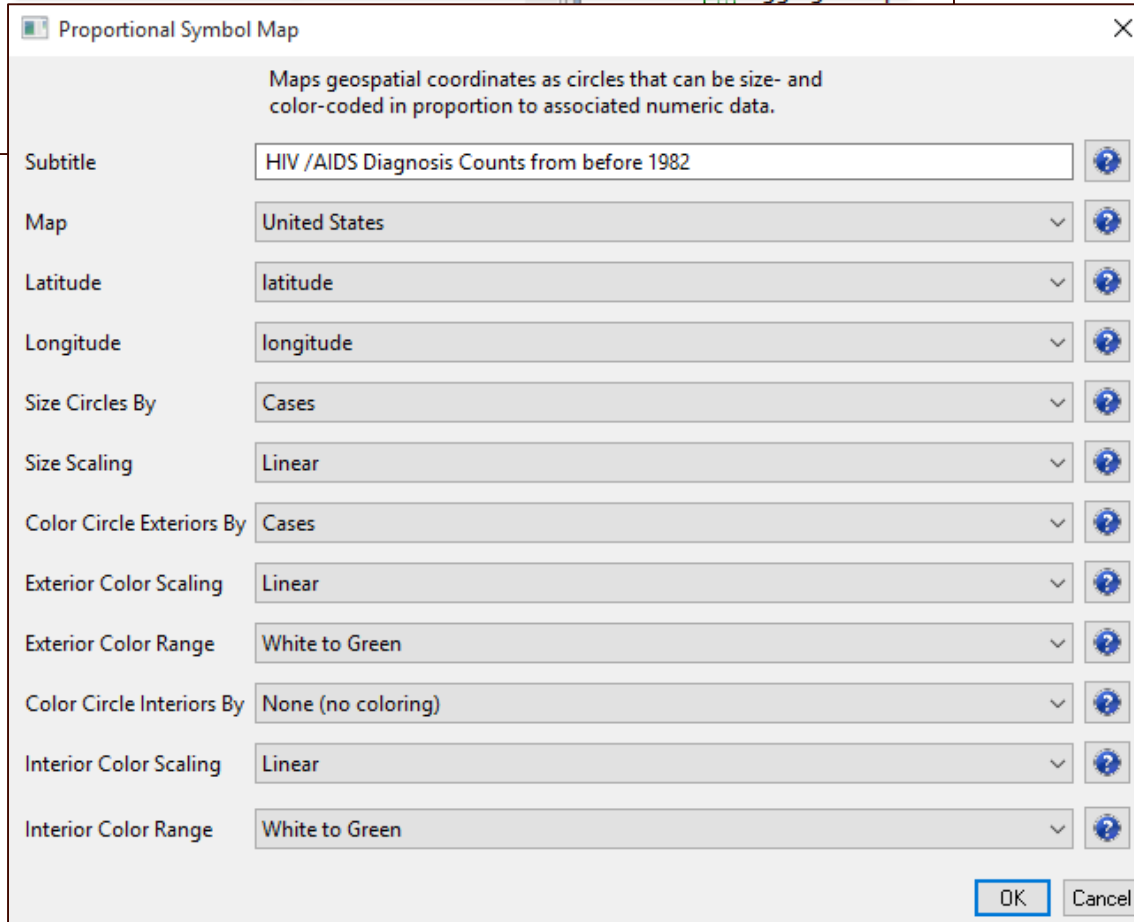
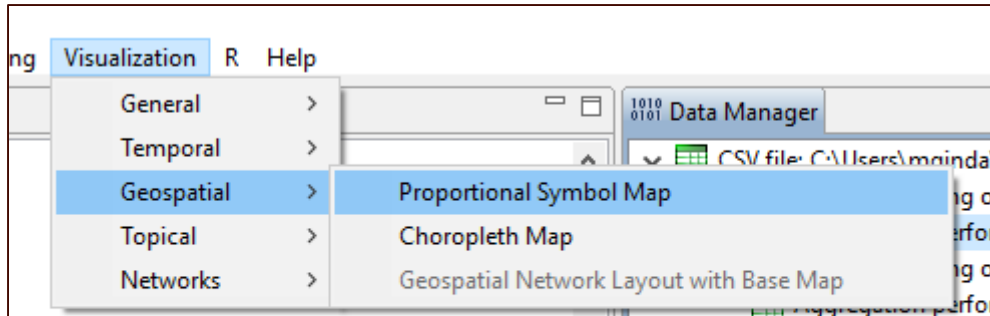


Set the following parameters for the Aggregate Data algorithm, and select “OK”:

	A	B	C	D	E
1	Location	latitude	longitude	Cases	Count
2	Riverside-S Berndino, CA	33.95335	-117.396	2	1
3	New York, NY	40.71278	-74.0059	200	2
4	Baltimore, MD	39.29038	-76.6122	3	2
5	Tampa-Saint Petersburg, FL	27.84987	-82.4828	2	2
6	El Paso, TX	31.76188	-106.485	1	1
7	San Francisco, CA	37.77493	-122.419	6	2
8	Bergen-Passaic, NJ	40.85677	-74.1285	4	2
9	Miami, FL	25.76168	-80.1918	40	2
10	Houston, TX	29.76043	-95.3698	2	2
11	New Orleans, LA	29.95107	-90.0715	6	2
12	Philadelphia, PA	39.95258	-75.1652	11	2
13	Chicago, IL	41.87811	-87.6298	8	2
14	Detroit, MI	42.33143	-83.0458	3	2
15	Gary, IN	41.59337	-87.3464	2	2
16	Los Angeles, CA	34.05223	-118.244	17	2
17	Washington, DC	38.90719	-77.0369	10	2
18	Raleigh-Durham, NC	35.89917	-78.8636	1	1
19	Newark, NJ	40.73566	-74.1724	34	2
20	Boston, MA	42.36008	-71.0589	8	2
21	Atlanta, GA	33.749	-84.388	9	2
22	Saint Louis, MO	38.627	-90.1994	2	2
23	Middlesex, NJ	40.5726	-74.4927	1	1
24	Akron, OH	41.08144	-81.519	1	1
25	San Antonio, TX	29.42412	-98.4936	1	1
26	Oklahoma City, OK	35.46756	-97.5164	1	1
27	Tulsa, OK	36.15398	-95.9928	1	1
28	Jersey City, NJ	40.72816	-74.0776	10	1
29	Colorado Springs, CO	38.83388	-104.821	1	1

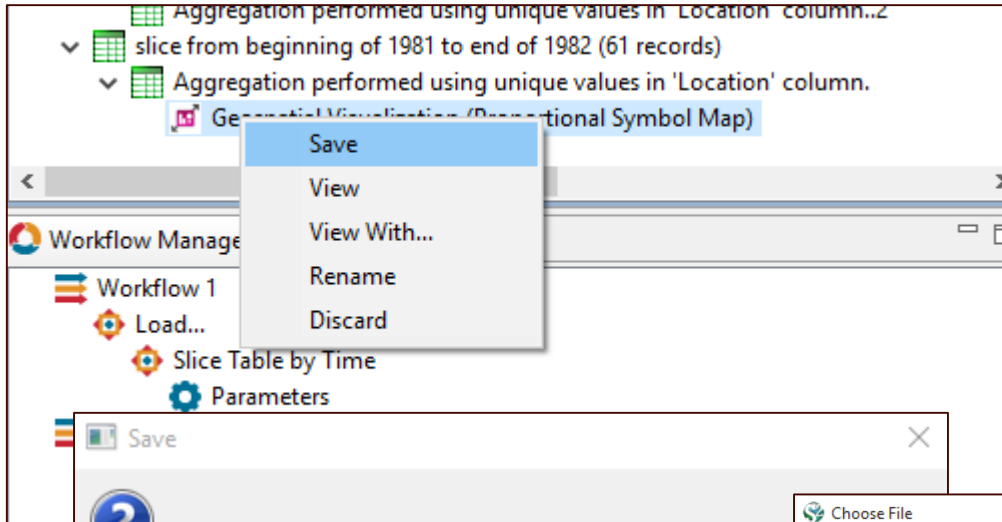
Aggregated data may be viewed by right clicking on the output file in the data manager, and selecting **View** or **View With**.

In Excel, the aggregated data for 1981-1982 looks like this.



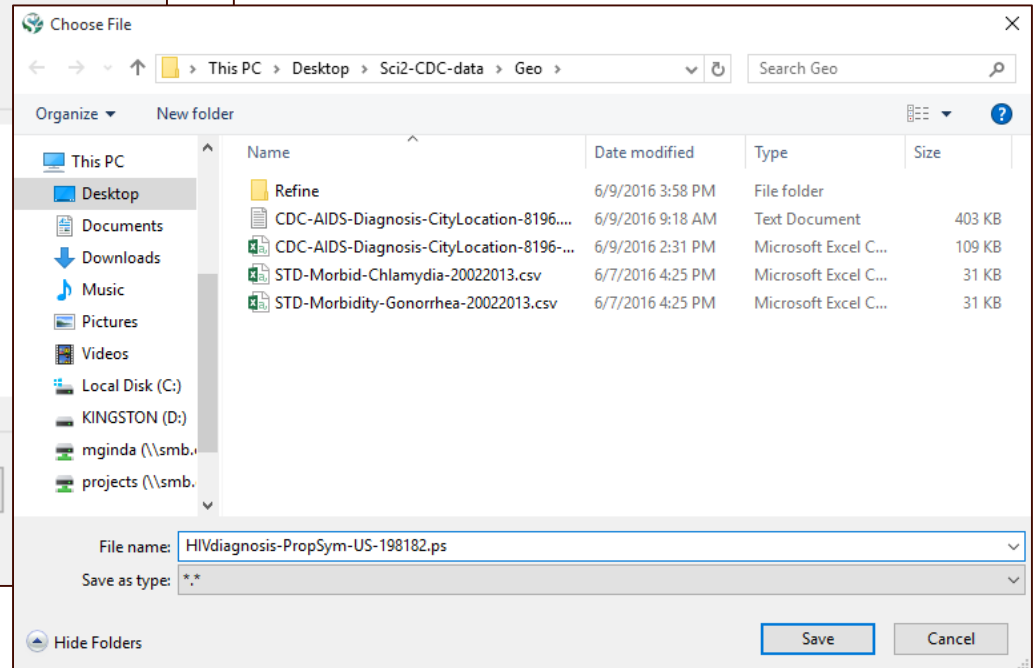
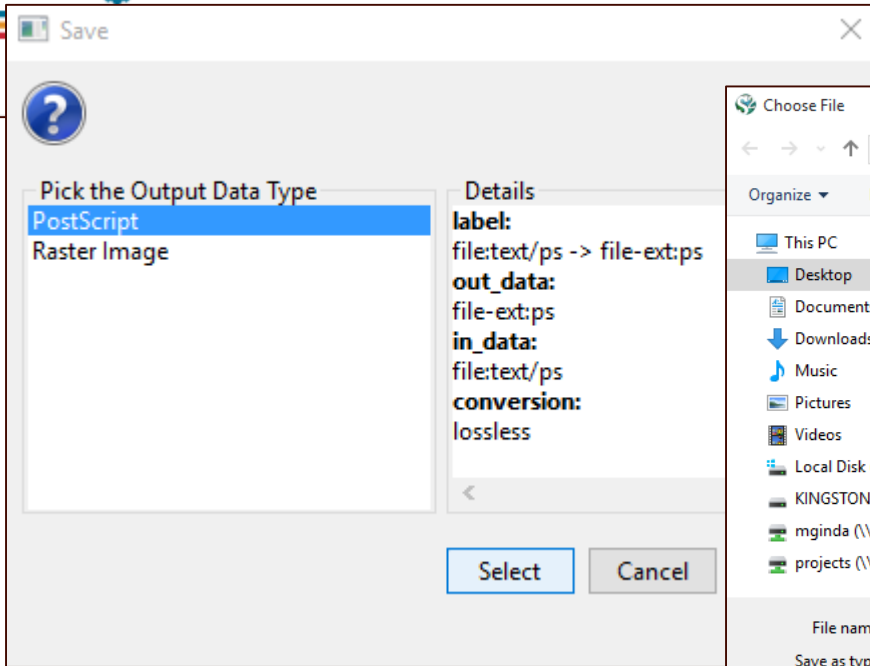
Last, is the visualization of the geospatial maps. First, set-up a new workflow in the **Workflow Manager**, as shown previously.

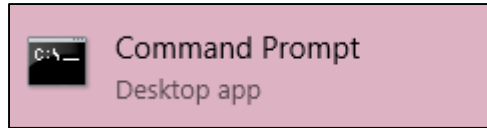
Then navigate to **Visualization > Geospatial > Proportional Symbol Map** and enter the following parameters, and select "OK".



The output from the visualization algorithm will be a Post Script file that can be turned into a PDF.

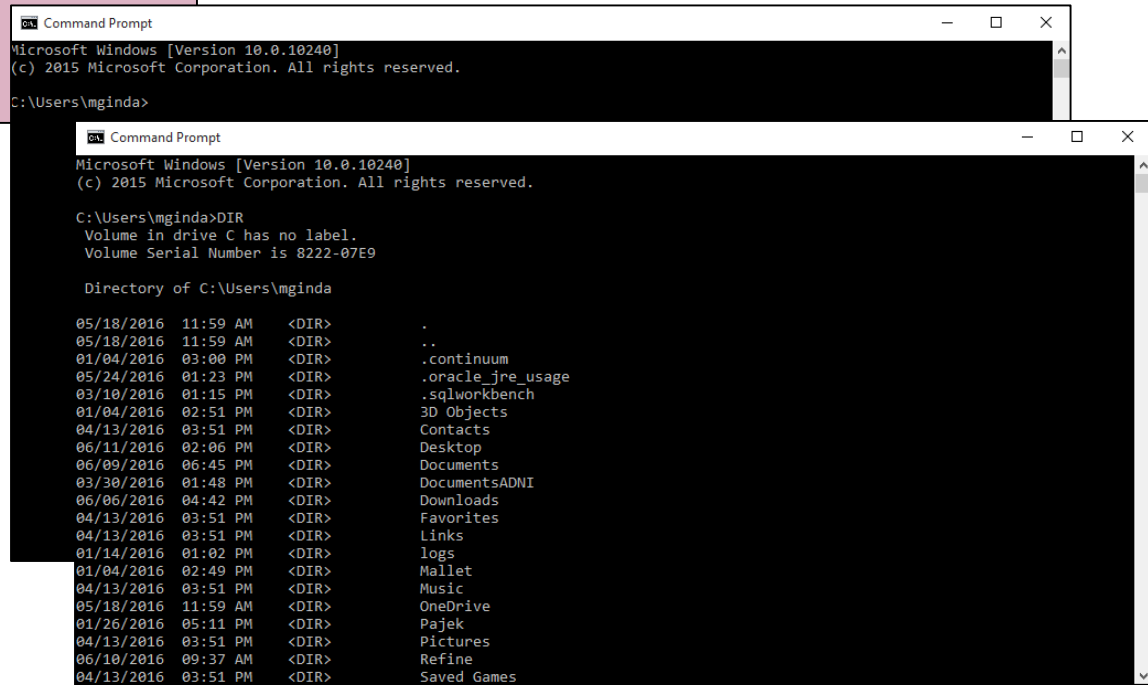
To save the file, select the PS file by right clicking it, and selecting “Save.” Pick the Output Data Type as “PostScript”, and give the map a name.





From the Windows Start menu, search for CMD, or Command Prompt.

Open a new window, and enter the DIR to locate your current directory. You will need to navigate to the directory where you saved your map PostScript files.



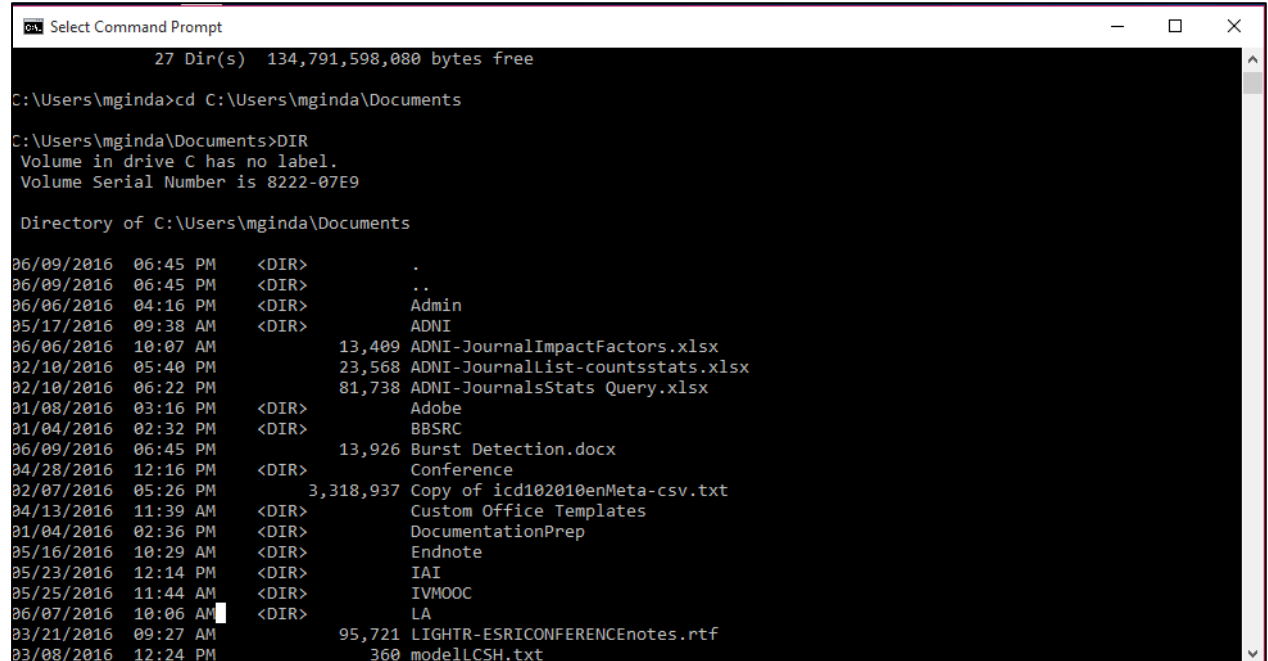
A common place to save a file may be the Desktop or the Documents folder.

To navigate to the directory you saved your PostScript file(s) in, enter

```
cd C:\Users\[username]\Desktop
```

or

```
cd C:\Users\[username]\Documents
```



```
Select Command Prompt
27 Dir(s) 134,791,598,080 bytes free

C:\Users\mginda>cd C:\Users\mginda\Documents

C:\Users\mginda\Documents>DIR
Volume in drive C has no label.
Volume Serial Number is 8222-07E9

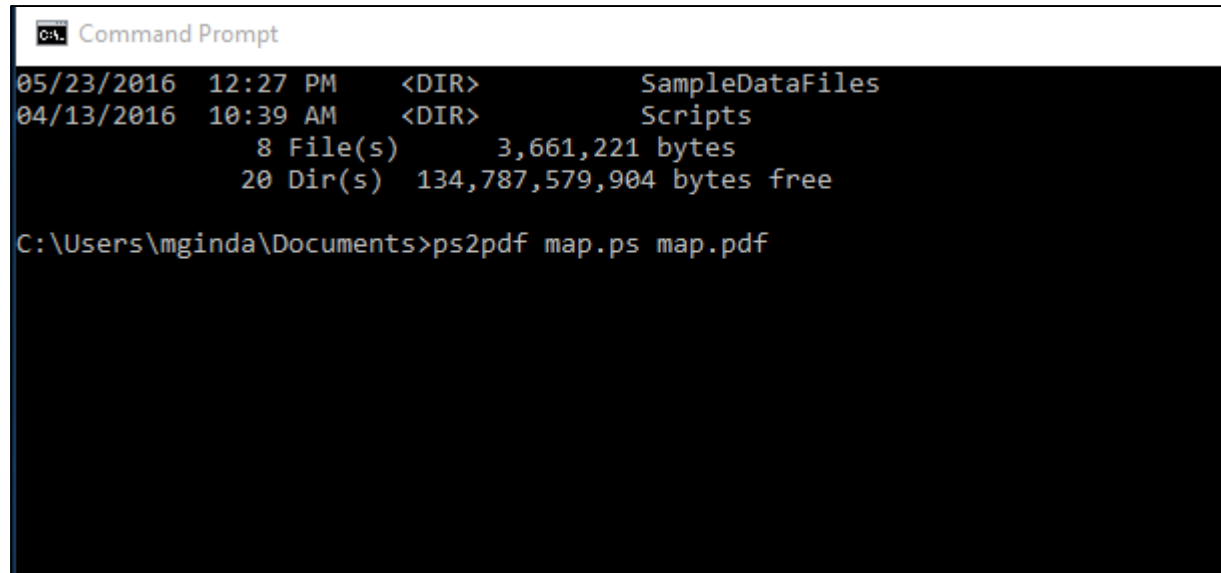
Directory of C:\Users\mginda\Documents

06/09/2016 06:45 PM <DIR>      .
06/09/2016 06:45 PM <DIR>      ..
06/06/2016 04:16 PM <DIR>      Admin
05/17/2016 09:38 AM <DIR>      ADNI
06/06/2016 10:07 AM           13,409 ADNI-JournalImpactFactors.xlsx
02/10/2016 05:40 PM           23,568 ADNI-JournalList-countsstats.xlsx
02/10/2016 06:22 PM           81,738 ADNI-JournalsStats Query.xlsx
01/08/2016 03:16 PM <DIR>      Adobe
01/04/2016 02:32 PM <DIR>      BBSRC
06/09/2016 06:45 PM           13,926 Burst Detection.docx
04/28/2016 12:16 PM <DIR>      Conference
02/07/2016 05:26 PM          3,318,937 Copy of icd102010enMeta-csv.txt
04/13/2016 11:39 AM <DIR>      Custom Office Templates
01/04/2016 02:36 PM <DIR>      DocumentationPrep
05/16/2016 10:29 AM <DIR>      Endnote
05/23/2016 12:14 PM <DIR>      IAI
05/25/2016 11:44 AM <DIR>      IVMOOC
06/07/2016 10:06 AM <DIR>      LA
03/21/2016 09:27 AM           95,721 LIGHTR-ESRICONFERENCEnotes.rtf
03/08/2016 12:24 PM           360 modelLCSH.txt
```

Once you have navigated to the proper directory, you can run the PS2PDF program. In the command line run:

```
ps2pdf [options] input.[e]ps output.pdf
```

If you leave off the output file name, the resulting PDF file will have the same name as the original file with a .PDF extension into the same directory you navigated to.



```
Command Prompt
05/23/2016 12:27 PM <DIR> SampleDataFiles
04/13/2016 10:39 AM <DIR> Scripts
      8 File(s)      3,661,221 bytes
     20 Dir(s)  134,787,579,904 bytes free

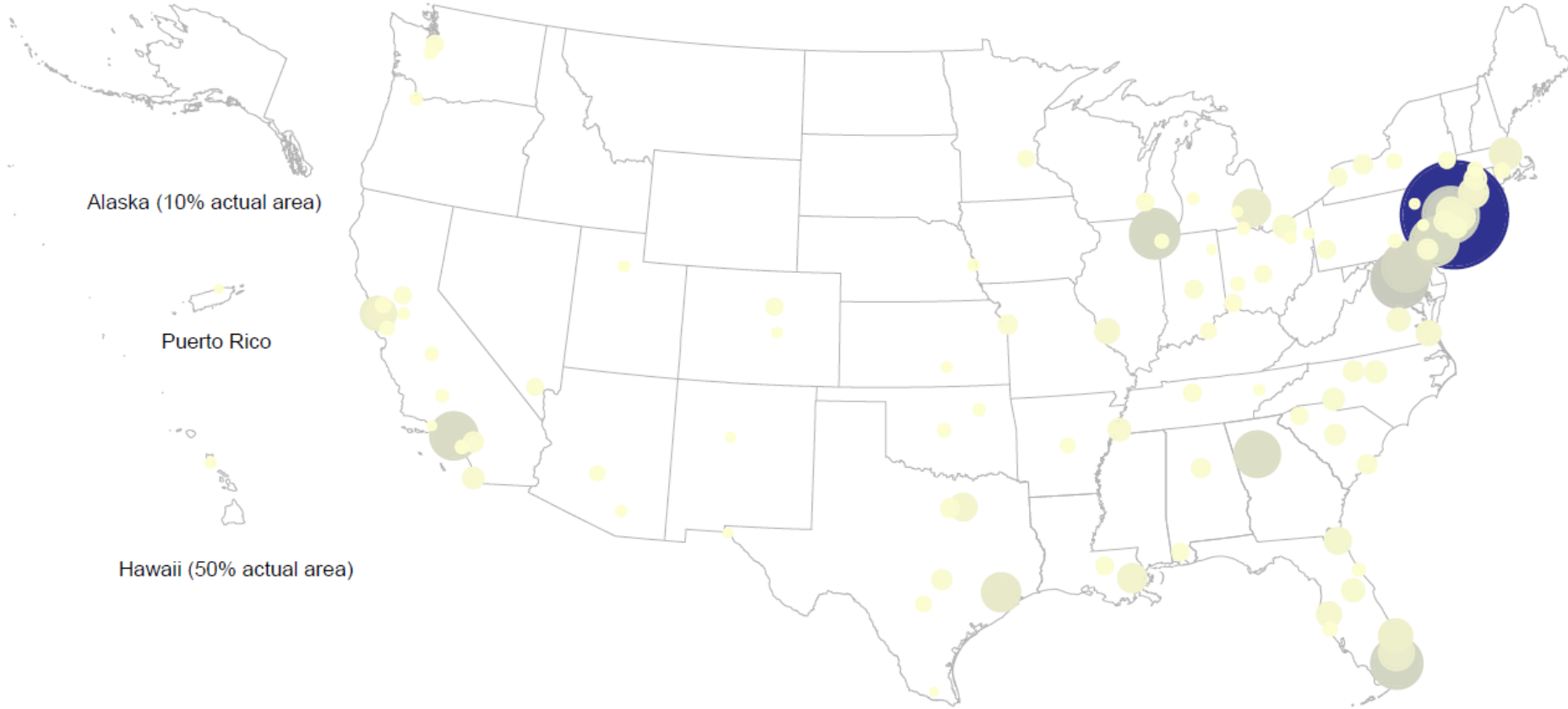
C:\Users\mginda\Documents>ps2pdf map.ps map.pdf
```

You can view the PDF file with Adobe Reader. More information on this script tool is available in the [PS2PDF documentation site](#).

Geospatial Visualization (Proportional Symbol Map)

Generated from Aggregation performed using unique values in 'Location' column.

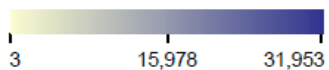
Jun 09, 2016 | 05:19:34 PM EDT



Legend

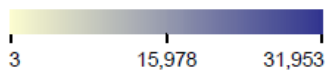
Interior Color (Linear)

Cases



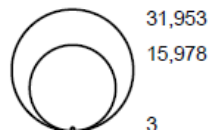
Exterior Color (Linear)

Cases



Area (Linear)

Cases



How to Read this Map

This *proportional symbol map* shows 52 U.S. states and other jurisdictions using the Albers equal-area conic projection with Alaska, Puerto Rico, and Hawaii inset. Each dataset record is represented by a circle centered at its geolocation. The area, interior color, and exterior color of each circle may represent numeric attribute values. Minimum and maximum data values are given in the legend.

Questions?

8:00 Welcome and Overview of Tutorial and Attendees**8:30 Visualization Framework and Workflow Design**

- Overview of the Visualization framework
- Overview of Graphical variables, and color selection

9:15 Tool Overview and Trouble Shooting

- Sci2 Overview – scientometric analysis tool
- Open Refine – data parsing, transformation, and editing tool
- Gephi – network visualization

9:45 Geospatial Analysis

- Overview of geospatial analysis and mapping
- Geospatial Analysis: Geocoding with OpenRefine
- Geospatial Analysis: Proportional Symbol Map using CDC

11:00 Topical/Temporal Analysis: Burst Detection

- Overview of burst analysis and introductory workflow
- Burst Detection with CDC Grants

12:30 Lunch**1:30 Network Analysis**

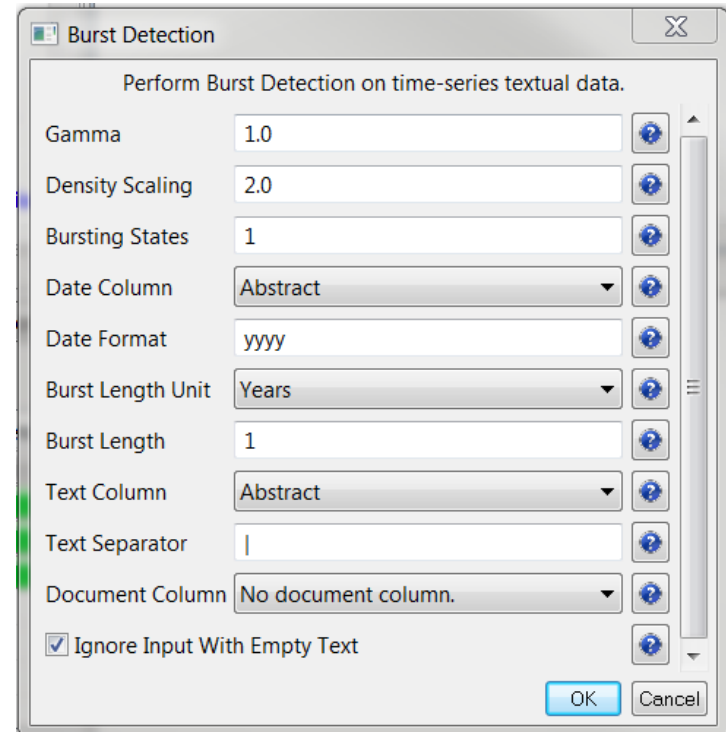
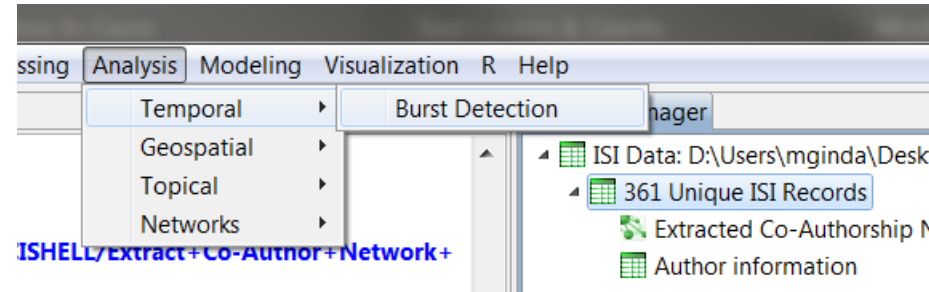
- Introduction to Network Analysis
- ~~Network Analysis: Bimodal networks with Morbidity Data~~
- Network Analysis: Co-authorship Network with CDC publications

4:00 Wrap-up

4:30 Adjourn

- Science evolves over time
- Temporal analysis seeks to study this evolution by examining patterns, trends, seasonality, outliers, and bursts of activity
- Time series data can be thought of as either discrete or continuous
- Many scholarly datasets can be understood as a discrete time series with events or observations (publications etc.) that happen at regularly spaced intervals (journal publication cycles etc.)

- Sci2 uses an implementation of Kleinberg's burst detection algorithm (Kleinberg 2002) to study bursts in usage of words in scholarly data
- Algorithm does not calculate the frequency of individual words, instead bursts is a measure of rate of use.
- Algorithm uses probabilistic Markov model to determine the rate at which use of a word increases or decreases, identifying bursts in usage of a word during a period of time.



Kleinberg, J. (2002). [Bursty and Hierarchical Structure in Streams](#). Proceedings from the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Canada: ACM.

Gamma – the higher this value, the smaller the list of generated bursts.

Density Scaling – determines how much “more bursty” each level is beyond the previous one.

Bursting states – determines how many bursting states there will be, beyond the non-bursting states.

Date Column – name of the column in the original data with date/time when events/topics happen.

Date Format – specifies how the date column will be interpreted.

Burst Length Unit – specifies how to divide the date range into burstable units.

Burst Length – specifies the number of burstable units per burstable period.

Text Column – the name of the column with values (delimiter and tokens) to be computed for bursting results.

Text Separator – delimits the tokens in the text column.

Questions?

8:00 Welcome and Overview of Tutorial and Attendees**8:30 Visualization Framework and Workflow Design**

- Overview of the Visualization framework
- Overview of Graphical variables, and color selection

9:15 Tool Overview and Trouble Shooting

- Sci2 Overview – scientometric analysis tool
- Open Refine – data parsing, transformation, and editing tool
- Gephi – network visualization

9:45 Geospatial Analysis

- Overview of geospatial analysis and mapping
- Geospatial Analysis: Geocoding with OpenRefine
- Geospatial Analysis: Proportional Symbol Map using CDC

11:00 Topical/Temporal Analysis: Burst Detection

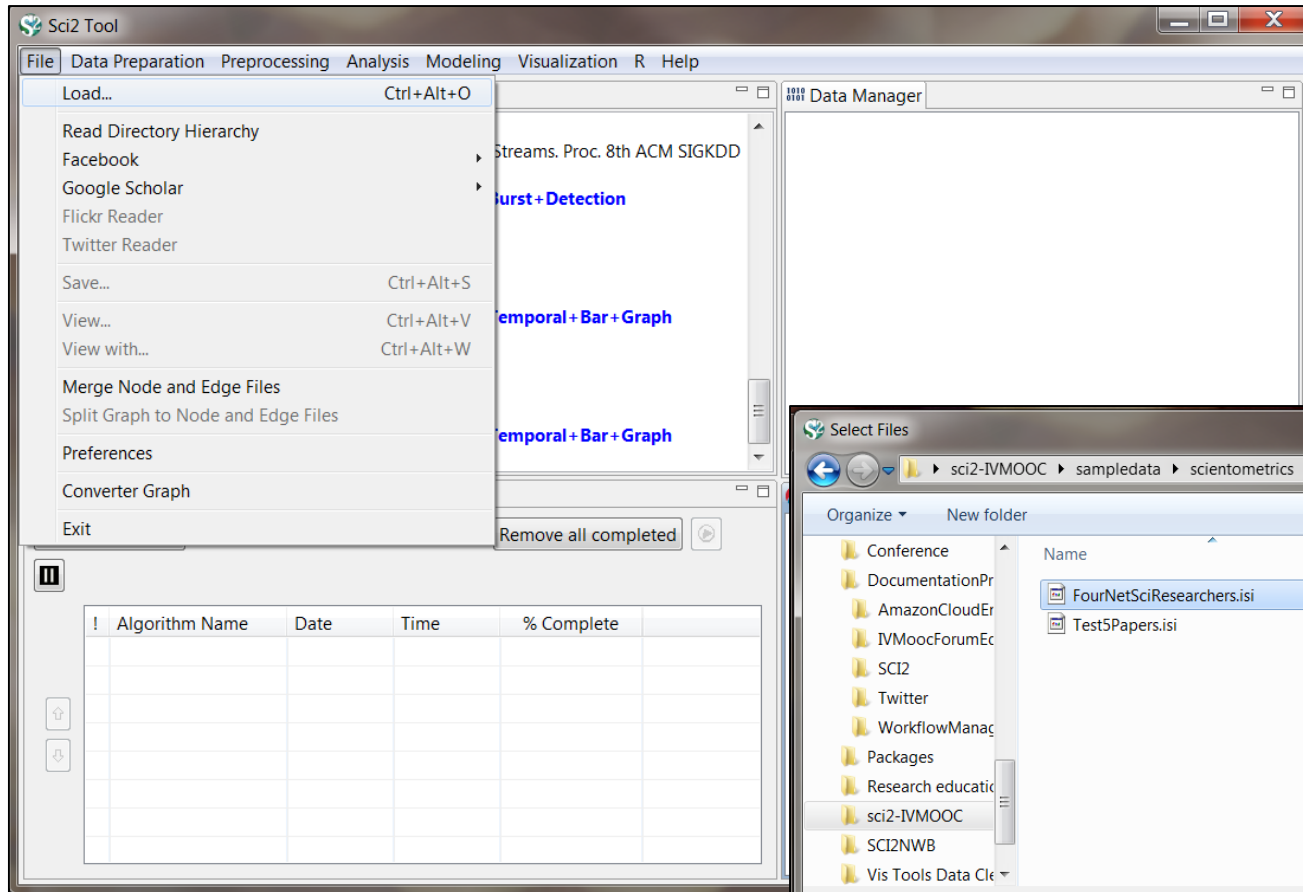
- Overview of burst analysis and introductory workflow
- **Burst Detection with CDC Grants**

12:30 Lunch**1:30 Network Analysis**

- Introduction to Network Analysis
- Network Analysis: Bimodal networks with Morbidity Data
- Network Analysis: Co-authorship Network with CDC publications

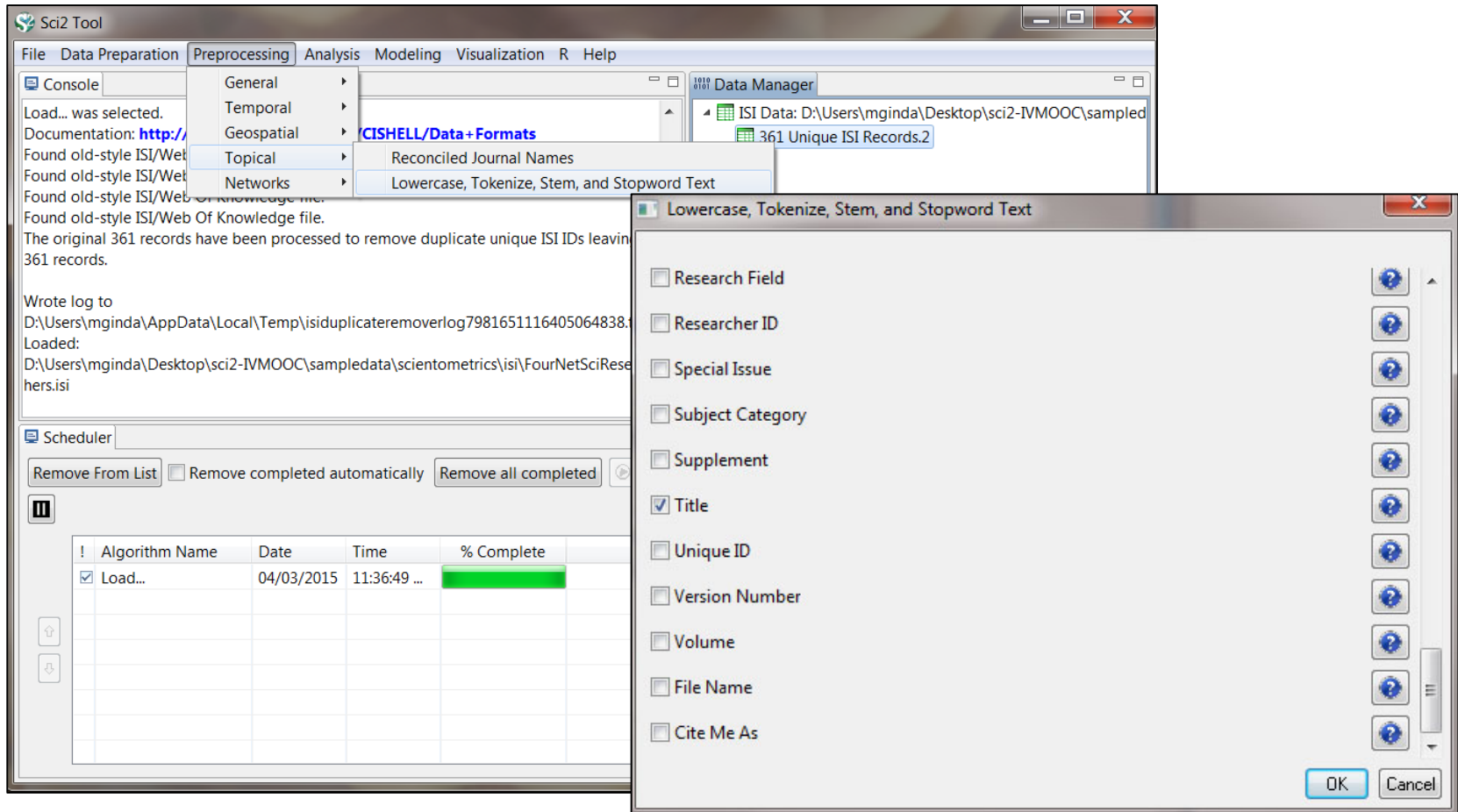
4:00 Wrap-up

4:30 Adjourn



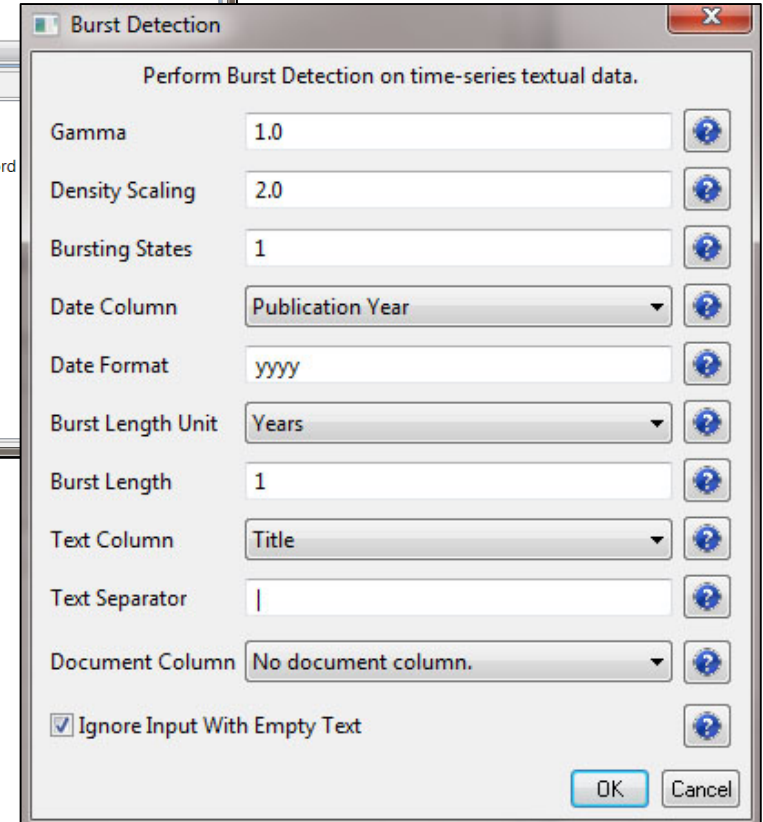
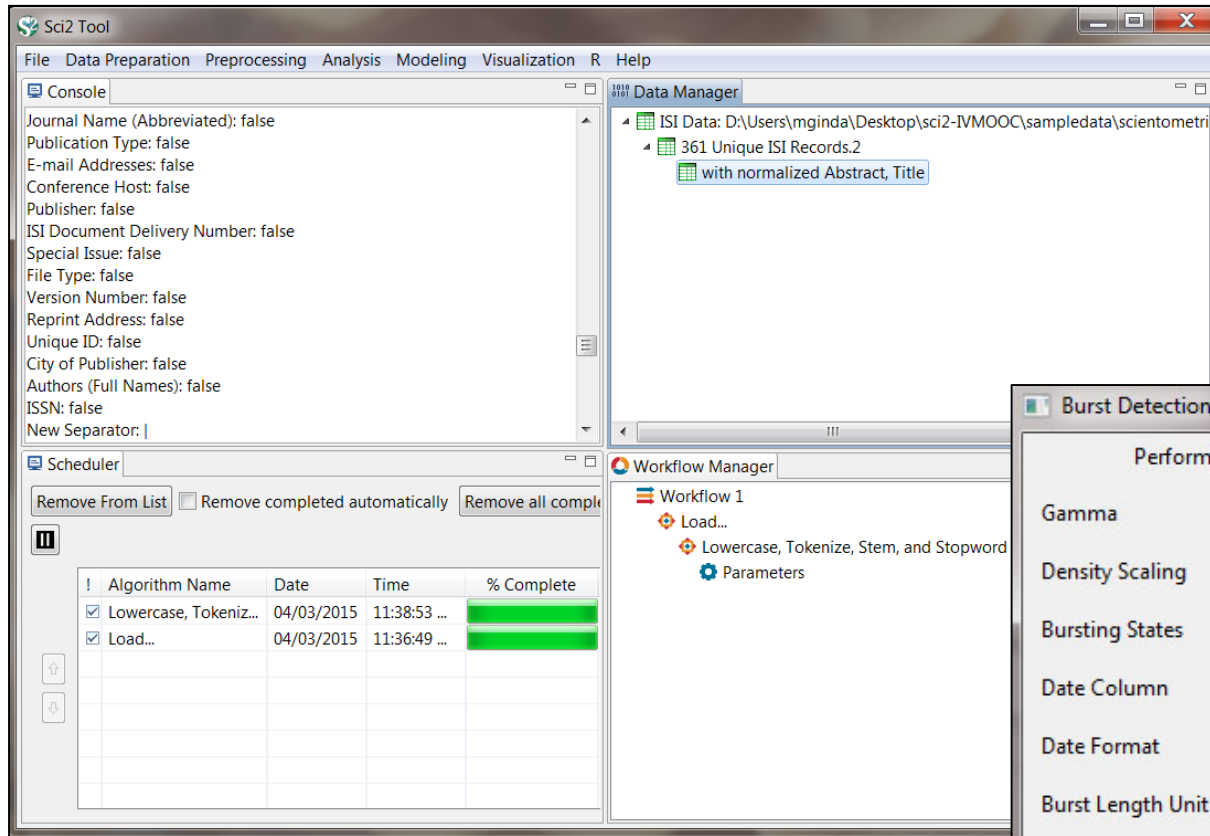
Load *FourNetSciResearchers.isi*

Located in Sci2 Directory -> sampledata
-> scientometrics -> isi

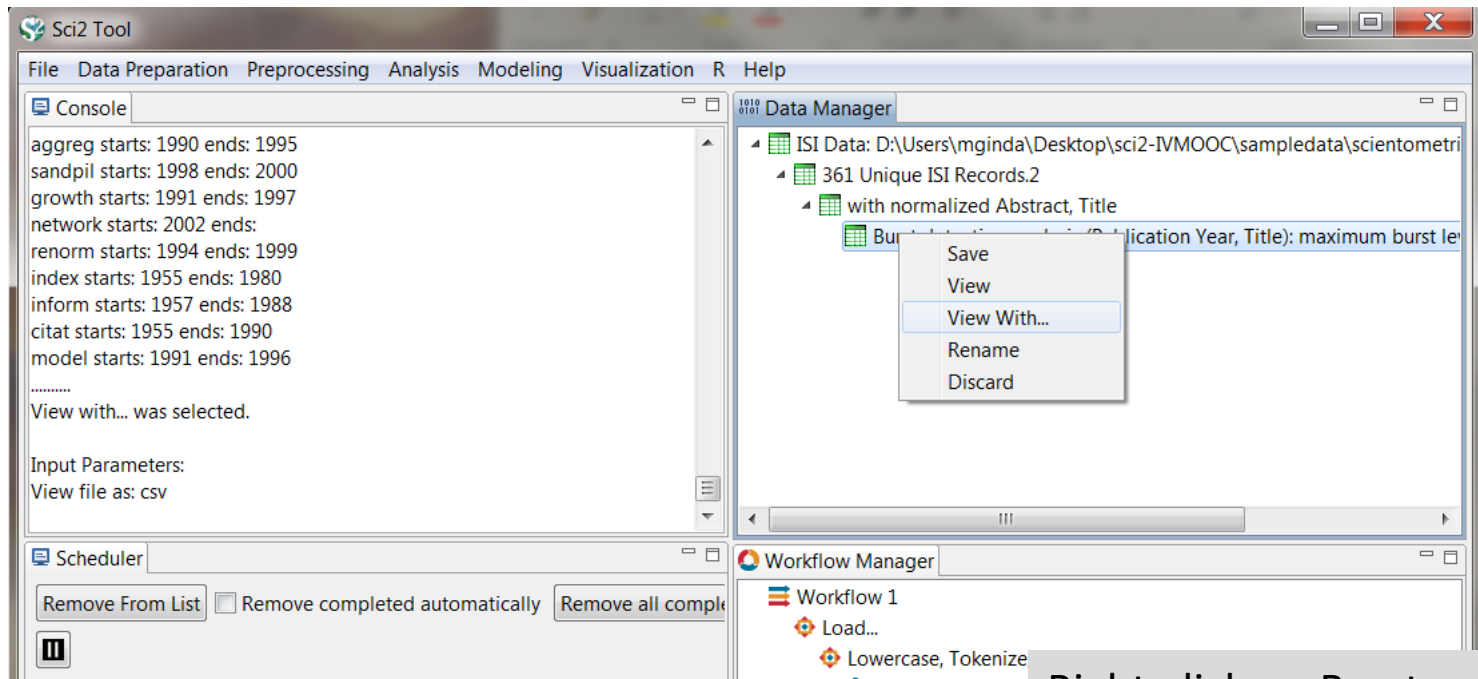


Select *Preprocessing* > *Topical* > *Lowercase, Tokenize, Stem, and Stopword Text*

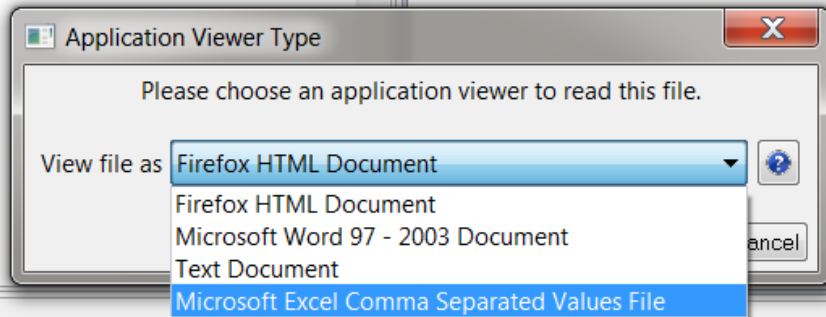
Then select **Title** from the input parameters



Highlight the table 'with normalized Title'
 Select *Analysis > Temporal > Burst Detection*
 Then set the parameters to what is shown to the right

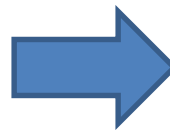


Right-click on Burst detection analysis (Publication Year, Title): maximum burst level 1 in the data manager and view the file in the spreadsheet program of your choice



Missing end dates indicate the continuation of a burst in a given data set. Add the End date of 2014 to those records missing and End date.

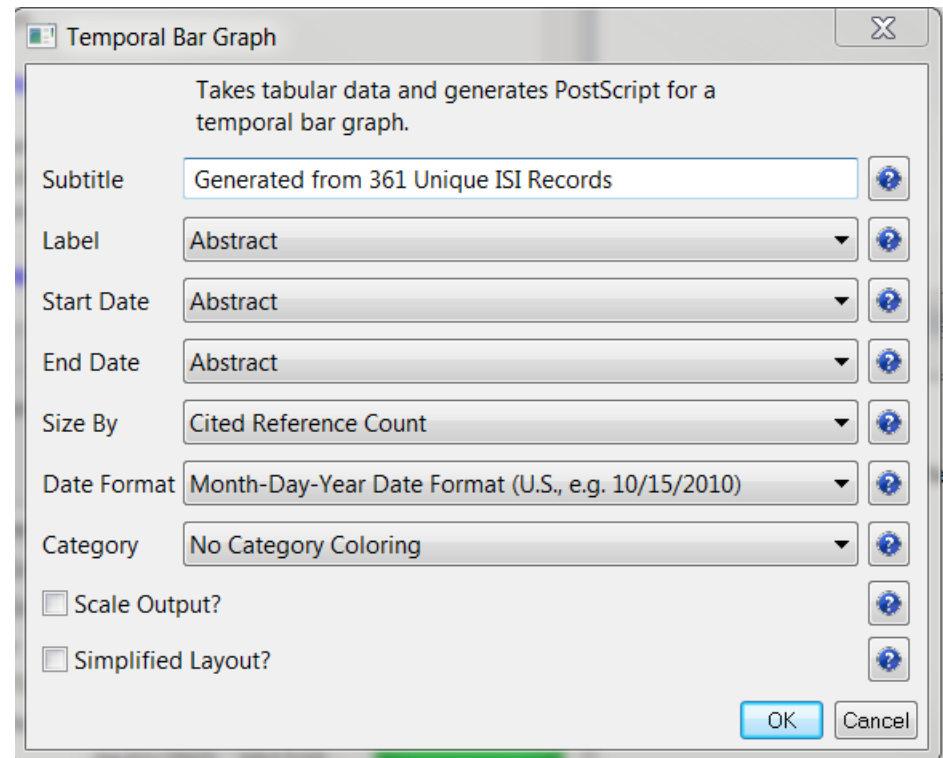
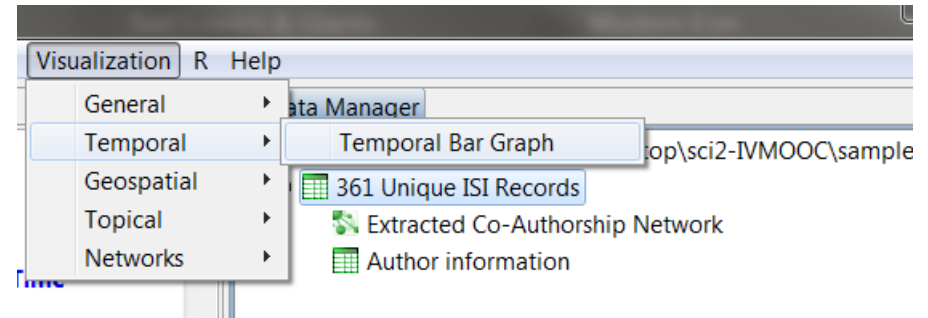
	A	B	C	D	E	F	G
1	Word	Level	Weight	Length	Start	End	
2	transform	1	4.447253	8	1988	1995	
3	analysi	1	5.170601	22	1972	1993	
4	scienc	1	6.959986	31	1955	1985	
5	critic	1	5.934392	6	1993	1998	
6	complex	1	8.823672	8	2000		
7	chemic	1	4.026779	22	1957	1978	
8	self	1	5.584457	10	1990	1999	
9	fractal	1	4.654846	8	1990	1997	
10	transit	1	4.855824	4	1997	2000	
11	protein	1	4.017034	5	2003		
12	aggreg	1	4.791958	6	1990	1995	
13	sandpil	1	4.789993	3	1998	2000	
14	growth	1	5.953631	7	1991	1997	
15	network	1	21.68094	6	2002		
16	renorm	1	4.301231	6	1994	1999	
17	index	1	9.816979	26	1955	1980	
18	inform	1	4.867829	32	1957	1988	
19	citat	1	6.541599	36	1955	1990	
20	model	1	9.214359	6	1991	1996	
21							

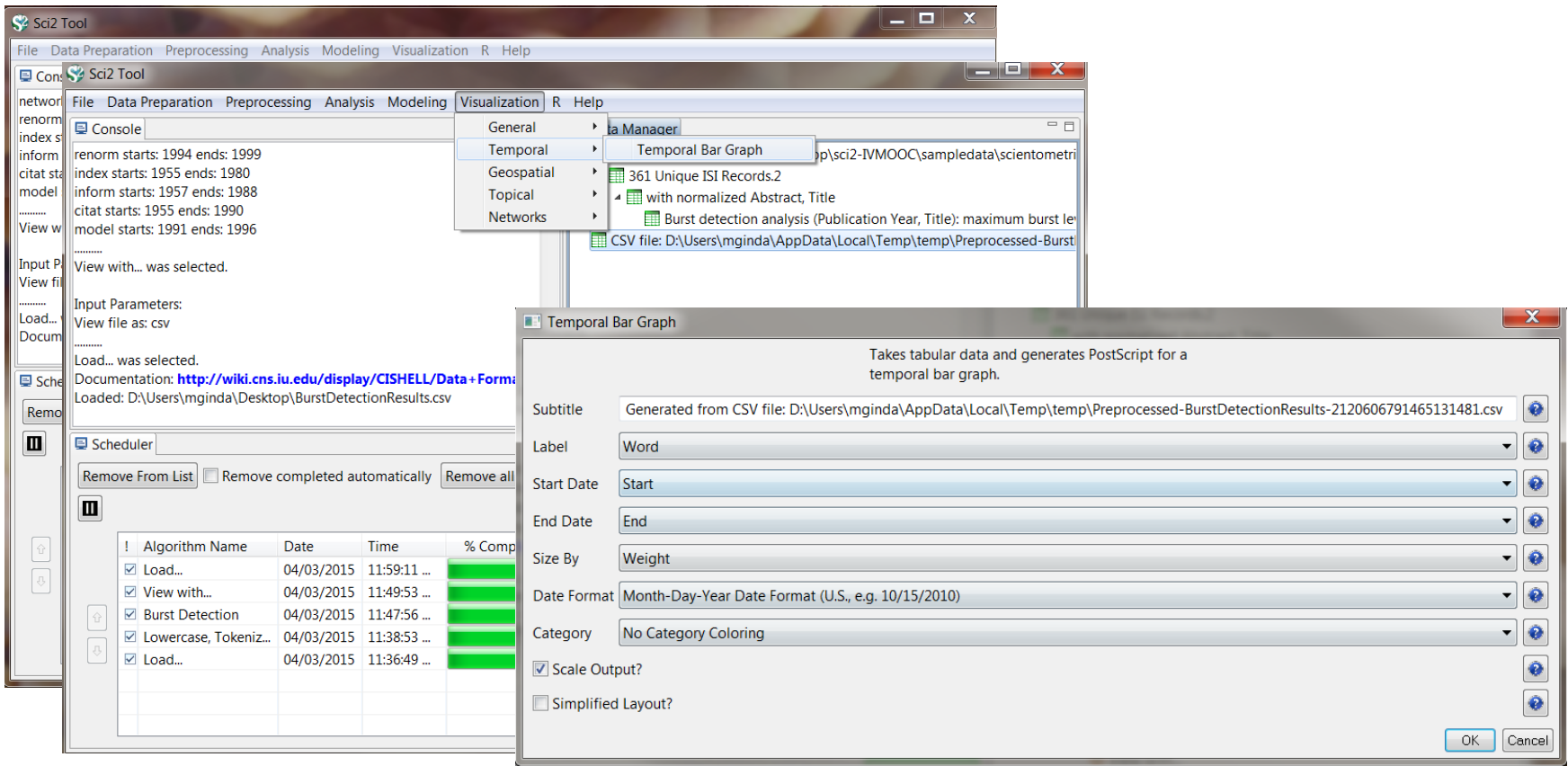


	A	B	C	D	E	F	G	H
1	Word	Level	Weight	Length	Start	End		
2	transform	1	4.447253	8	1988	1995		
3	analysi	1	5.170601	22	1972	1993		
4	scienc	1	6.959986	31	1955	1985		
5	critic	1	5.934392	6	1993	1998		
6	complex	1	8.823672	8	2000	2014		
7	chemic	1	4.026779	22	1957	1978		
8	self	1	5.584457	10	1990	1999		
9	fractal	1	4.654846	8	1990	1997		
10	transit	1	4.855824	4	1997	2000		
11	protein	1	4.017034	5	2003	2014		
12	aggreg	1	4.791958	6	1990	1995		
13	sandpil	1	4.789993	3	1998	2000		
14	growth	1	5.953631	7	1991	1997		
15	network	1	21.68094	6	2002	2014		
16	renorm	1	4.301231	6	1994	1999		
17	index	1	9.816979	26	1955	1980		
18	inform	1	4.867829	32	1957	1988		
19	citat	1	6.541599	36	1955	1990		
20	model	1	9.214359	6	1991	1996		
21								

Save the file as a .CSV file and load it back into Sci2,
selecting the Standard CSV format

- Visualizes numeric data over time
- It accepts a CSV file as input, including NSF grant data
- Start and end dates for each record are necessary to use the temporal bar graph visualization algorithm
- The output of the visualization consists of labeled horizontal bars that correspond to records in the original dataset.





The screenshot shows the Sci2 Tool interface with the 'Temporal Bar Graph' dialog box open. The dialog box contains the following settings:

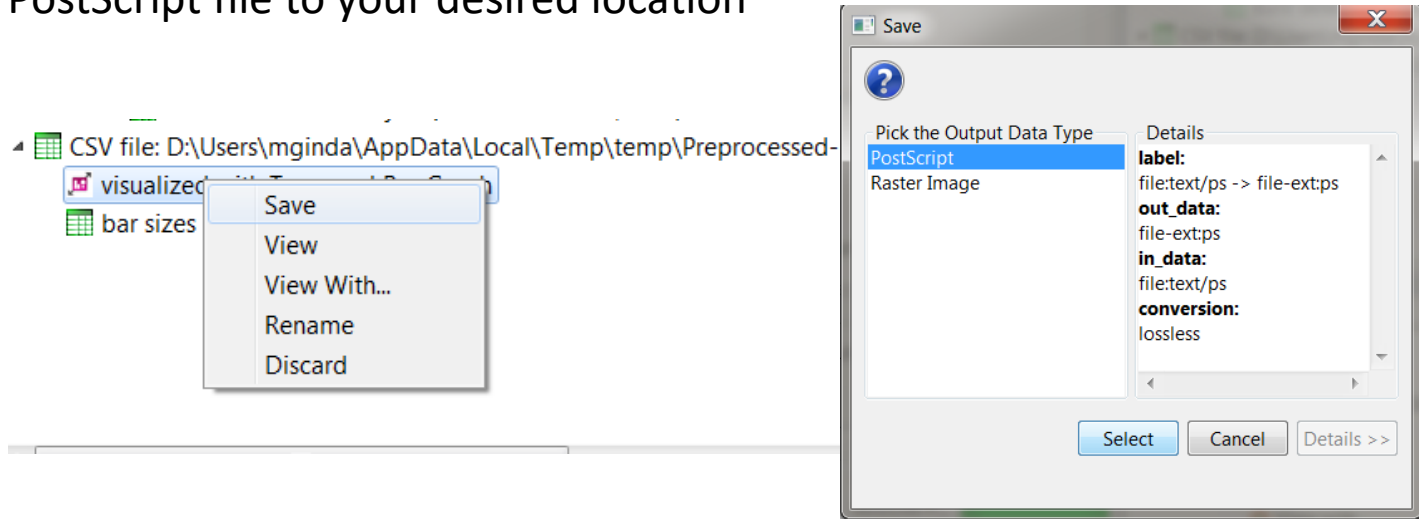
- Subtitle: Generated from CSV file: D:\Users\mginda\AppData\Local\Temp\temp\Preprocessed-BurstDetectionResults-2120606791465131481.csv
- Label: Word
- Start Date: Start
- End Date: End
- Size By: Weight
- Date Format: Month-Day-Year Date Format (U.S., e.g. 10/15/2010)
- Category: No Category Coloring
- Scale Output?
- Simplified Layout?

The background interface shows the 'Visualization' menu path: Visualization > Temporal > Temporal Bar Graph. A table in the Scheduler window shows the following data:

Algorithm Name	Date	Time	% Comp
Load...	04/03/2015	11:59:11 ...	100%
View with...	04/03/2015	11:49:53 ...	100%
Burst Detection	04/03/2015	11:47:56 ...	100%
Lowercase, Tokeniz...	04/03/2015	11:38:53 ...	100%
Load...	04/03/2015	11:36:49 ...	100%

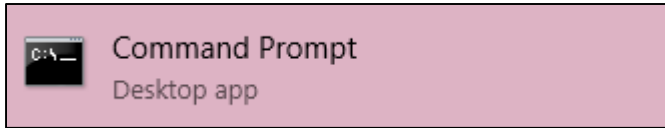
- Load updated Burst Detection result file into Sci2 as a standard CSV format.
- Select the newly loaded file in the data manager and the *Visualization > Temporal > Temporal Bar Graph* in the menu bar.
- Set the parameter values to those shown to the right

Right-click on the visualized with **Temporal Bar Graph** file in the Data Manager and save the PostScript file to your desired location



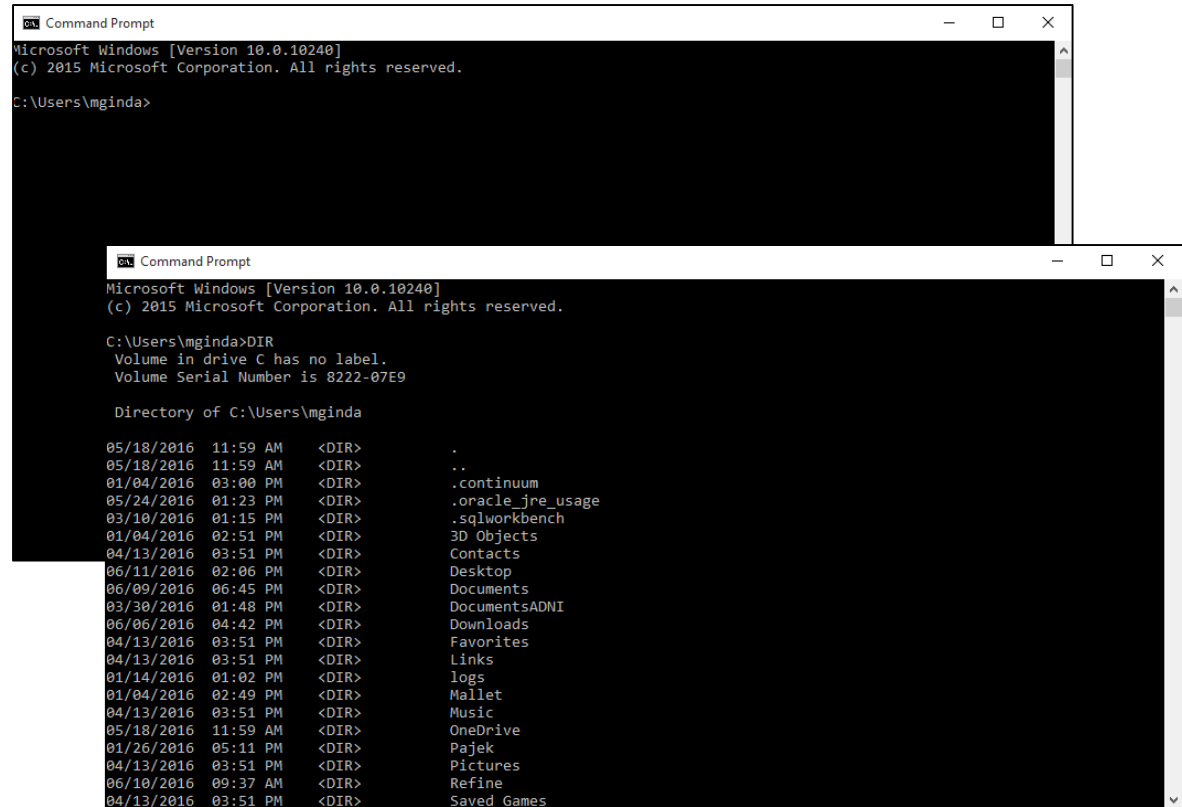
To convert the PS file output to a PDF, we will use GhostScripts command line tool, PS2PDF. Open Windows' Command Prompt. Next

If you do not have a program or script to convert PostScript files using PS2PDF.com.



From the Windows Start menu, search for CMD, or Command Prompt.

Open a new window, and enter the DIR to locate your current directory. You will need to navigate to the directory where you saved your burst detection results PostScript files.



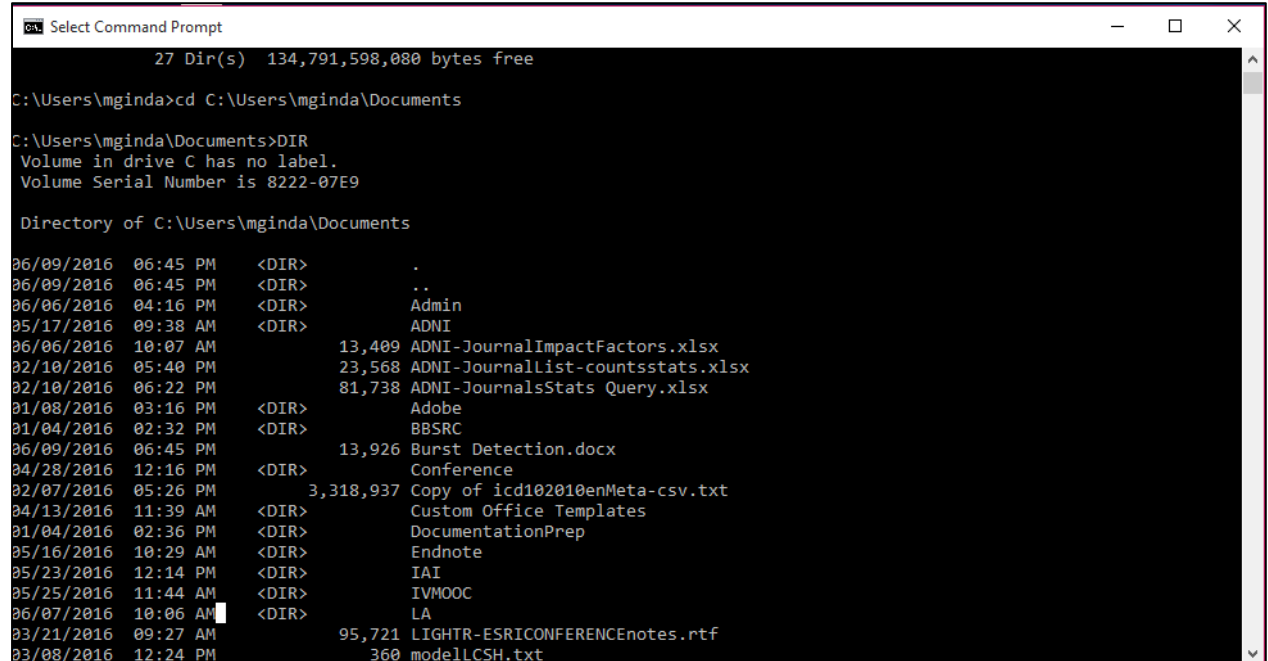
A common place to save a file may be the Desktop or the Documents folder.

To navigate to the directory you saved your PostScript file(s) in, enter

```
cd C:\Users\[username]\Desktop
```

or

```
cd C:\Users\[username]\Documents
```



```
Select Command Prompt
27 Dir(s) 134,791,598,080 bytes free

C:\Users\mginda>cd C:\Users\mginda\Documents

C:\Users\mginda\Documents>DIR
Volume in drive C has no label.
Volume Serial Number is 8222-07E9

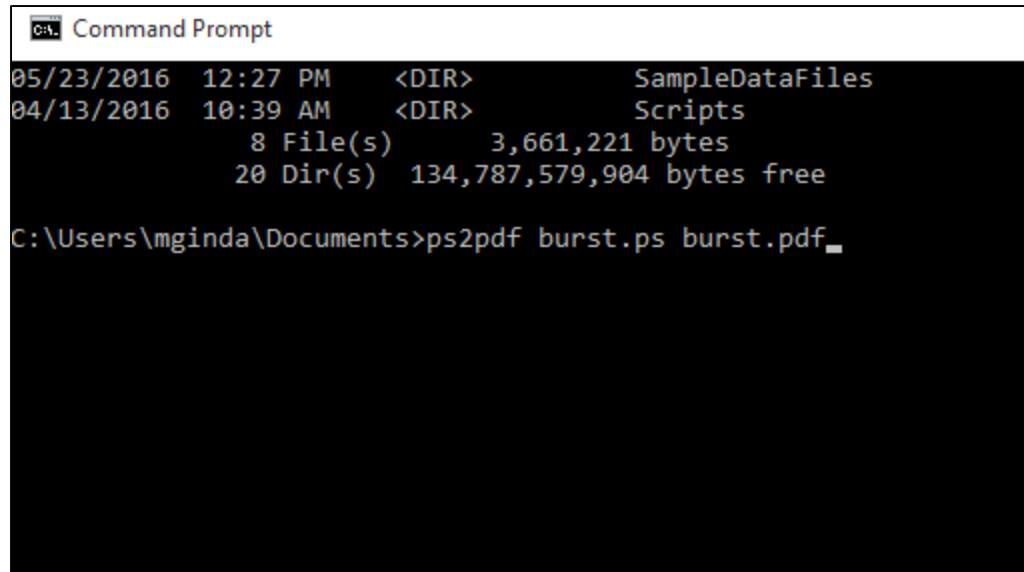
Directory of C:\Users\mginda\Documents

06/09/2016 06:45 PM <DIR>      .
06/09/2016 06:45 PM <DIR>      ..
06/06/2016 04:16 PM <DIR>      Admin
05/17/2016 09:38 AM <DIR>      ADNI
06/06/2016 10:07 AM           13,409 ADNI-JournalImpactFactors.xlsx
02/10/2016 05:40 PM           23,568 ADNI-JournalList-countsstats.xlsx
02/10/2016 06:22 PM           81,738 ADNI-JournalsStats Query.xlsx
01/08/2016 03:16 PM <DIR>      Adobe
01/04/2016 02:32 PM <DIR>      BBSRC
06/09/2016 06:45 PM           13,926 Burst Detection.docx
04/28/2016 12:16 PM <DIR>      Conference
02/07/2016 05:26 PM      3,318,937 Copy of icd102010enMeta-csv.txt
04/13/2016 11:39 AM <DIR>      Custom Office Templates
01/04/2016 02:36 PM <DIR>      DocumentationPrep
05/16/2016 10:29 AM <DIR>      Endnote
05/23/2016 12:14 PM <DIR>      IAI
05/25/2016 11:44 AM <DIR>      IVMOOC
06/07/2016 10:06 AM <DIR>      LA
03/21/2016 09:27 AM           95,721 LIGHTR-ESRICONFERENCEnotes.rtf
03/08/2016 12:24 PM           360  modelLCSH.txt
```

Once you have navigated to the proper directory, you can run the PS2PDF program. In the command line run:

```
ps2pdf [options] input.[e]ps output.pdf
```

If you leave off the output file name, the resulting PDF file will have the same name as the original file with a .PDF extension into the same directory you navigated to.



```
Command Prompt
05/23/2016 12:27 PM <DIR> SampleDataFiles
04/13/2016 10:39 AM <DIR> Scripts
           8 File(s)      3,661,221 bytes
          20 Dir(s)  134,787,579,904 bytes free

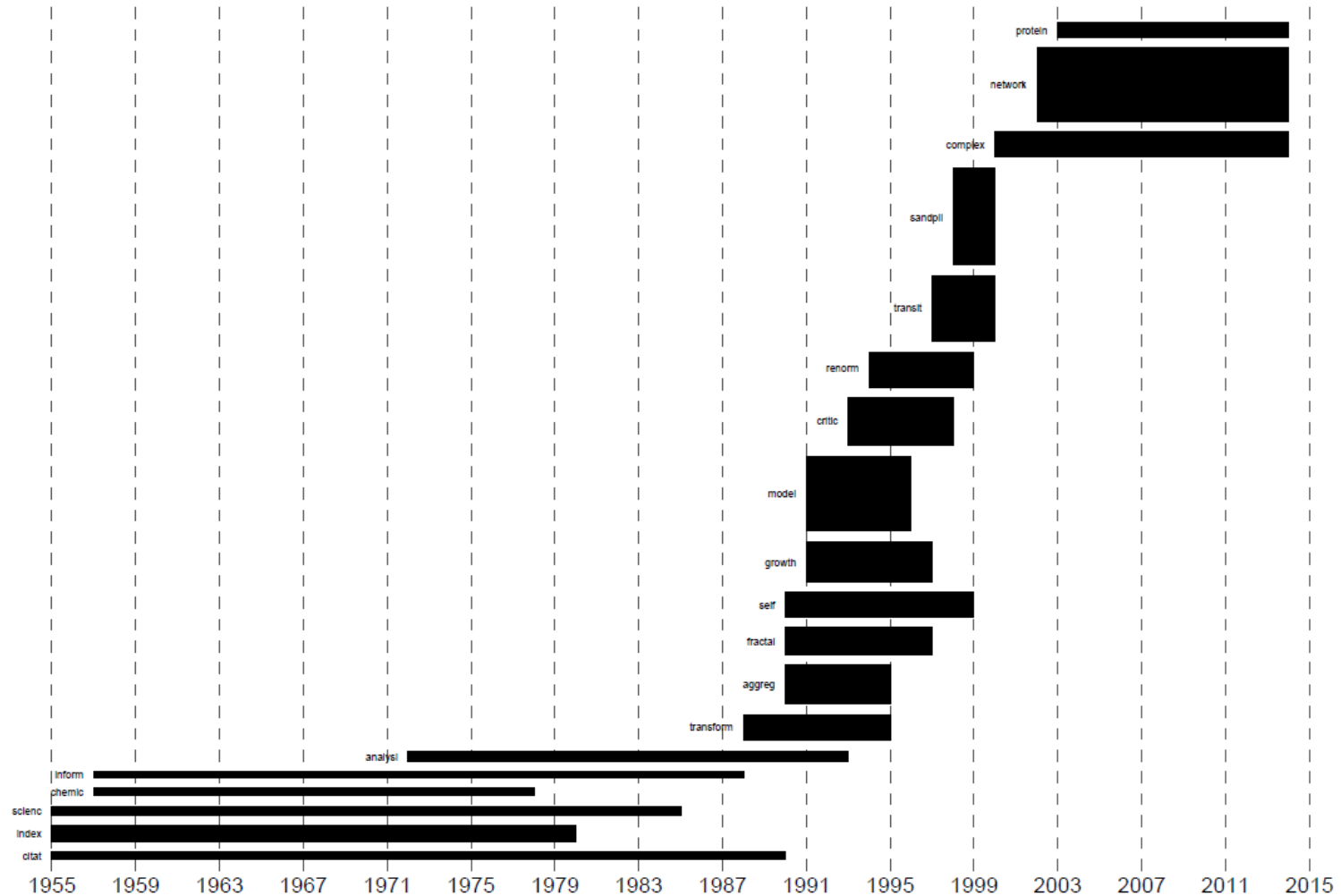
C:\Users\mginda\Documents>ps2pdf burst.ps burst.pdf_
```

You can view the PDF file with Adobe Reader. More information on this script tool is available in the [PS2PDF documentation site](#).

Temporal Visualization

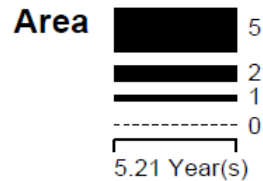
(Generated from CSV file: D:\Users\mginda\AppData\Local\Temp\temp\Preprocessed-BurstDetectionResults-4411812064438418826.csv)

April 03, 2015 | 12:05 PM EDT



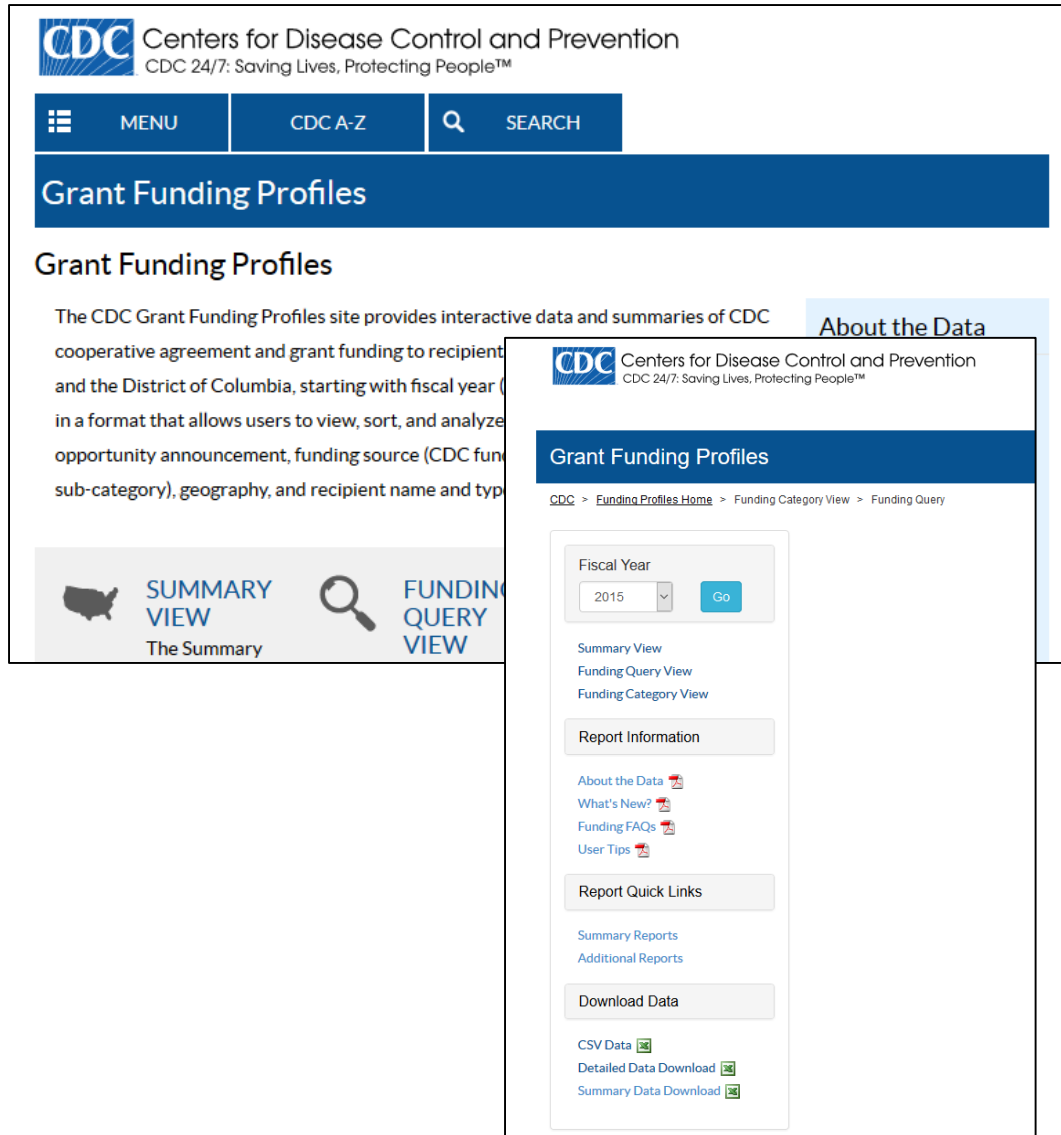
Legend

Area size: Weight
Minimum = 4
Maximum = 22
Text label: Word



How To Read This Map

This *temporal bar graph* visualization represents each record as a horizontal bar with a specific start and end date and a text label on its left side. The area of each bar encodes a numerical attribute value, e.g., total amount of funding. Bars may be colored to present categorical attribute values of records.



The screenshot shows the CDC Grant Funding Profiles website. At the top left is the CDC logo and the text "Centers for Disease Control and Prevention CDC 24/7: Saving Lives, Protecting People™". Below this is a navigation bar with "MENU", "CDC A-Z", and "SEARCH". The main heading is "Grant Funding Profiles".

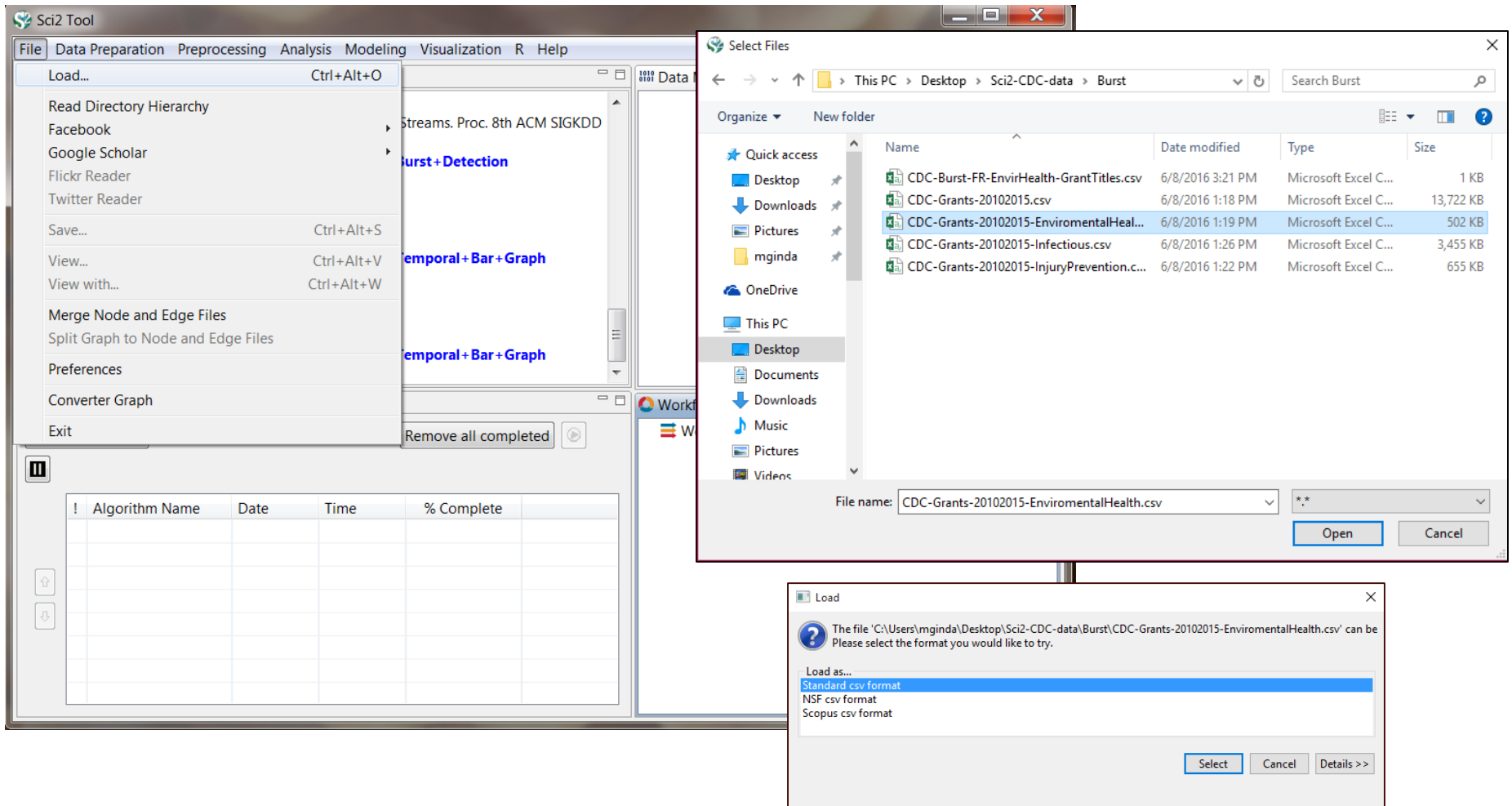
Below the heading, there is a sub-heading "Grant Funding Profiles" and a paragraph: "The CDC Grant Funding Profiles site provides interactive data and summaries of CDC cooperative agreement and grant funding to recipient organizations and the District of Columbia, starting with fiscal year 2010 in a format that allows users to view, sort, and analyze funding data by opportunity announcement, funding source (CDC funding sub-category), geography, and recipient name and type." To the right of this paragraph is a link for "About the Data".

Below the paragraph are two main options: "SUMMARY VIEW The Summary" (with a map icon) and "FUNDING QUERY VIEW" (with a magnifying glass icon). A smaller inset window shows a detailed view of the "FUNDING QUERY VIEW" interface, which includes a "Fiscal Year" dropdown menu set to "2015" and a "Go" button. Below this are links for "Summary View", "Funding Query View", and "Funding Category View". There are also sections for "Report Information", "Report Quick Links", and "Download Data", with the latter containing links for "CSV Data", "Detailed Data Download", and "Summary Data Download".

CDC Grant funding data was collected for each fiscal year between 2010-2015 from the Grant Funding Profiles page for CDC.

The funding data include actions awarded (i.e., obligated funds) domestically in each federal fiscal year (October 1st of one year to September 30th of the next year) from CDC's annual appropriation. International funding are not found here.

This data is available to download as in the CSV format.



The screenshot shows the Sci2 Tool interface with a 'File' menu open, a 'Select Files' dialog box, and a 'Load' dialog box. The 'Select Files' dialog shows a list of files in the 'Burst' directory, with 'CDC-Grants-20102015-EnviromentalHealth.csv' selected. The 'Load' dialog shows the file format selection options.

Select Files Dialog:

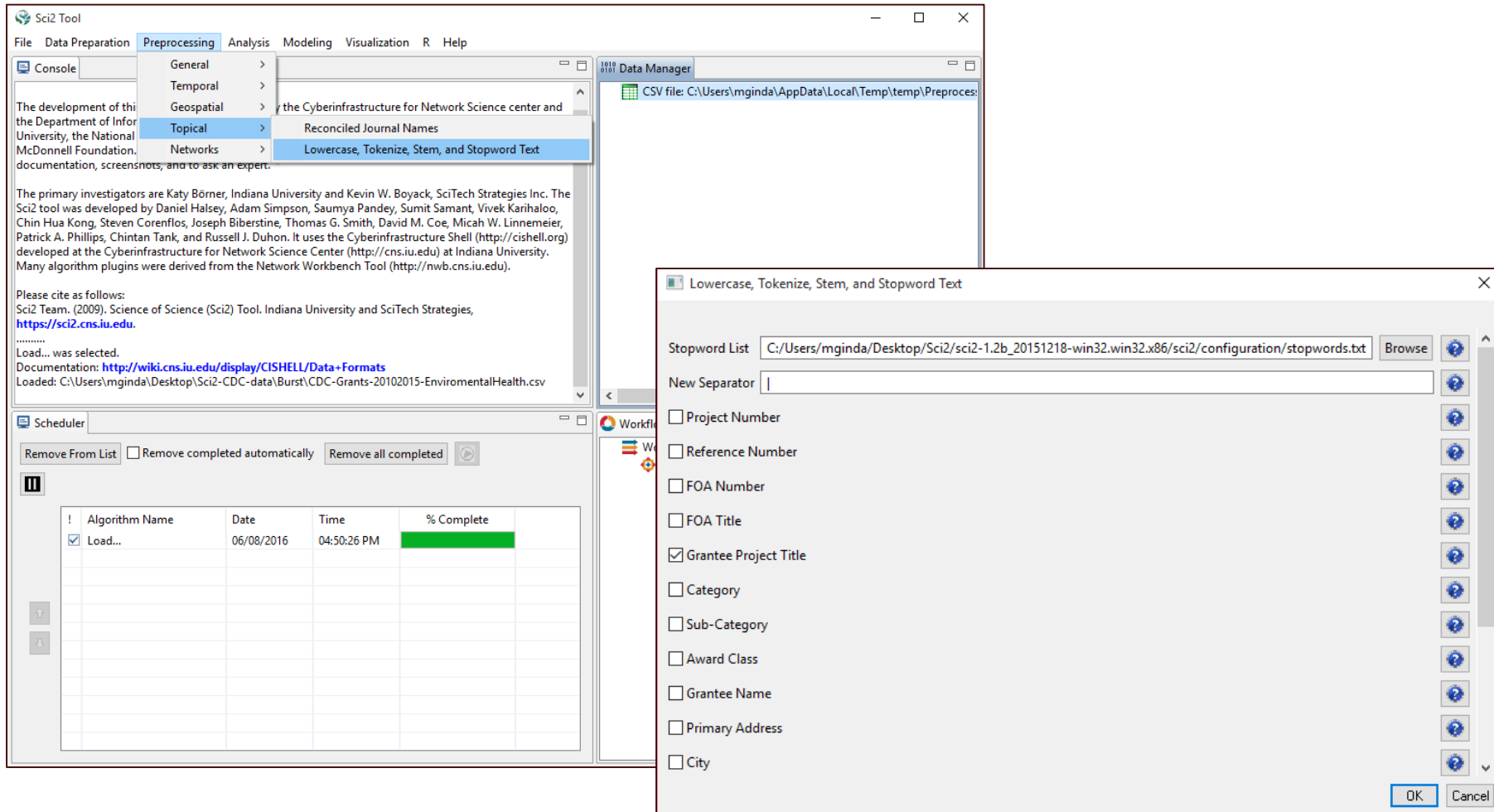
Name	Date modified	Type	Size
CDC-Burst-FR-EnvirHealth-GrantTitles.csv	6/8/2016 3:21 PM	Microsoft Excel C...	1 KB
CDC-Grants-20102015.csv	6/8/2016 1:18 PM	Microsoft Excel C...	13,722 KB
CDC-Grants-20102015-EnviromentalHeal...	6/8/2016 1:19 PM	Microsoft Excel C...	502 KB
CDC-Grants-20102015-Infectious.csv	6/8/2016 1:26 PM	Microsoft Excel C...	3,455 KB
CDC-Grants-20102015-InjuryPrevention.c...	6/8/2016 1:22 PM	Microsoft Excel C...	655 KB

Load Dialog:

The file 'C:\Users\mginda\Desktop\Sci2-CDC-data\Burst\CDC-Grants-20102015-EnviromentalHealth.csv' can be loaded. Please select the format you would like to try.

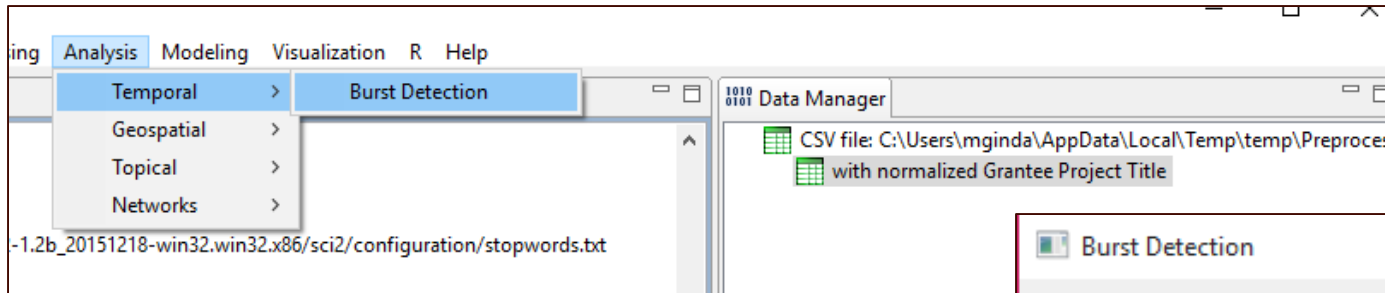
- Standard csv format
- NSF csv format
- Scopus csv format

Load ***CDC-Grants-20102015-EnviromentalHealth.csv***
 Located in CDC data directory: Sci2-CDC-data-> burst



Select the loaded data file in the Data Manager and then navigate to *Preprocessing* > *Topical* > *Lowercase, Tokenize, Stem, and Stopword Text*

Then select **Grantee Project Title** from the input parameters



Highlight the table 'with normalized Grantee Project Title'

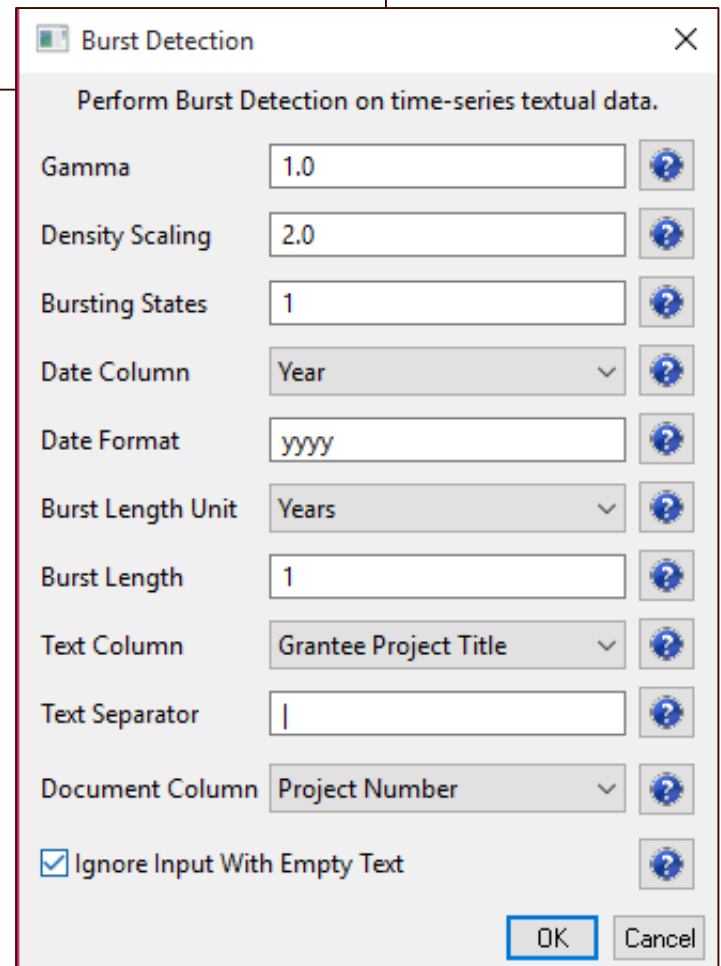
Select *Analysis > Temporal > Burst Detection*

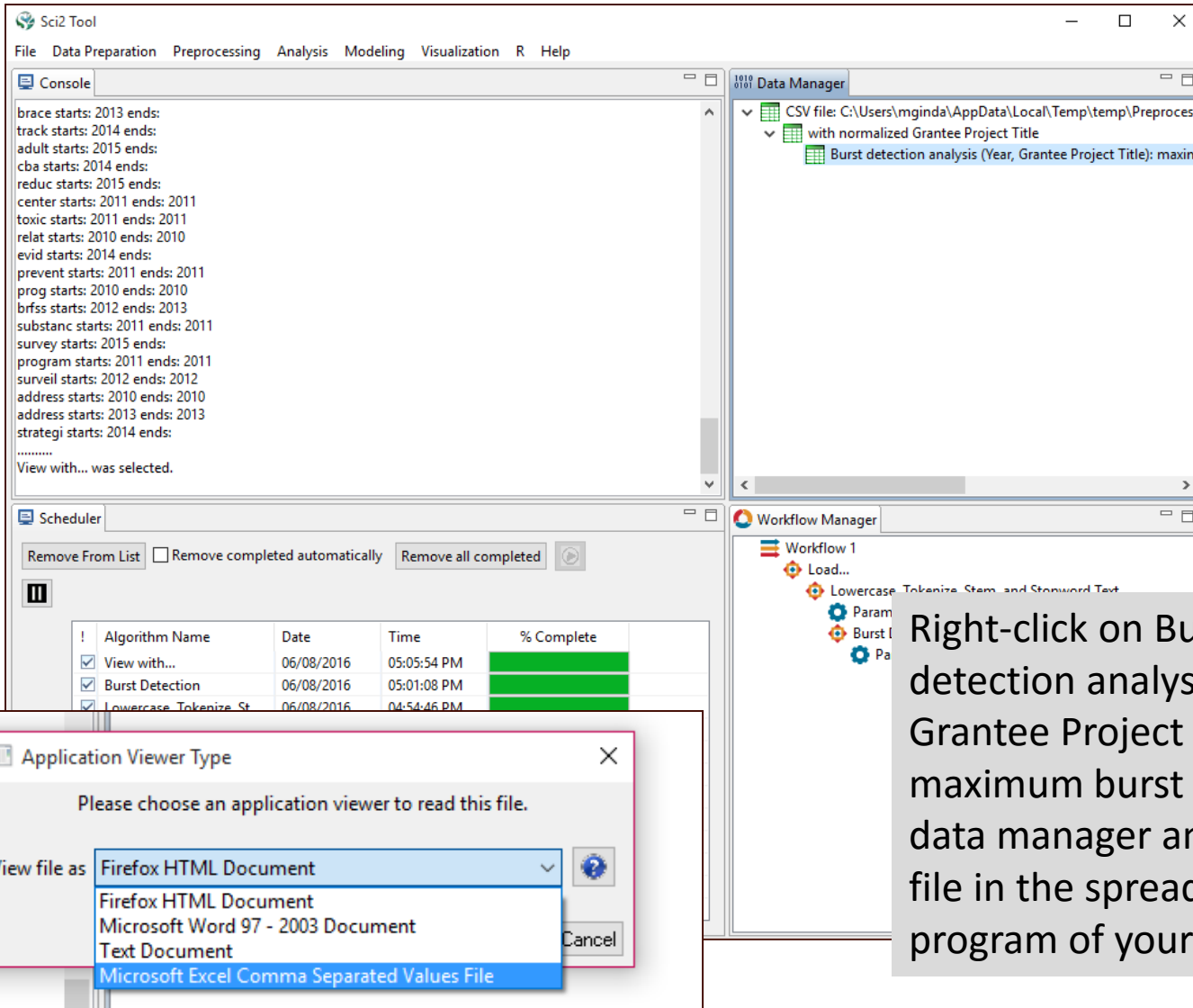
For our initial burst analysis we will use the base **Gamma** and **Density Scaling** Parameters.

For the **Date Column** select "Year"

For **Text Column** select "Grant Project Title"

For **Document Column** select "Project Number"





The screenshot shows the Sci2 Tool interface with several panels:

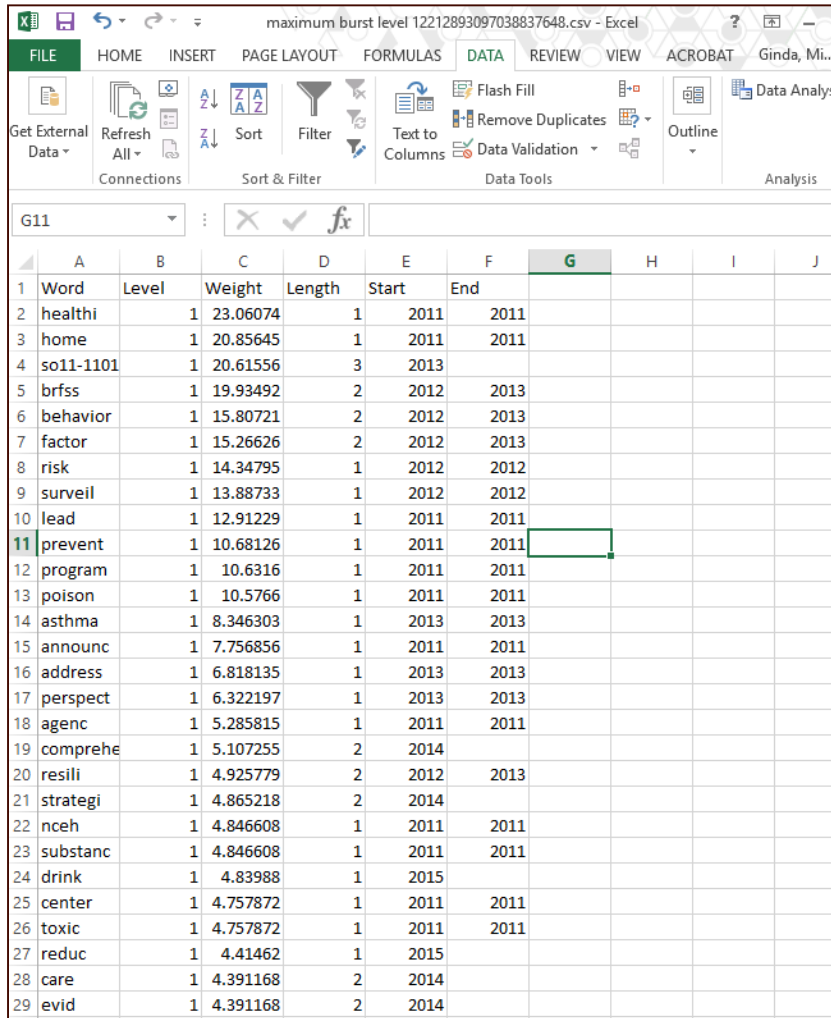
- Console:** Lists various datasets with their start and end years, such as "brace starts: 2013 ends:", "track starts: 2014 ends:", "adult starts: 2015 ends:", "cba starts: 2014 ends:", "reduc starts: 2015 ends:", "center starts: 2011 ends: 2011", "toxic starts: 2011 ends: 2011", "relat starts: 2010 ends: 2010", "evind starts: 2014 ends:", "prevent starts: 2011 ends: 2011", "prog starts: 2010 ends: 2010", "brfss starts: 2012 ends: 2013", "substanc starts: 2011 ends: 2011", "survey starts: 2015 ends:", "program starts: 2011 ends: 2011", "surveil starts: 2012 ends: 2012", "address starts: 2010 ends: 2010", "address starts: 2013 ends: 2013", "strategi starts: 2014 ends:". Below the list, it says "..... View with... was selected."
- Data Manager:** Shows a tree view of data files. The selected file is "Burst detection analysis (Year, Grantee Project Title): maximum burst level 1 in the data manager and view the file in the spreadsheet program of your choice".
- Scheduler:** Contains a table of tasks and their completion status.
- Workflow Manager:** Shows a workflow named "Workflow 1" with steps like "Load...", "Lowercase, Tokenize, Stem, and Stopword Text", "Param...", "Burst I...", and "Pa...".

An "Application Viewer Type" dialog box is open in the foreground, prompting the user to choose an application viewer to read the file. The dialog has a "View file as" dropdown menu with the following options:

- Firefox HTML Document
- Firefox HTML Document
- Microsoft Word 97 - 2003 Document
- Text Document
- Microsoft Excel Comma Separated Values File

The "Microsoft Excel Comma Separated Values File" option is currently selected.

Right-click on Burst detection analysis (Year, Grantee Project Title): maximum burst level 1 in the data manager and view the file in the spreadsheet program of your choice



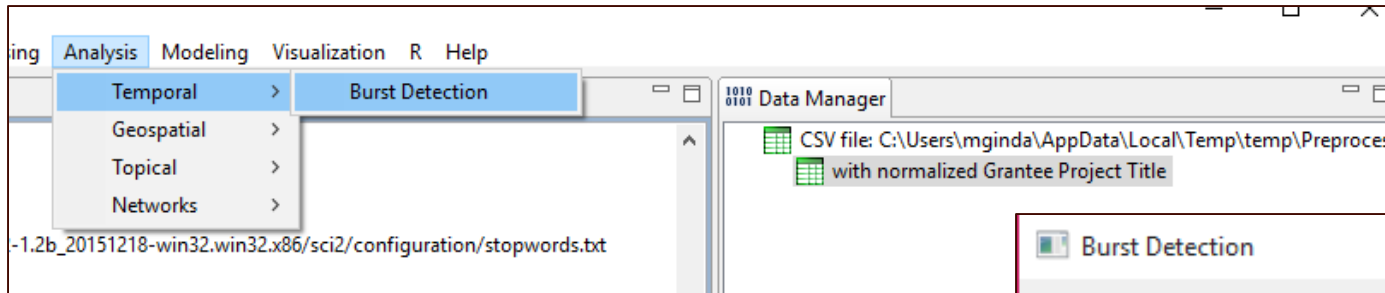
	A	B	C	D	E	F	G	H	I	J
1	Word	Level	Weight	Length	Start	End				
2	healthi	1	23.06074	1	2011	2011				
3	home	1	20.85645	1	2011	2011				
4	so11-1101	1	20.61556	3	2013					
5	brfss	1	19.93492	2	2012	2013				
6	behavior	1	15.80721	2	2012	2013				
7	factor	1	15.26626	2	2012	2013				
8	risk	1	14.34795	1	2012	2012				
9	surveil	1	13.88733	1	2012	2012				
10	lead	1	12.91229	1	2011	2011				
11	prevent	1	10.68126	1	2011	2011				
12	program	1	10.6316	1	2011	2011				
13	poison	1	10.5766	1	2011	2011				
14	asthma	1	8.346303	1	2013	2013				
15	announc	1	7.756856	1	2011	2011				
16	address	1	6.818135	1	2013	2013				
17	perspect	1	6.322197	1	2013	2013				
18	agenc	1	5.285815	1	2011	2011				
19	comprehe	1	5.107255	2	2014					
20	resili	1	4.925779	2	2012	2013				
21	strategi	1	4.865218	2	2014					
22	nceh	1	4.846608	1	2011	2011				
23	substanc	1	4.846608	1	2011	2011				
24	drink	1	4.83988	1	2015					
25	center	1	4.757872	1	2011	2011				
26	toxic	1	4.757872	1	2011	2011				
27	reduc	1	4.41462	1	2015					
28	care	1	4.391168	2	2014					
29	evid	1	4.391168	2	2014					

As before, the missing end dates indicate the continuation of a burst in a given data set.

Before we add the End date of 2015 to those records missing a value, lets review the results.

What are the primary bursting terms?

How many are there, and tuning the parameter values help reduce the number of terms returned?



Highlight the table ‘with normalized Grantee Project Title’

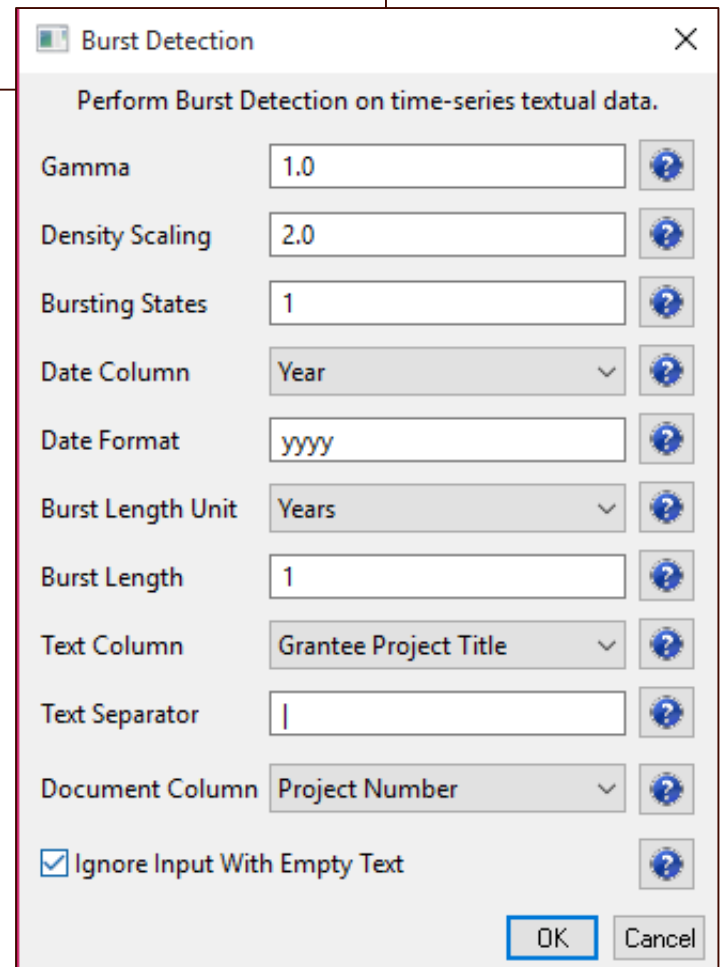
Select *Analysis > Temporal > Burst Detection*

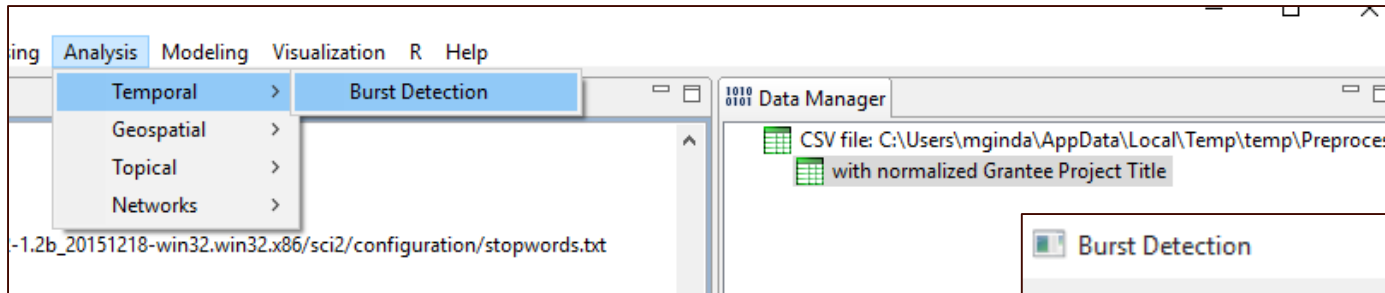
For our initial burst analysis we will use the base **Gamma** and **Density Scaling** Parameters.

For the **Date Column** select “Year”

For **Text Column** select “Grant Project Title”

For **Document Column** select “Project Number”





Highlight the table ‘with normalized Grantee Project Title’

Select *Analysis > Temporal > Burst Detection*

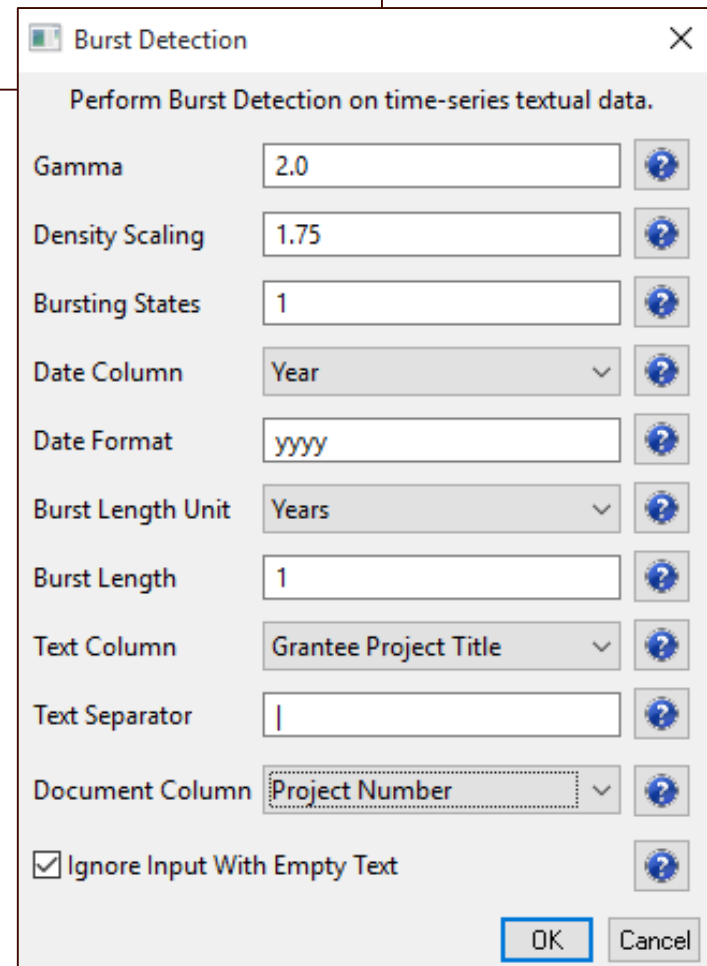
For the final burst analysis use the base **Gamma** enter “2.0”

For **Density Scaling** enter “1.75”

For the **Date Column** select “Year”

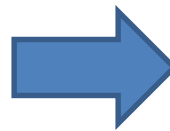
For **Text Column** select “Grant Project Title”

For **Document Column** select “Project Number”



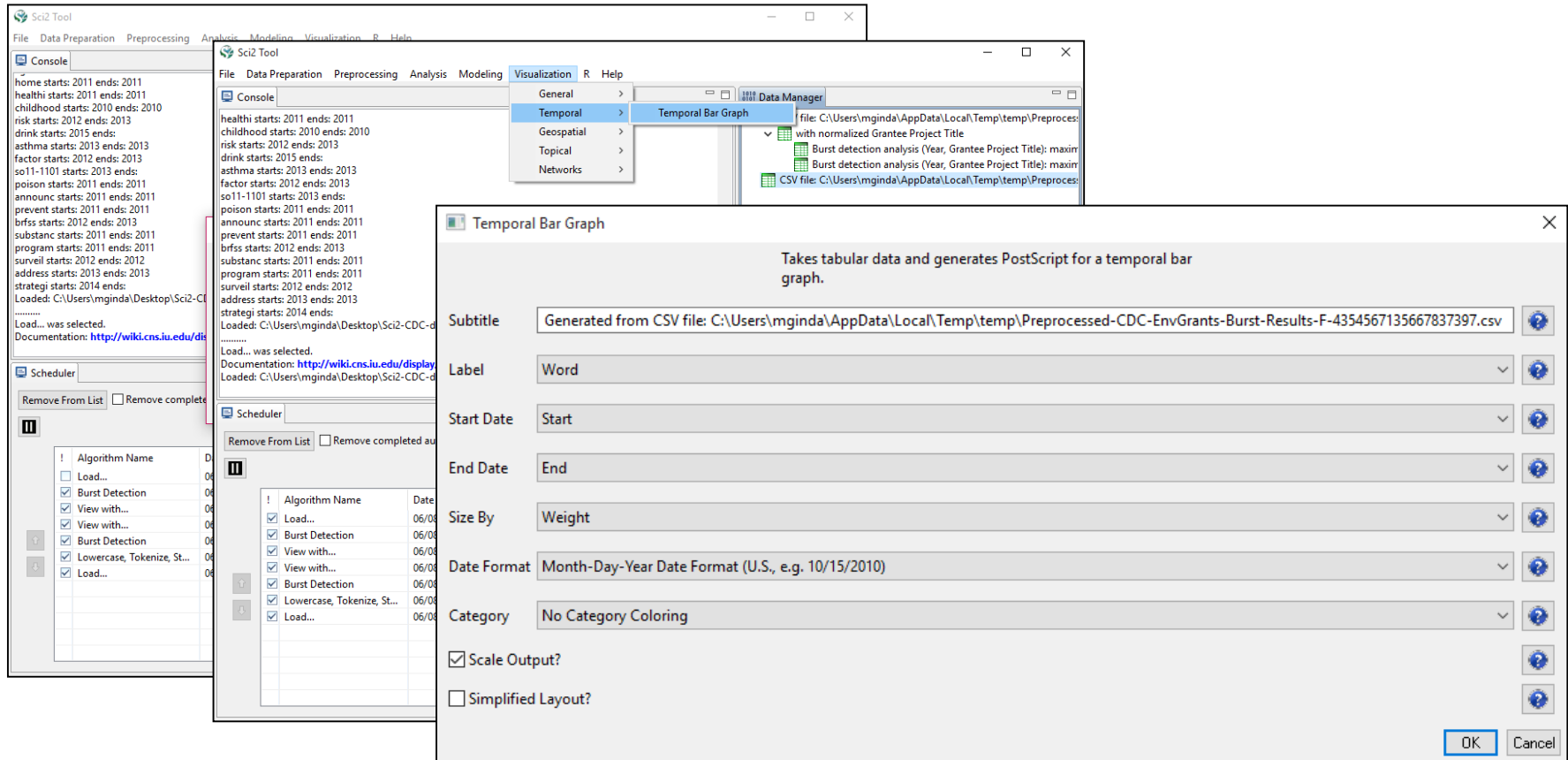
Missing end dates indicate the continuation of a burst in a given data set. Add the End date of 2015 to those records missing and End date.

	A	B	C	D	E	F
1	Word	Level	Weight	Length	Start	End
2	perspect	1	5.43949	1	2013	2013
3	behavior	1	15.12261	2	2012	2013
4	resili	1	4.874477	4	2012	
5	comprehe	1	4.397597	2	2014	
6	lead	1	11.27075	1	2011	2011
7	nkeh	1	3.953627	1	2011	2011
8	agenc	1	4.323344	1	2011	2011
9	home	1	17.00882	1	2011	2011
10	healthi	1	18.82999	1	2011	2011
11	childhood	1	4.011145	1	2010	2010
12	risk	1	12.29086	2	2012	2013
13	drink	1	4.114564	1	2015	
14	asthma	1	7.934085	1	2013	2013
15	factor	1	14.74289	2	2012	2013
16	so11-1101	1	18.24718	3	2013	
17	poison	1	9.396071	1	2011	2011
18	announc	1	6.423941	1	2011	2011
19	prevent	1	9.470077	1	2011	2011
20	brfss	1	17.50042	2	2012	2013
21	substanc	1	3.953627	1	2011	2011
22	program	1	10.39794	1	2011	2011
23	surveil	1	11.64216	1	2012	2012
24	address	1	6.092743	1	2013	2013
25	strategi	1	4.131878	2	2014	
26						



	A	B	C	D	E	F
1	Word	Level	Weight	Length	Start	End
2	perspect	1	5.43949	1	2013	2013
3	behavior	1	15.12261	2	2012	2013
4	resili	1	4.874477	4	2012	2015
5	comprehe	1	4.397597	2	2014	2015
6	lead	1	11.27075	1	2011	2011
7	nkeh	1	3.953627	1	2011	2011
8	agenc	1	4.323344	1	2011	2011
9	home	1	17.00882	1	2011	2011
10	healthi	1	18.82999	1	2011	2011
11	childhood	1	4.011145	1	2010	2010
12	risk	1	12.29086	2	2012	2013
13	drink	1	4.114564	1	2015	2015
14	asthma	1	7.934085	1	2013	2013
15	factor	1	14.74289	2	2012	2013
16	so11-1101	1	18.24718	3	2013	2015
17	poison	1	9.396071	1	2011	2011
18	announc	1	6.423941	1	2011	2011
19	prevent	1	9.470077	1	2011	2011
20	brfss	1	17.50042	2	2012	2013
21	substanc	1	3.953627	1	2011	2011
22	program	1	10.39794	1	2011	2011
23	surveil	1	11.64216	1	2012	2012
24	address	1	6.092743	1	2013	2013
25	strategi	1	4.131878	2	2014	2015

Save the file as a .CSV file and load it back into Sci2,
selecting the Standard CSV format

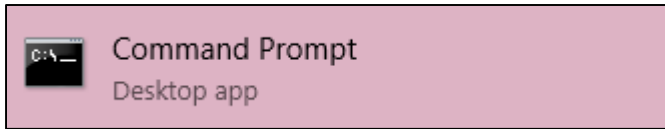


The screenshot shows the Sci2 Tool interface with the 'Temporal Bar Graph' dialog box open. The dialog box contains the following configuration options:

- Title:** Takes tabular data and generates PostScript for a temporal bar graph.
- Subtitle:** Generated from CSV file: C:\Users\mginda\AppData\Local\Temp\Preprocessed-CDC-EnvGrants-Burst-Results-F-4354567135667837397.csv
- Label:** Word
- Start Date:** Start
- End Date:** End
- Size By:** Weight
- Date Format:** Month-Day-Year Date Format (U.S., e.g. 10/15/2010)
- Category:** No Category Coloring
- Scale Output?
- Simplified Layout?

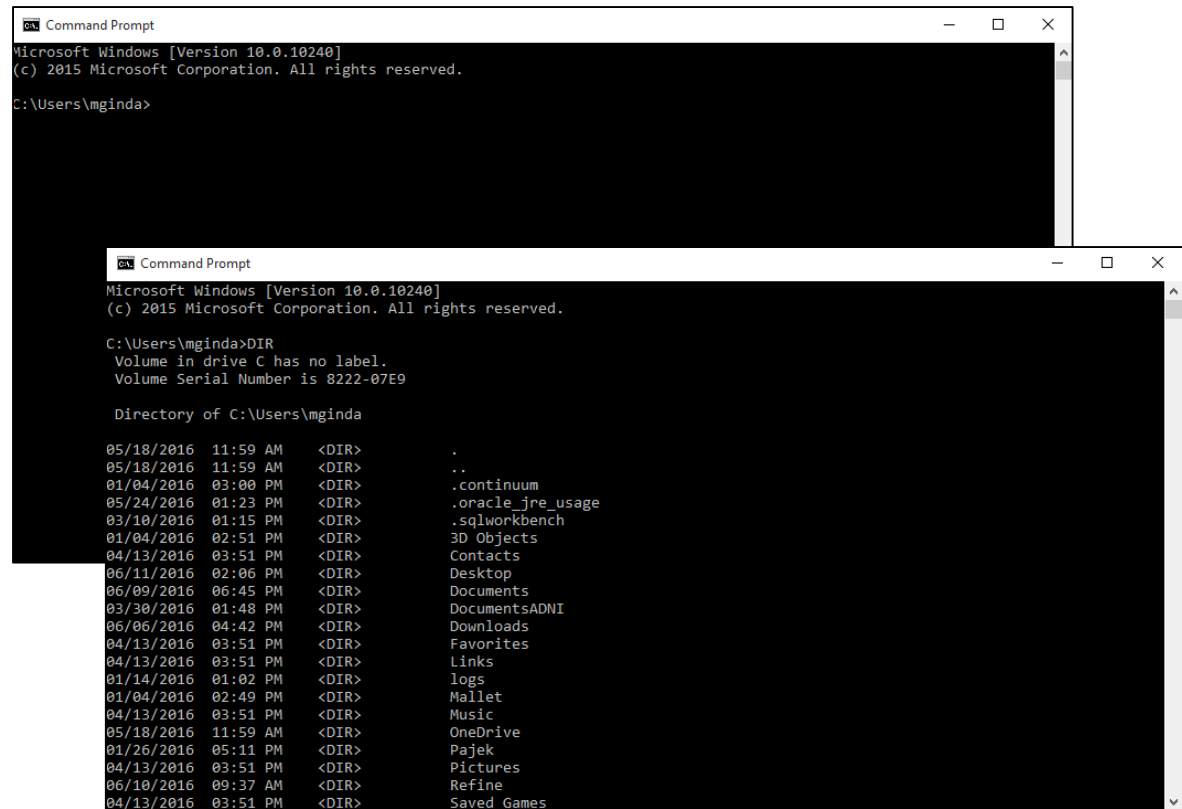
The background shows the Sci2 Tool interface with the 'Visualization' menu open, highlighting 'Temporal Bar Graph'. The 'Data Manager' window shows a list of files, including the CSV file used in the dialog box. The 'Console' window displays a list of CDC grants with their start and end dates.

- Load updated Burst Detection result file into Sci2 as a standard CSV format.
- Select the newly loaded file in the data manager and the *Visualization > Temporal > Temporal Bar Graph* in the menu bar.
- Set the parameter values to those shown to the right



From the Windows Start menu, search for CMD, or Command Prompt.

Open a new window, and enter the DIR to locate your current directory. You will need to navigate to the directory where you saved your burst detection results PostScript files.



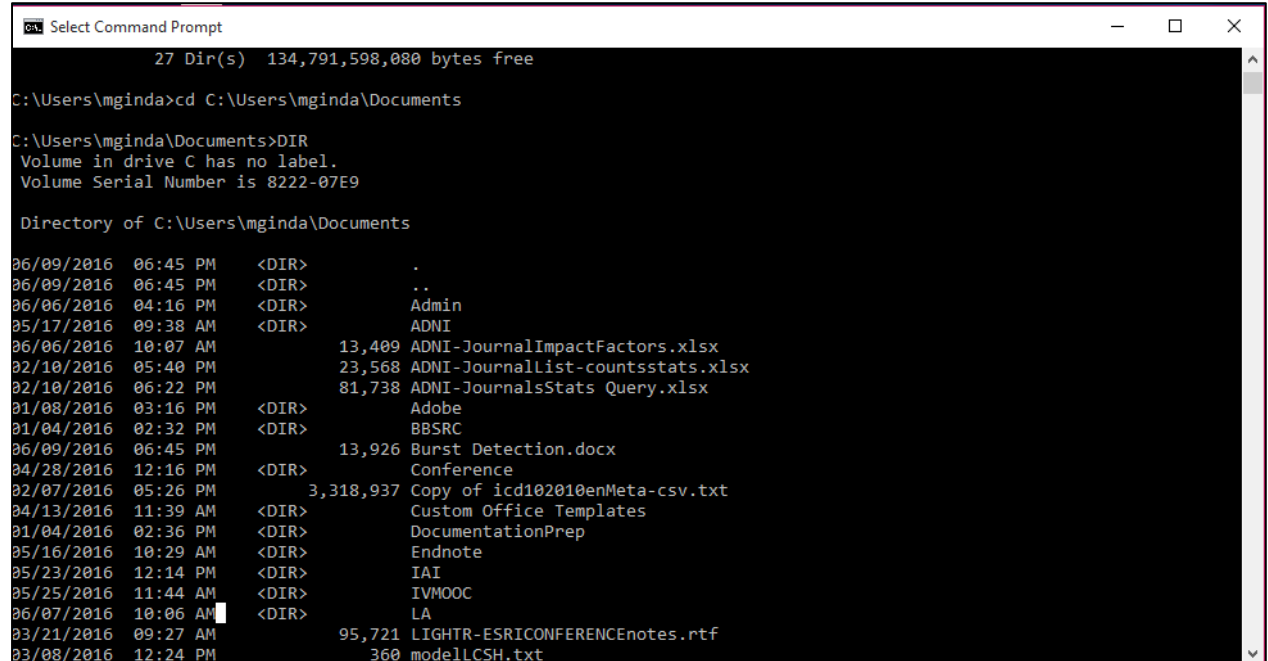
A common place to save a file may be the Desktop or the Documents folder.

To navigate to the directory you saved your PostScript file(s) in, enter

```
cd C:\Users\[username]\Desktop
```

or

```
cd C:\Users\[username]\Documents
```



```
Select Command Prompt
27 Dir(s) 134,791,598,080 bytes free

C:\Users\mginda>cd C:\Users\mginda\Documents

C:\Users\mginda\Documents>DIR
Volume in drive C has no label.
Volume Serial Number is 8222-07E9

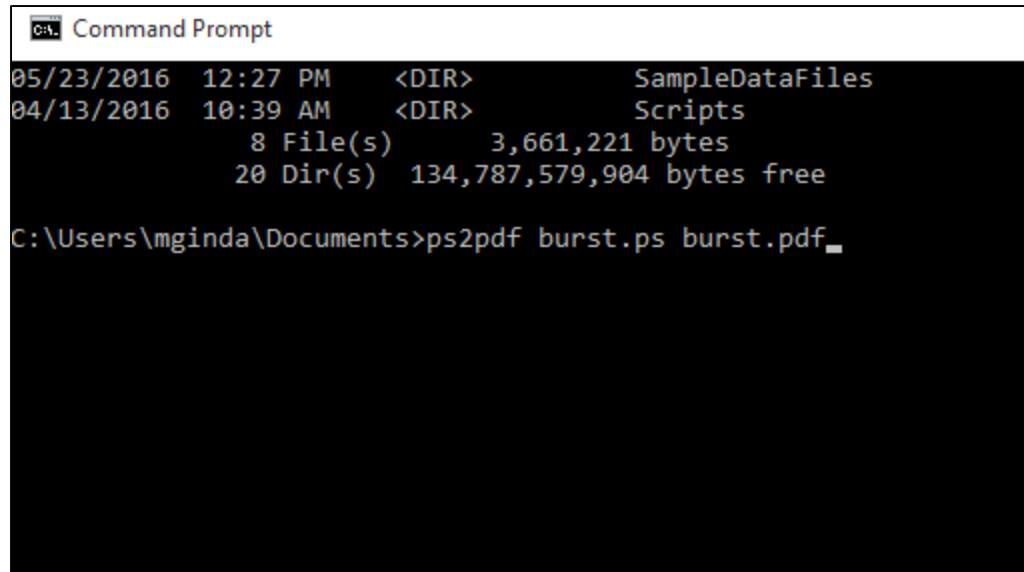
Directory of C:\Users\mginda\Documents

06/09/2016  06:45 PM    <DIR>          .
06/09/2016  06:45 PM    <DIR>          ..
06/06/2016  04:16 PM    <DIR>          Admin
05/17/2016  09:38 AM    <DIR>          ADNI
06/06/2016  10:07 AM                13,409 ADNI-JournalImpactFactors.xlsx
02/10/2016  05:40 PM                23,568 ADNI-JournalList-countsstats.xlsx
02/10/2016  06:22 PM                81,738 ADNI-JournalsStats Query.xlsx
01/08/2016  03:16 PM    <DIR>          Adobe
01/04/2016  02:32 PM    <DIR>          BBSRC
06/09/2016  06:45 PM                13,926 Burst Detection.docx
04/28/2016  12:16 PM    <DIR>          Conference
02/07/2016  05:26 PM          3,318,937 Copy of icd102010enMeta-csv.txt
04/13/2016  11:39 AM    <DIR>          Custom Office Templates
01/04/2016  02:36 PM    <DIR>          DocumentationPrep
05/16/2016  10:29 AM    <DIR>          Endnote
05/23/2016  12:14 PM    <DIR>          IAI
05/25/2016  11:44 AM    <DIR>          IVMOOC
06/07/2016  10:06 AM    <DIR>          LA
03/21/2016  09:27 AM                95,721 LIGHTR-ESRICONFERENCEnotes.rtf
03/08/2016  12:24 PM                360  modelLCSH.txt
```

Once you have navigated to the proper directory, you can run the PS2PDF program. In the command line run:

```
ps2pdf [options] input.[e]ps output.pdf
```

If you leave off the output file name, the resulting PDF file will have the same name as the original file with a .PDF extension into the same directory you navigated to.



```
Command Prompt
05/23/2016 12:27 PM <DIR> SampleDataFiles
04/13/2016 10:39 AM <DIR> Scripts
           8 File(s)      3,661,221 bytes
          20 Dir(s)  134,787,579,904 bytes free

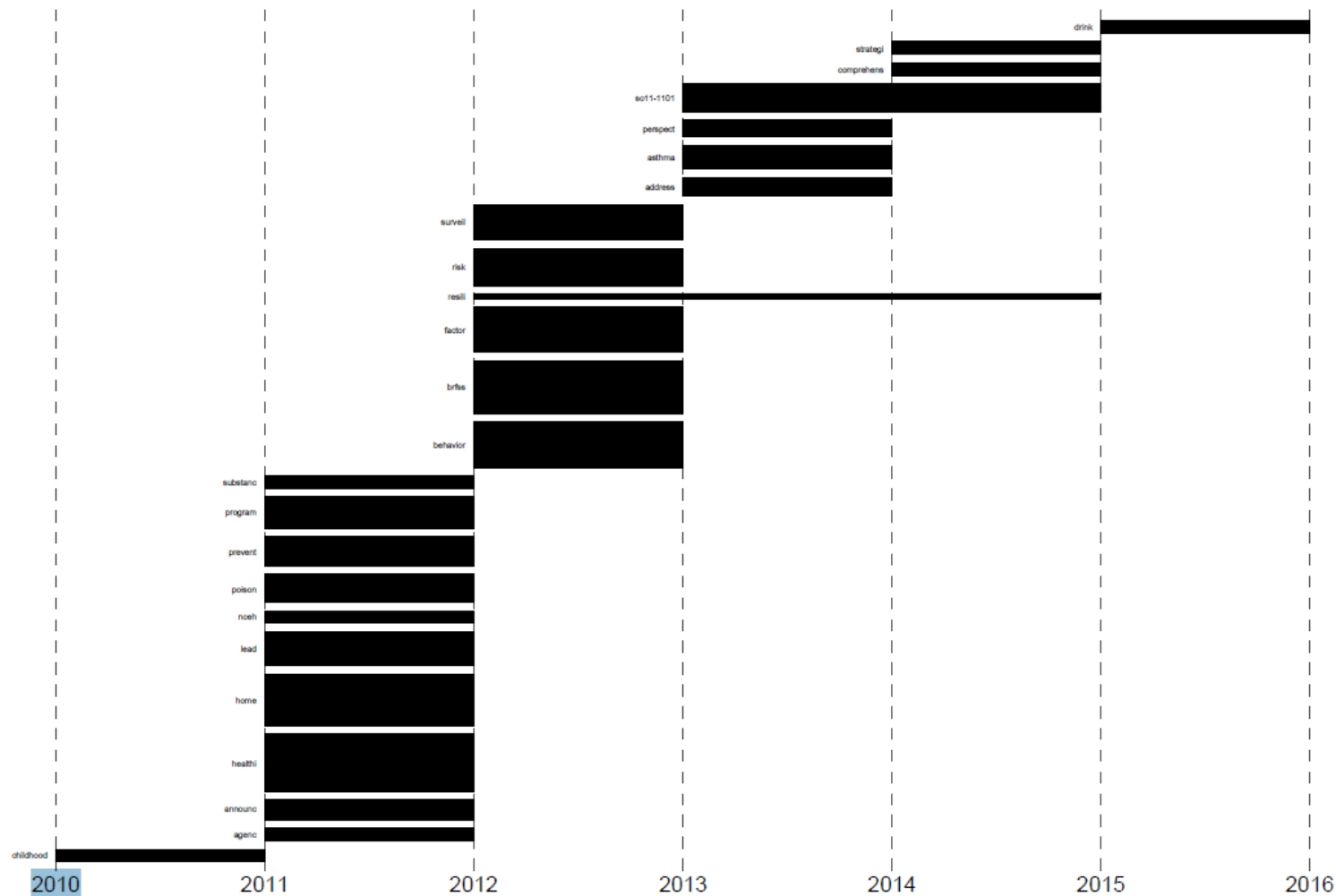
C:\Users\mginda\Documents>ps2pdf burst.ps burst.pdf_
```

You can view the PDF file with Adobe Reader. More information on this script tool is available in the [PS2PDF documentation site](#).

Temporal Visualization

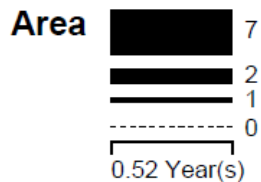
(Generated from CSV file: C:\Users\mginda\AppData\Local\Temp\temp\Preprocessed-CDC-EnvGrants-Burst-Results-F-4354567135667837397.csv)

June 08, 2016 | 5:43 PM EDT



Legend

Area size: Weight
 Minimum = 4
 Maximum = 19
 Text label: Word



How To Read This Map

This *temporal bar graph* visualization represents each record as a horizontal bar with a specific start and end date and a text label on its left side. The area of each bar encodes a numerical attribute value, e.g., total amount of funding. Bars may be colored to present categorical attribute values of records.

Questions?

8:00 Welcome and Overview of Tutorial and Attendees**8:30 Visualization Framework and Workflow Design**

- Overview of the Visualization framework
- Overview of Graphical variables, and color selection

9:15 Tool Overview and Trouble Shooting

- Sci2 Overview – scientometric analysis tool
- Open Refine – data parsing, transformation, and editing tool
- Gephi – network visualization

9:45 Geospatial Analysis

- Overview of geospatial analysis and mapping
- Geospatial Analysis: Geocoding with OpenRefine
- Geospatial Analysis: Proportional Symbol Map using CDC

11:00 Topical/Temporal Analysis: Burst Detection

- Overview of burst analysis and introductory workflow
- Burst Detection with CDC Grants

12:30 Lunch**1:30 Network Analysis**

- **Introduction to Network Analysis**
- Network Analysis: Co-authorship Network with CDC publications
- ~~Network Analysis: Bimodal networks with Morbidity Data~~

4:00 Wrap-up

4:30 Adjourn

What is a Network?

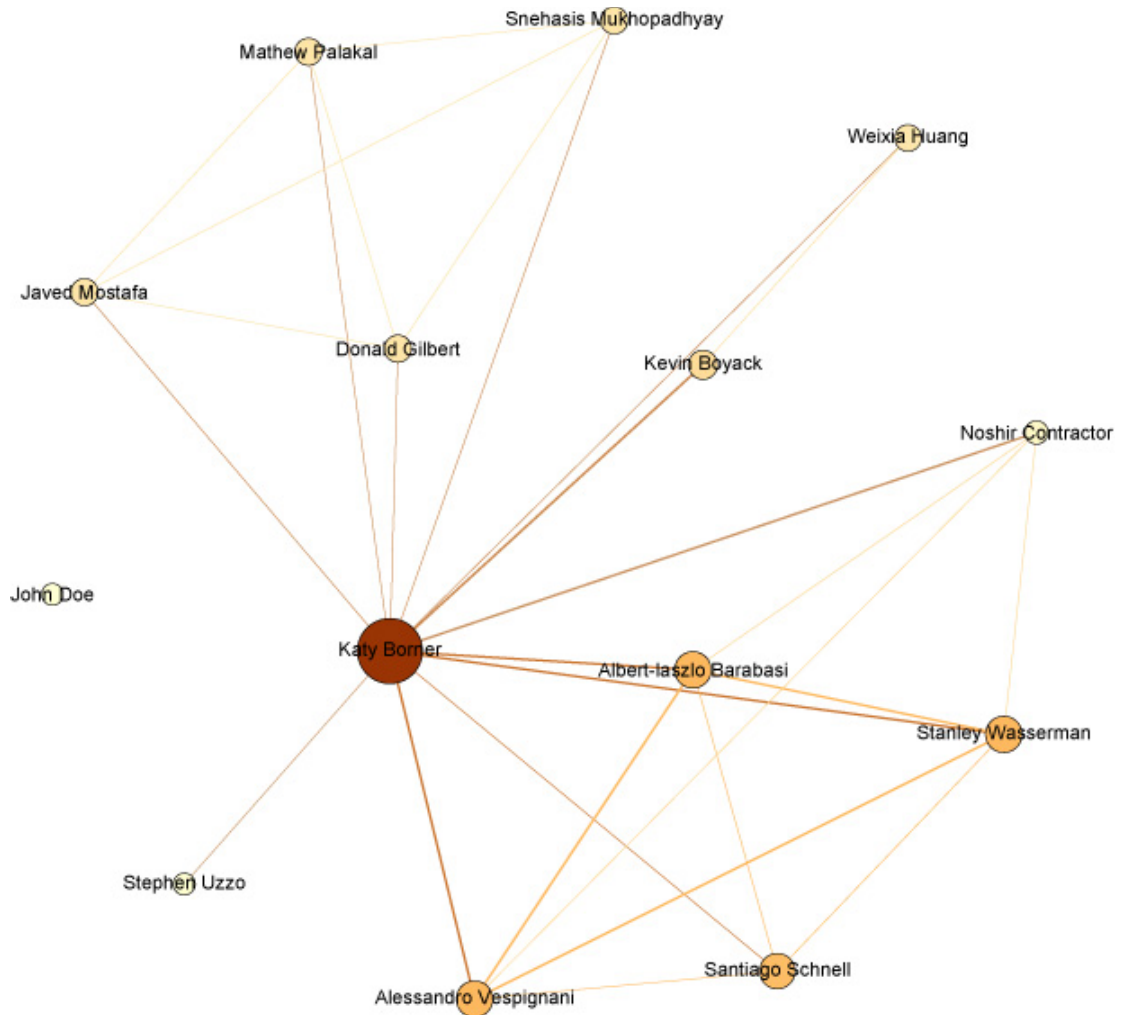
- Graph – network visualized
- Nodes
- Edges
- Components

Representations

- Matrices
- Graphs
- Edge and Node Lists

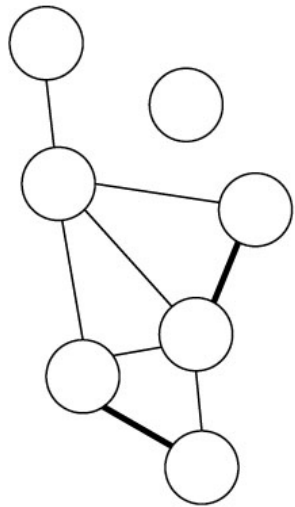
Data Formats

- Tabular
- XML
- Text
- JSON



General types of networks

Undirected Networks



Nodes:



Edges:



Node Degree:

Number of edges
connected to nodes

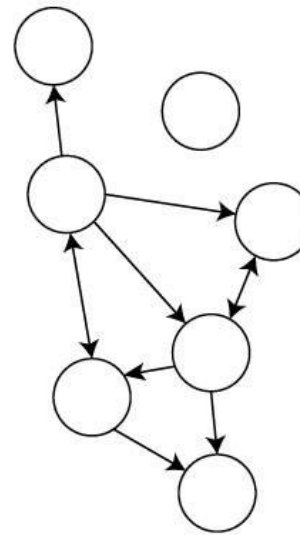
Isolates:

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

Other types of networks and graphs:

- Hierarchical networks (tree networks)
- Bipartite Networks
- Multigraphs
- Hypergraphs

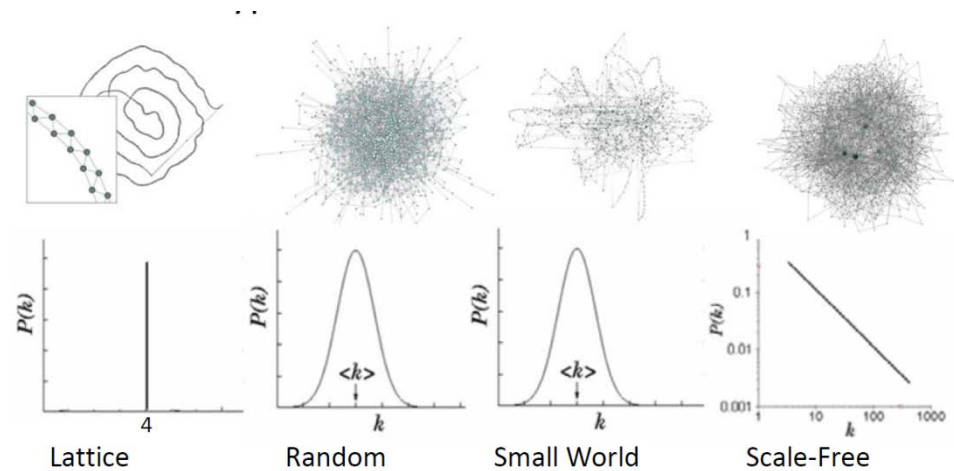
Graph Features

General Topologies

- Random Graphs network
- Watts-Strogatz // Small World network
 - gene networks, food chains, voter networks, power grids
- Barabasi-Albert Scale Free network
 - Internet, Citation Networks, Social Network

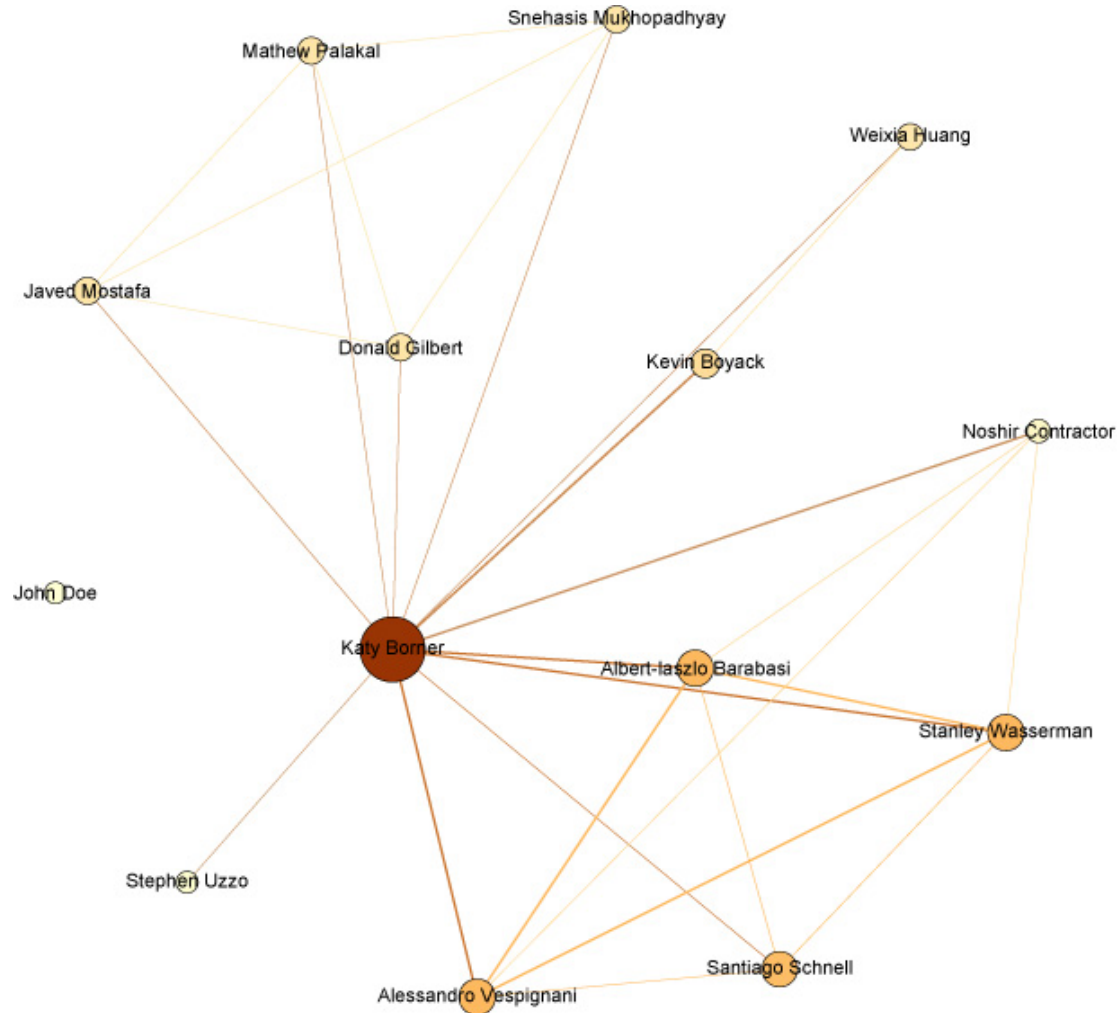
Measurements

- Node and Edge Counts
- Network Components
- Giant Component
- Avg. degree distribution
- Avg. Clustering
- Density
- Avg. Path Length
- Diameter



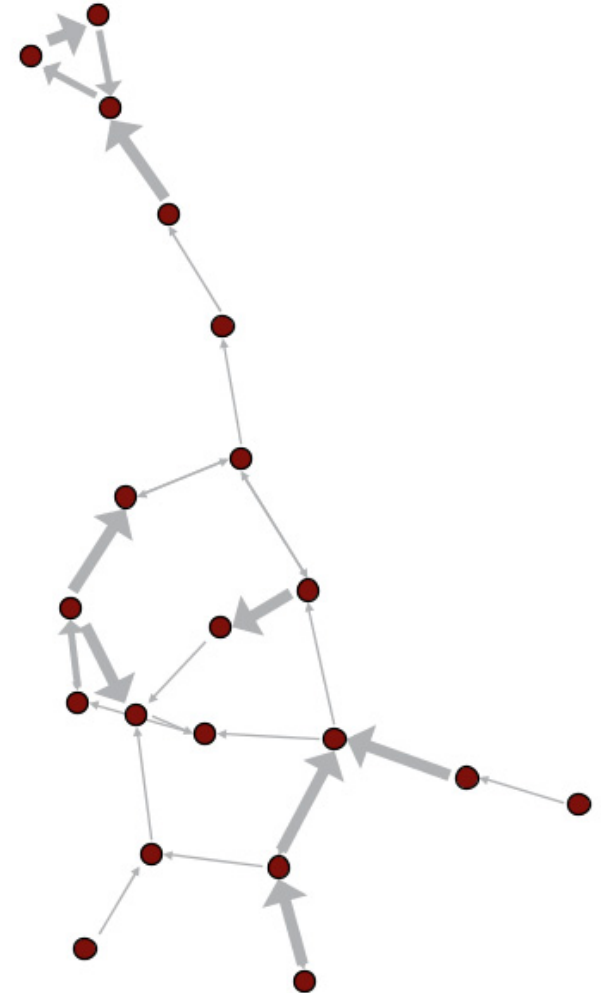
Node Metrics

- Degree
- Isolate nodes
- Degree Centrality
- Betweenness
- Closeness centrality



Graph Metrics - Edges

- Shortest paths – shortest distance between two nodes
- Weight – strength of tie
- Directionality – is the connection one-way or two-way (in-degree vs. out-degree)?
- Bridge – deleting would change structure



8:00 Welcome and Overview of Tutorial and Attendees**8:30 Visualization Framework and Workflow Design**

- Overview of the Visualization framework
- Overview of Graphical variables, and color selection

9:15 Tool Overview and Trouble Shooting

- Sci2 Overview – scientometric analysis tool
- Open Refine – data parsing, transformation, and editing tool
- Gephi – network visualization

9:45 Geospatial Analysis

- Overview of geospatial analysis and mapping
- Geospatial Analysis: Geocoding with OpenRefine
- Geospatial Analysis: Proportional Symbol Map using CDC

11:00 Topical/Temporal Analysis: Burst Detection

- Overview of burst analysis and introductory workflow
- Burst Detection with CDC Grants

12:30 Lunch**1:30 Network Analysis**

- Introduction to Network Analysis
- Network Analysis: Co-authorship Network with CDC publications
- ~~Network Analysis: Bimodal networks with Morbidity Data~~

4:00 Wrap-up

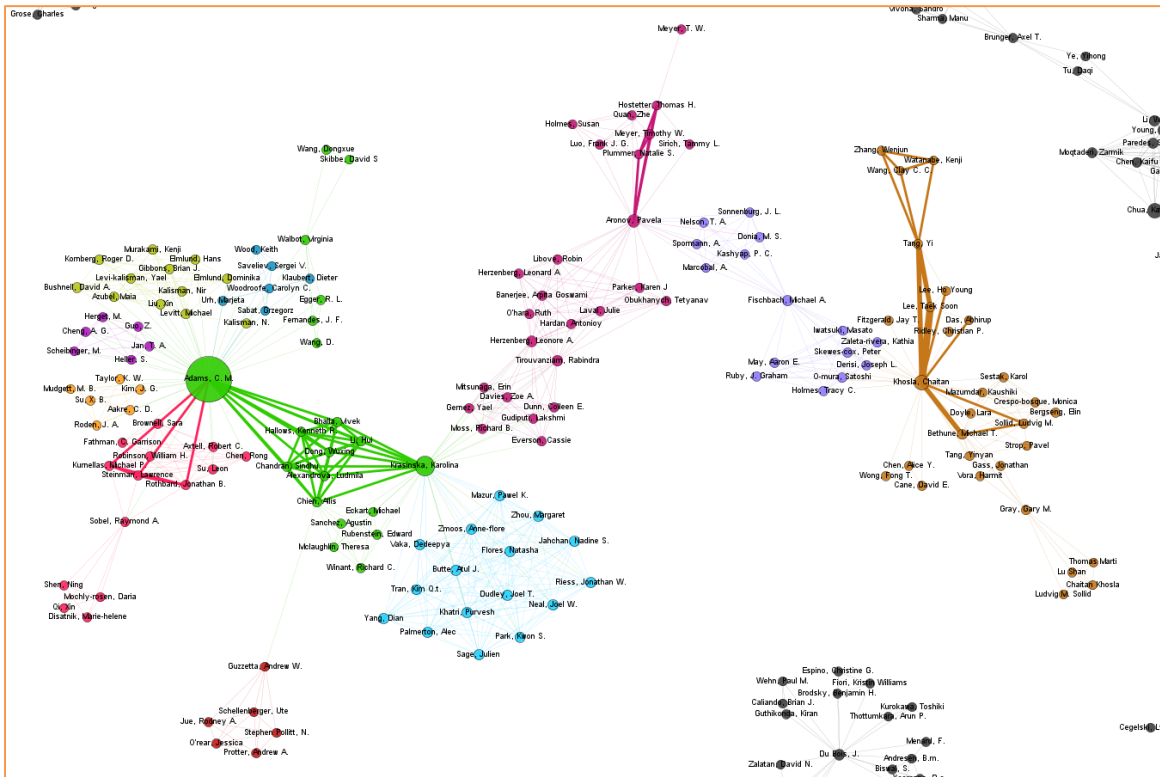
4:30 Adjourn

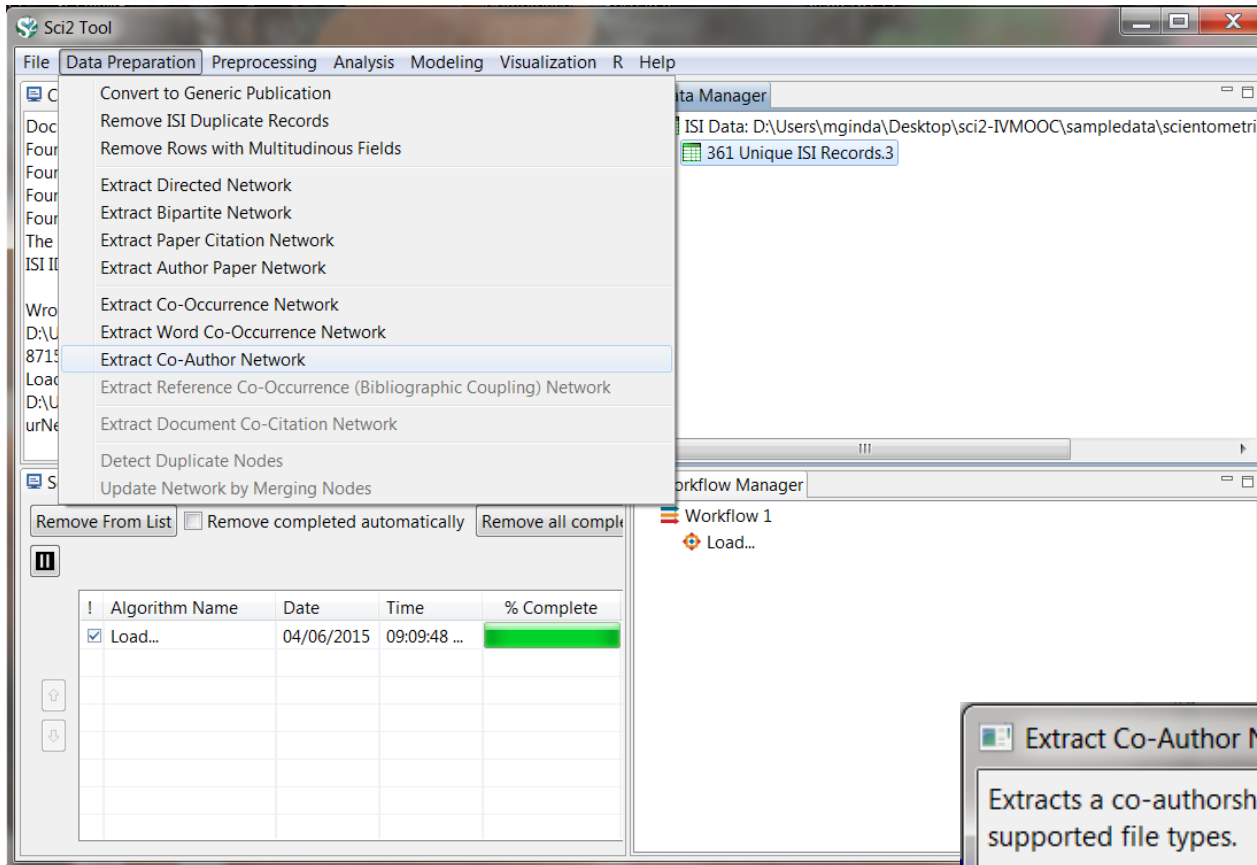
What is the purpose of looking at co-author networks? What can they tell us?

This is a visualization of a citation dataset from a researcher that administers a Core research facility at Stanford University.

The objective of the researcher who provided this data set was to understand

1. Which researchers using her lab were publishing articles?
2. Which researchers collaborate frequently in the facility?
3. Who has the most citation impact?

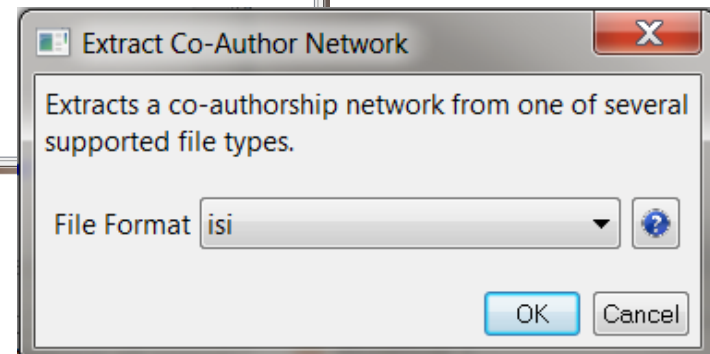


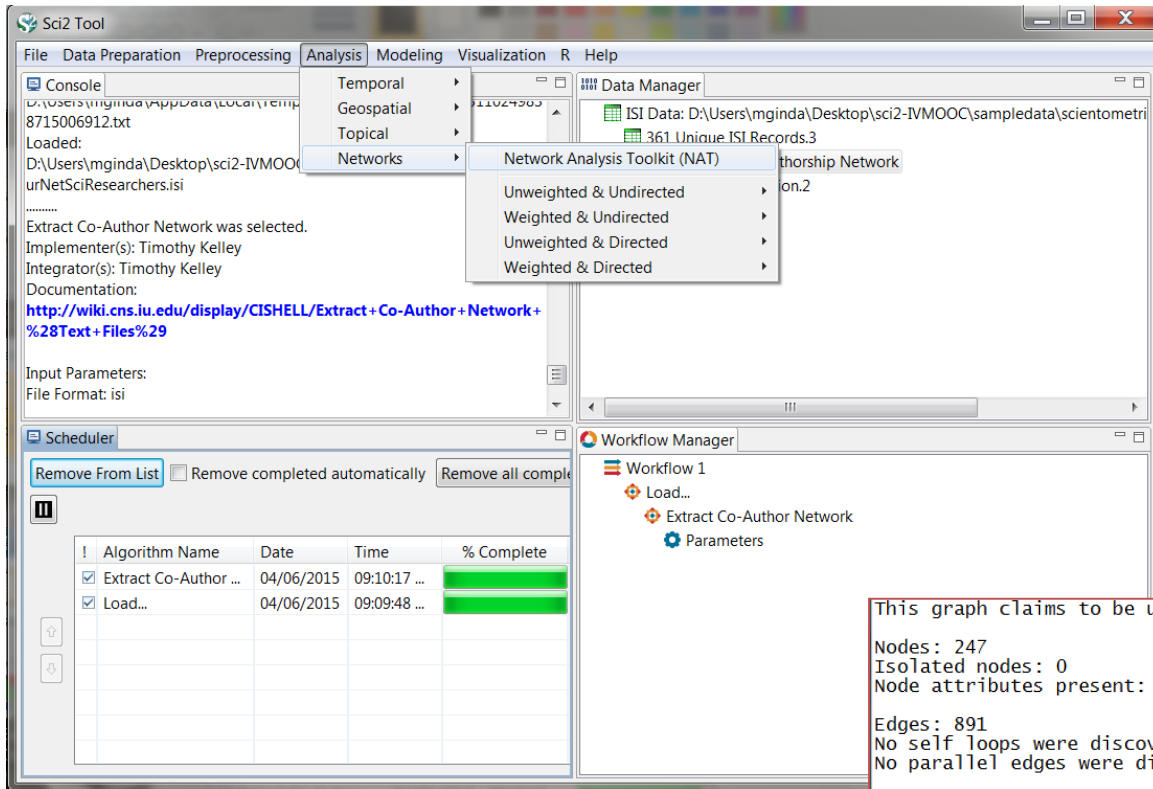


Load *FourNetSciResearchers.isi* located in Sci2 Directory

Select the data 361 Unique ISI Records in the Data manager and the algorithm *Extract Co-Author Network* in the Data Preparation menu.

A pop-up window will appear; select the format ISI.





After creating your initial network, it is a good idea to get a brief overview of its statistical properties.

Sci2 has a built-in analysis toolkit to perform these basic statistics.

Select the network output file in the data manager, and then in the menu select *Analysis -> Networks -> Network Analysis Toolkit (NAT)*

The output should read:

```

This graph claims to be undirected.
Nodes: 247
Isolated nodes: 0
Node attributes present: label, number_of_authored_works, times_cited

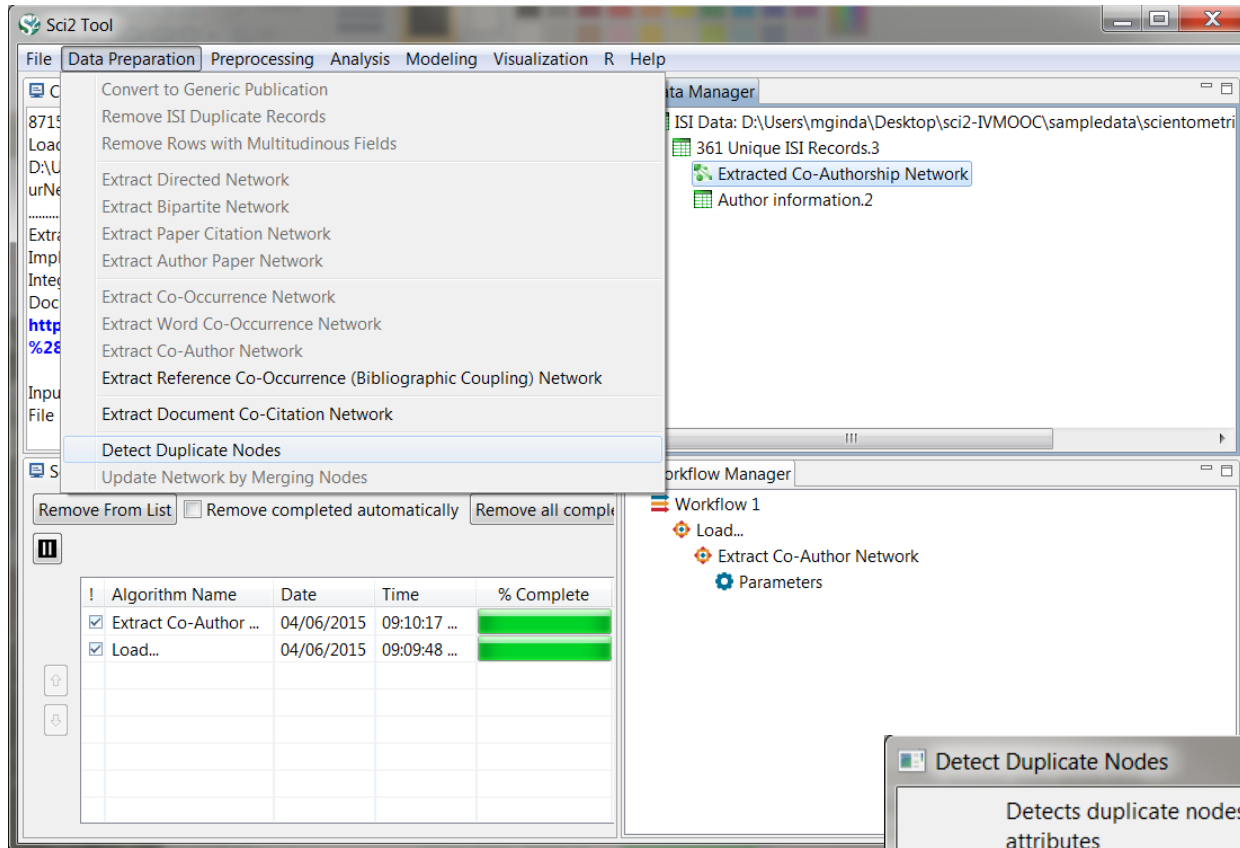
Edges: 891
No self loops were discovered.
No parallel edges were discovered.

Edge attributes:
Did not detect any nonnumeric attributes.
Numeric attributes:
           min    max    mean
number_...  1     33    1.76094
weight      1     33    1.76094

This network seems to be valued.

Average degree: 7.2146
This graph is not weakly connected.
There are 3 weakly connected components. (0 isolates)
The largest connected component consists of 194 nodes.
Did not calculate strong connectedness because this graph was not directed.

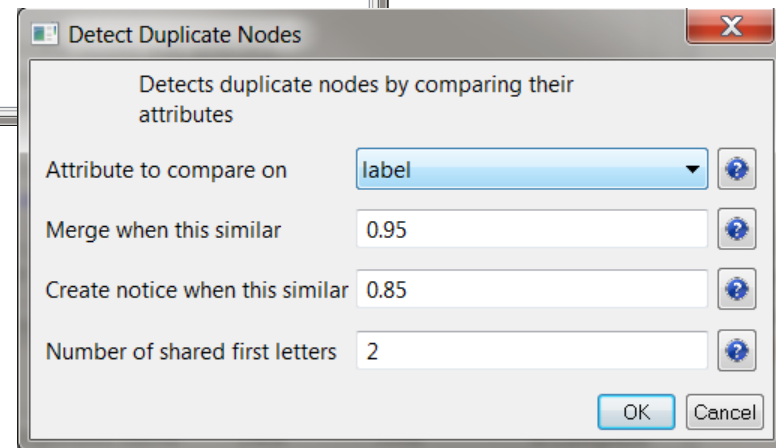
Density (disregarding weights): 0.0293
Additional Densities by Numeric Attribute
    
```

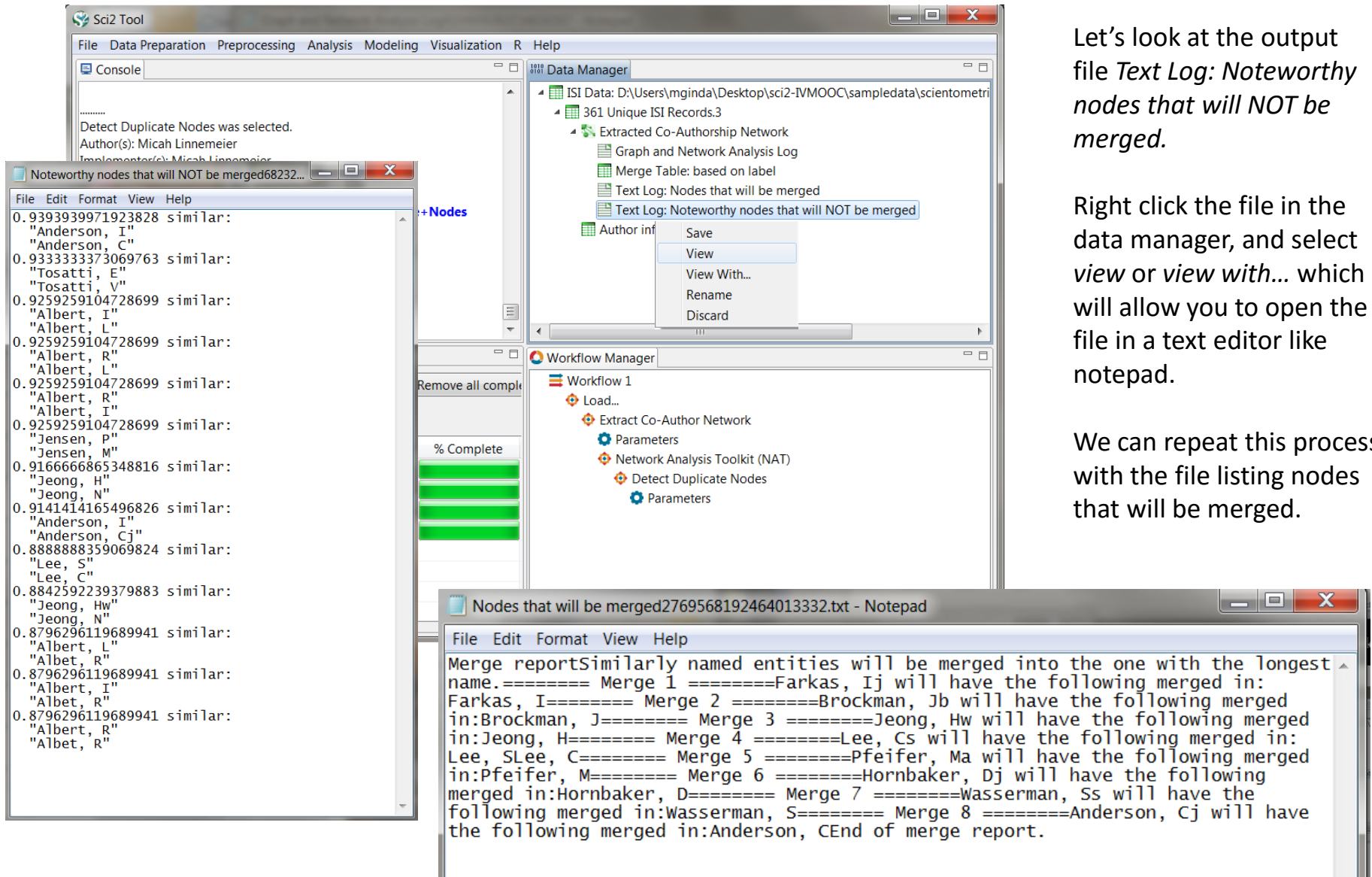


One of the challenges of a co-author network is determining if your data set has duplicate names (e.g. John P. Smith and J P Smith).

To detect duplicate nodes, we will want to select the network in the data manager, and then select *Data Preparation -> Detect Duplicate Nodes*.

A pop-up window will appear, for this demo we will keep the input parameters.





The screenshot displays the Sci2 Tool interface with several windows open:

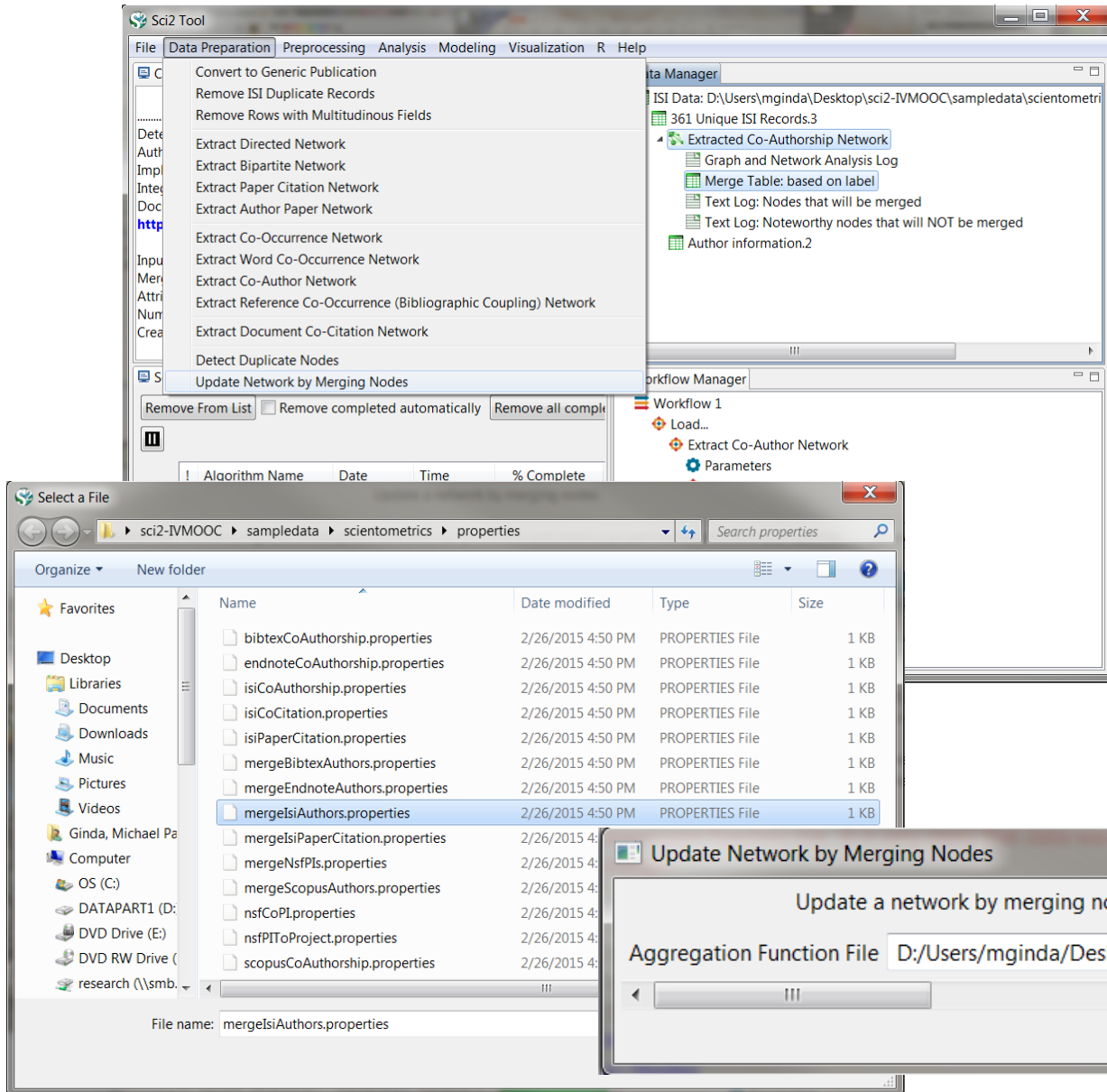
- Console:** Shows a message: "Detect Duplicate Nodes was selected. Author(s): Micah Linnemeier. Implementer(s): Micah Linnemeier."
- Data Manager:** Shows a tree view of data files. A context menu is open over the file "Text Log: Noteworthy nodes that will NOT be merged", with options: Save, View, View With..., Rename, and Discard.
- Workflow Manager:** Shows a workflow named "Workflow 1" with steps: Load..., Extract Co-Author Network (Parameters), Network Analysis Toolkit (NAT) (Parameters), and Detect Duplicate Nodes (Parameters).
- Noteworthy nodes that will NOT be merged68232...:** A list of nodes with similarity scores and names:
 - 0.9393939971923828 similar: "Anderson, I", "Anderson, C"
 - 0.9333333373069763 similar: "Tosatti, E", "Tosatti, V"
 - 0.9259259104728699 similar: "Albert, I", "Albert, L"
 - 0.9259259104728699 similar: "Albert, R", "Albert, L"
 - 0.9259259104728699 similar: "Albert, R", "Albert, I"
 - 0.9259259104728699 similar: "Jensen, P", "Jensen, M"
 - 0.9166666865348816 similar: "Jeong, H", "Jeong, N"
 - 0.9141414165496826 similar: "Anderson, I", "Anderson, Cj"
 - 0.8888888359069824 similar: "Lee, S", "Lee, C"
 - 0.8842592239379883 similar: "Jeong, Hw", "Jeong, N"
 - 0.8796296119689941 similar: "Albert, L", "Albet, R"
 - 0.8796296119689941 similar: "Albert, I", "Albet, R"
 - 0.8796296119689941 similar: "Albert, R", "Albet, R"
- Nodes that will be merged2769568192464013332.txt - Notepad:** Contains a merge report:


```
Merge reportSimilarly named entities will be merged into the one with the longest name.
===== Merge 1 =====Farkas, Ij will have the following merged in:
Farkas, I===== Merge 2 =====Brockman, Jb will have the following merged in:
Brockman, J===== Merge 3 =====Jeong, Hw will have the following merged in:
Jeong, H===== Merge 4 =====Lee, Cs will have the following merged in:
Lee, SLee, C===== Merge 5 =====Pfeifer, Ma will have the following merged in:
Pfeifer, M===== Merge 6 =====Hornbaker, Dj will have the following merged in:
Hornbaker, D===== Merge 7 =====Wasserman, Ss will have the following merged in:
Wasserman, S===== Merge 8 =====Anderson, Cj will have the following merged in:
Anderson, CEnd of merge report.
```

Let's look at the output file *Text Log: Noteworthy nodes that will NOT be merged*.

Right click the file in the data manager, and select *view* or *view with...* which will allow you to open the file in a text editor like notepad.

We can repeat this process with the file listing nodes that will be merged.

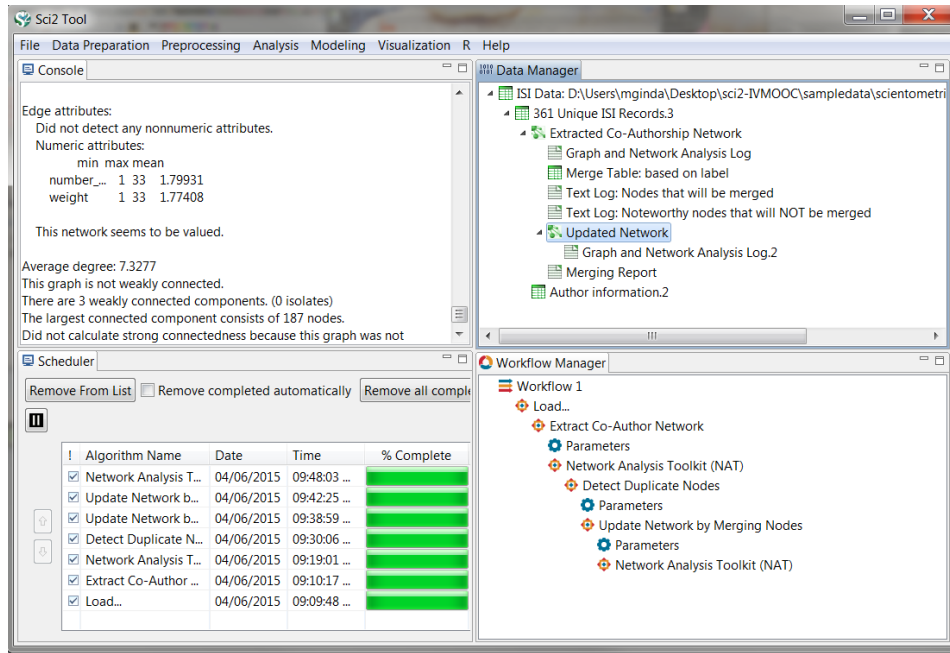


After we've identifying our duplicate nodes, we need to merge these duplicates.

Select the network file and the *Merge Table: based on label* file in the data manager, and then select *Data Preparation -> Update Network by Merging Nodes*

A box will appear that allows us to use an aggregation function file (property files). Select browse, and navigate to the *Sci2 directory sampledata -> scientometrics -> properties* and select *MergelsiAuthors.properties*

Select open, and then OK.



We've now updated out network, so lets re-run the network analysis toolkit algorithm to see how our network has been effected by our work.

What changes do you notice to the network statistics?

This graph claims to be undirected.

Original Network

Nodes: 247
Isolated nodes: 0
Node attributes present: label, number_of_authored_works, times_cited

Edges: 891
No self loops were discovered.
No parallel edges were discovered.

Edge attributes:
Did not detect any nonnumeric attributes.
Numeric attributes:

	min	max	mean
number_...	1	33	1.76094
weight	1	33	1.76094

This network seems to be valued.

Average degree: 7.2146
This graph is not weakly connected.
There are 3 weakly connected components. (0 isolates)
The largest connected component consists of 194 nodes.
Did not calculate strong connectedness because this graph was not directed.

Density (disregarding weights): 0.0293
Additional Densities by Numeric Attribute

This graph claims to be undirected.

Revised Network

Nodes: 238
Isolated nodes: 0
Node attributes present: label, number_of_authored_works, times_cited

Edges: 872
No self loops were discovered.
No parallel edges were discovered.

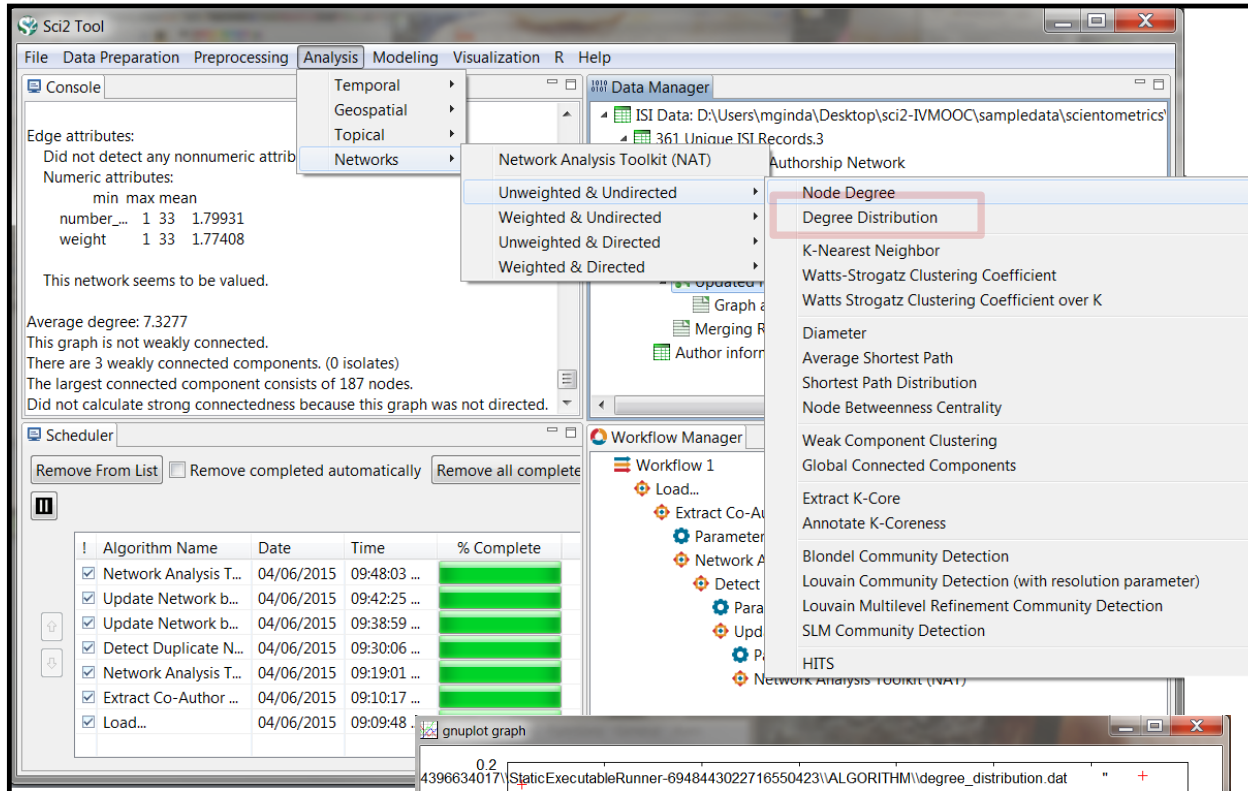
Edge attributes:
Did not detect any nonnumeric attributes.
Numeric attributes:

	min	max	mean
number_...	1	33	1.79931
weight	1	33	1.77408

This network seems to be valued.

Average degree: 7.3277
This graph is not weakly connected.
There are 3 weakly connected components. (0 isolates)
The largest connected component consists of 187 nodes.
Did not calculate strong connectedness because this graph was not directed.

Density (disregarding weights): 0.0309
Additional Densities by Numeric Attribute



Sci2 Tool

File Data Preparation Preprocessing Analysis Modeling Visualization R Help

Console

Edge attributes:

Did not detect any nonnumeric attributes

Numeric attributes:

	min	max	mean
number...	1	33	1.79931
weight	1	33	1.77408

This network seems to be valued.

Average degree: 7.3277

This graph is not weakly connected.

There are 3 weakly connected components. (0 isolates)

The largest connected component consists of 187 nodes.

Did not calculate strong connectedness because this graph was not directed.

Scheduler

Remove From List Remove completed automatically Remove all complete

Algorithm Name	Date	Time	% Complete
<input checked="" type="checkbox"/> Network Analysis T...	04/06/2015	09:48:03 ...	<div style="width: 100%;"></div>
<input checked="" type="checkbox"/> Update Network b...	04/06/2015	09:42:25 ...	<div style="width: 100%;"></div>
<input checked="" type="checkbox"/> Update Network b...	04/06/2015	09:38:59 ...	<div style="width: 100%;"></div>
<input checked="" type="checkbox"/> Detect Duplicate N...	04/06/2015	09:30:06 ...	<div style="width: 100%;"></div>
<input checked="" type="checkbox"/> Network Analysis T...	04/06/2015	09:19:01 ...	<div style="width: 100%;"></div>
<input checked="" type="checkbox"/> Extract Co-Author ...	04/06/2015	09:10:17 ...	<div style="width: 100%;"></div>
<input checked="" type="checkbox"/> Load...	04/06/2015	09:09:48 ...	<div style="width: 100%;"></div>

Workflow Manager

Workflow 1

- Load...
- Extract Co-Author...
- Parameter...
- Network Analysis...
- Detect Duplicate...
- Parameter...
- Update Network...
- Parameter...
- Network Analysis Toolkit (NAT)

gnuplot graph

4396634017\StaticExecutableRunner-6948443022716550423\ALGORITHM\degree_distribution.dat

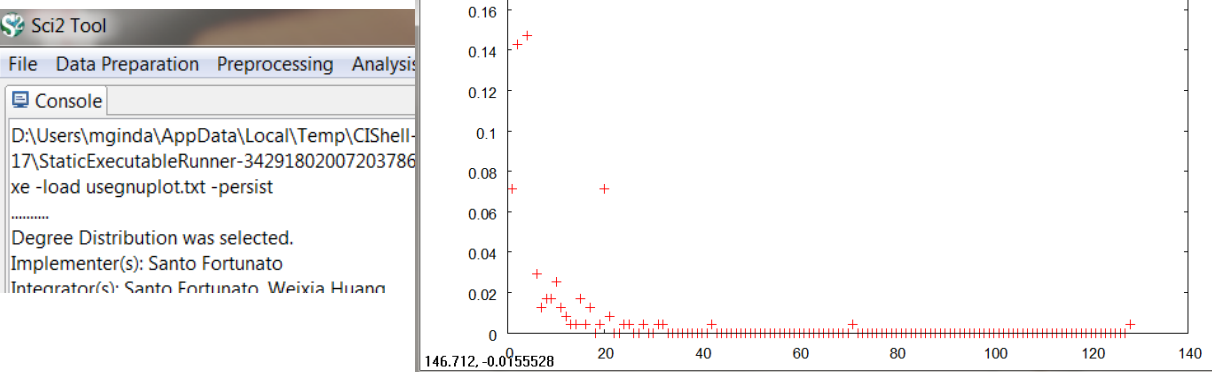
Let's start to analyze the updated network. To start, let's find the degree for each node.

Select the updated network in the data manager, and then select in the menu *Analysis -> Networks -> Unweighted & Undirected -> Node Degree*

A new network file will be output that has appended a degree to each node in your network file.

To see the distribution of node degrees, use the same menu path above, except you will need to select algorithm *Degree Distribution*. A pop-up window will appear, for now, just hit OK. Two data files will appear, we'll select the first.

To visualize this file, select *Visualization -> General -> GnuPlot*



Sci2 Tool

File Data Preparation Preprocessing Analysis

Console

```
D:\Users\mginda\AppData\Local\Temp\CIShell-17\StaticExecutableRunner-34291802007203786\exe -load usegnuplot.txt -persist
.....
Degree Distribution was selected.
Implementer(s): Santo Fortunato
Integrator(s): Santo Fortunato Weixia Huang
```

gnuplot graph

146.712, -0.0155528

0.2

0.18

0.16

0.14

0.12

0.1

0.08

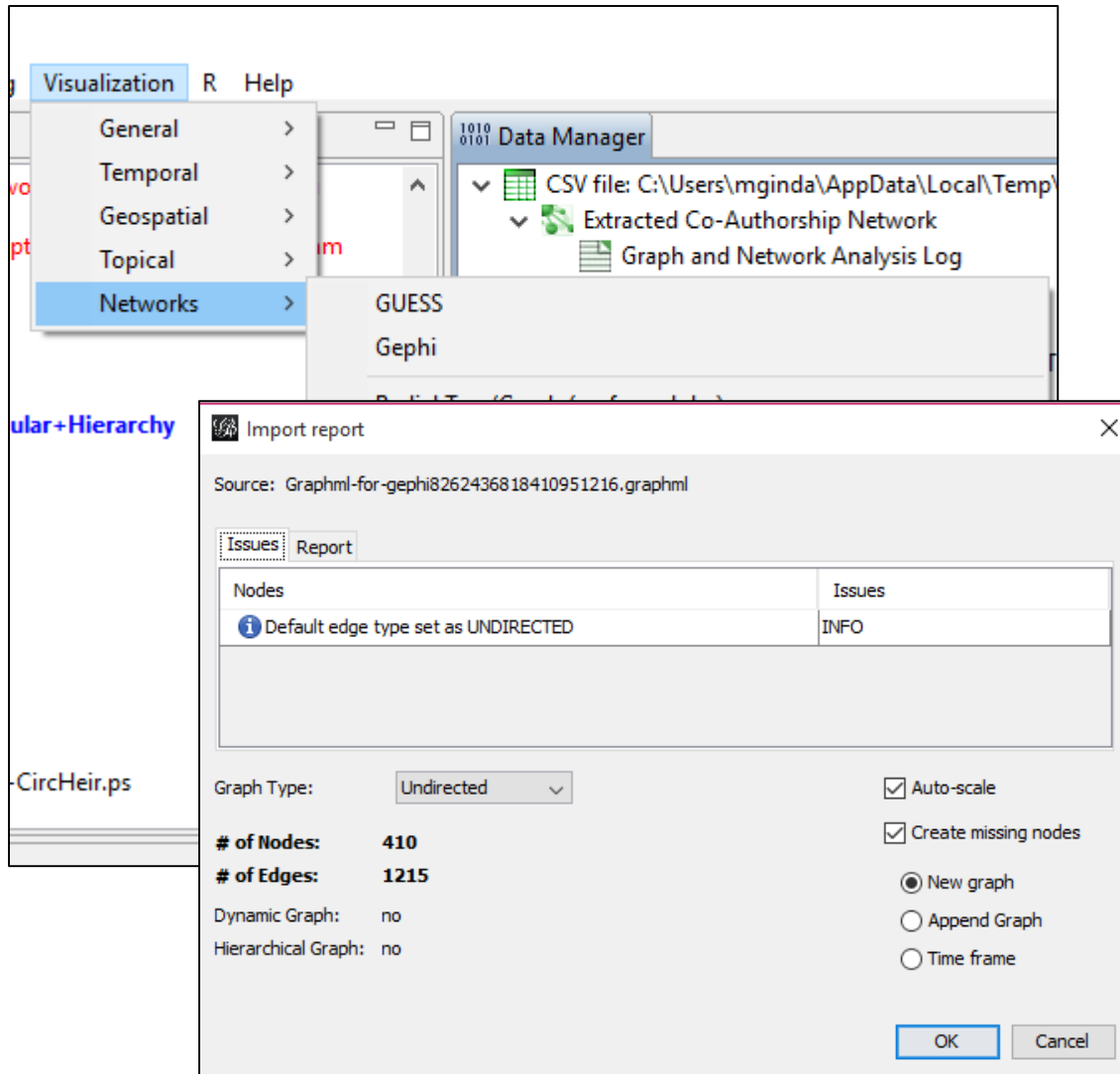
0.06

0.04

0.02

0

0 20 40 60 80 100 120 140



Next we will visualize the network in Gephi.

Navigate to **Visualization > Networks > Gephi**.

The algorithm is a bridge that passes the network data to Gephi. The program will automatically start. The tool produces an Import Report. It lets you select the network type, gives load errors, etc.

Next, is a brief walk through of Gephi's three main sections, and outline various functions and tools available.

Partition **Ranking** **Clustering**

Nodes Edges

---Choose a rank parameter

Pane lets user control size and color of nodes and edges based on network partitions, node and edge rankings, and clustering results and scale values to splines.

Layout

---Choose a layout

Run

Pane lets user select a network layout, and adjust the layout algorithm parameters

Graph

Tools let you select nodes and edges, and color nodes and edges based on paths.

Context

Nodes: 410
Edges: 1215
Undirected Graph

Basic network stats and filter stats.

Statistics **Filters**

Settings

Network Overview

- Average Degree Run
- Avg. Weighted Degree Run
- Network Diameter Run
- Graph Density Run
- HITS Run
- Modularity Run
- PageRank Run
- Connected Components Run

Node Overview

- Avg. Clustering Coefficient Run
- Eigenvector Centrality Run

Edge Overview

- Avg. Path Length Run

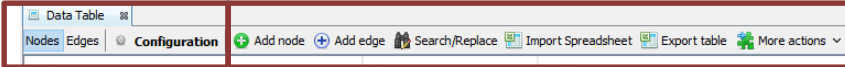
These tools let you re-center the network, and rest color, size and label attributes.

Presets... Reset

Workspace 0

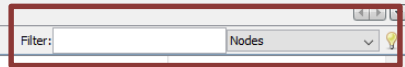
Tools let you adjust network labels attributes, and take snapshots of the graph viewer.

Navigate workspaces



Select the node or edge lists data, and configure the sheet

Add node edges, imports data (nodes and edges lists) to create new networks in blank workspaces, and exports data tables for a network.



A REGEX filter for nodes and edges table columns in data table.

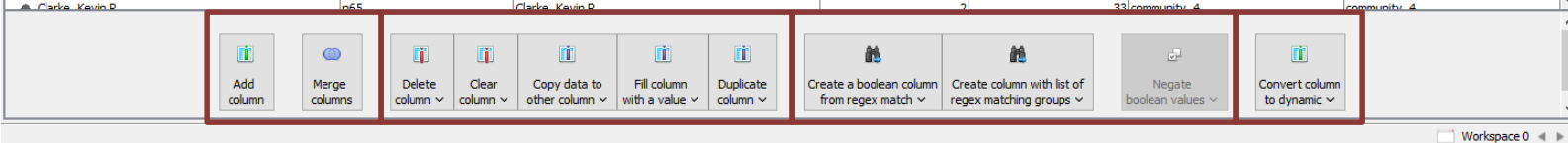
Nodes	ID	Label	number_of_authored...	times_cited	blonde_community_level_0	blonde_community_level_1
● Chaves, Sandra S.	n10	Chaves, Sandra S.	5	1389	community_0	community_0
● Gubareva, Larisa	n11	Gubareva, Larisa	5	48	community_0	community_0
● Hall, Henrietta	n12	Hall, Henrietta	2	48	community_0	community_0
● Wallis, Teresa	n13	Wallis, Teresa	4	17	community_0	community_0
● Villanueva, Julie	n14	Villanueva, Julie	4	48	community_0	community_0
● Xu, Xiyan	n15	Xu, Xiyan	4	48	community_0	community_0
● Bresee, Joseph	n16	Bresee, Joseph	10	48	community_0	community_0
● Cox, Nancy	n17	Cox, Nancy	10	1712	community_0	community_0
● Tappero, Jordan W.	n28	Tappero, Jordan W.	3	1995	community_3	community_3
● Nyenswah, Tolbert G.	n30	Nyenswah, Tolbert G.	6	58	community_1	community_1
● Montgomery, Joel M.	n31	Montgomery, Joel M.	5	307	community_1	community_1
● Neatherlin, John	n32	Neatherlin, John	3	42	community_1	community_1
● Singleton, James A.	n35	Singleton, James A.	10	25	community_54	community_45
● Flannery, Brendan	n36	Flannery, Brendan	2	933	community_2	community_2
● Fry, Alicia	n37	Fry, Alicia	3	51	community_2	community_2
● Pesik, Nicki	n41	Pesik, Nicki	2	310	community_3	community_2
● Brown, Clive M.	n42	Brown, Clive M.	3	52	community_3	community_3
● Aranas, Aaron E	n43	Aranas, Aaron E	3	61	community_3	community_3
● Cohen, Nicole J.	n48	Cohen, Nicole J.	3	54	community_16	community_3
● Hale, Christa	n51	Hale, Christa				
● Holton, Kelly		Holton, Kelly				
● Clarke, Kevin P.	n55	Clarke, Kevin P.	2	33	community_4	community_4

Adds & Merge data column tools

Column editing tools

Create columns fitting Boolean criteria and regex functions. Useful for filtering.

Converts data fields from standard to dynamic (temporal data fields)



The screenshot displays the Gephi 0.8.2 Preview window. On the left is the 'Preview Settings' panel, and on the right is the network graph visualization. The settings panel is organized into several sections:

- Presets:** A dropdown menu set to 'Default'.
- Nodes:**
 - Border Width: 1.0
 - Border Color: custom [0,0,0]
 - opacity: 100.0
- Node Labels:**
 - Show Labels:
 - Font: Arial 12 Plain
 - Proportional size:
 - Color: custom [0,0,0]
 - Shorten label:
 - Max characters: 30
 - Outline size: 0.0
 - Outline color: custom [255,255,255]
 - Outline opacity: 80.0
 - Box:
 - Box color: parent
 - Box opacity: 100.0
- Edges:**
 - Show Edges:
 - Thickness: 1.0
 - Rescale weight:
 - Color: mixed
 - Opacity: 100.0
 - Curved:
 - Radius: 0.0
- Edge Arrows:**
 - Size: 3.0
- Edge Labels:**
 - Show Labels:
 - Font: Arial 10 Plain
 - Color: original
 - Shorten label:
 - Max characters: 30
 - Outline size: 0.0
 - Outline color: custom [255,255,255]
 - Outline opacity: 80.0

The network graph on the right shows a dense, complex network of nodes and edges. The nodes are represented by small circles, and the edges are thin lines connecting them. The graph is rendered in a grayscale style.

At the bottom of the Preview window, there is a 'Preview ratio: 100%' indicator, a 'Refresh' button, and an 'Export: SVG/PDF/PNG' button. A 'Background Reset zoom - +' button is also visible at the bottom right.

Preset layouts, and a layout configuration saving feature

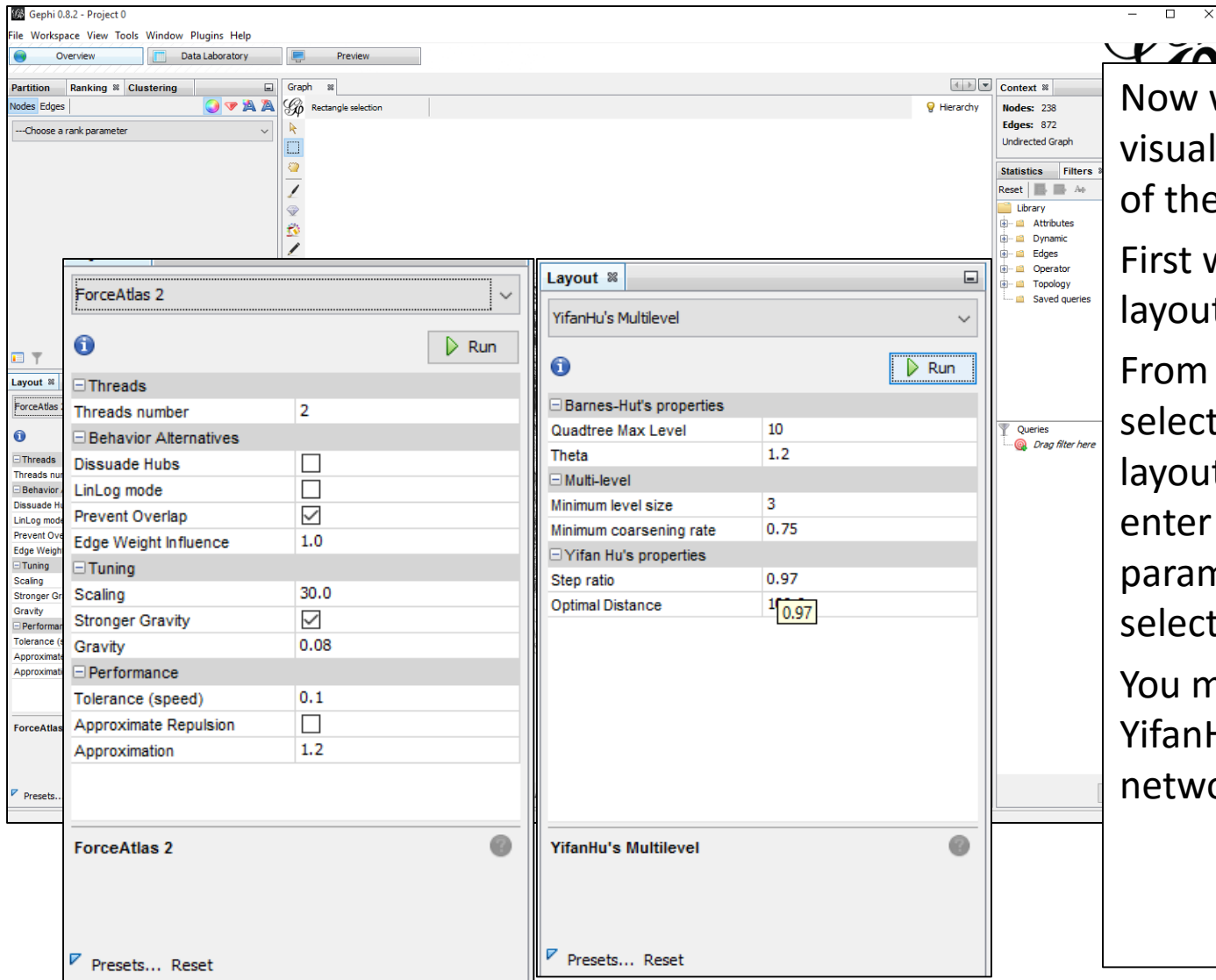
Node border and opacity attributes

Node labels attribute selection, including label size, color, length, etc.

Edge size, color, type, and opacity and scaling attributes.

Edge label attributes.

Refresh and exporting the Preview tools, and zoom resets.



The screenshot shows the Gephi 0.8.2 interface with two configuration panes open. The 'ForceAtlas 2' pane is on the left, and the 'YifanHu's Multilevel' pane is on the right. Both panes have a 'Run' button. The 'ForceAtlas 2' pane shows various settings for threads, behavior alternatives, tuning, and performance. The 'YifanHu's Multilevel' pane shows settings for Barnes-Hut's properties, Multi-level, and Yifan Hu's properties. The 'Optimal Distance' field in the Yifan Hu's properties section is highlighted with a yellow box and contains the value 0.97.

Category	Parameter	Value
Threads	Threads number	2
	Behavior Alternatives	
Behavior Alternatives	Dissuade Hubs	<input type="checkbox"/>
	LinLog mode	<input type="checkbox"/>
	Prevent Overlap	<input checked="" type="checkbox"/>
	Edge Weight Influence	1.0
Tuning	Scaling	30.0
	Stronger Gravity	<input checked="" type="checkbox"/>
	Gravity	0.08
Performance	Tolerance (speed)	0.1
	Approximate Repulsion	<input type="checkbox"/>
	Approximation	1.2

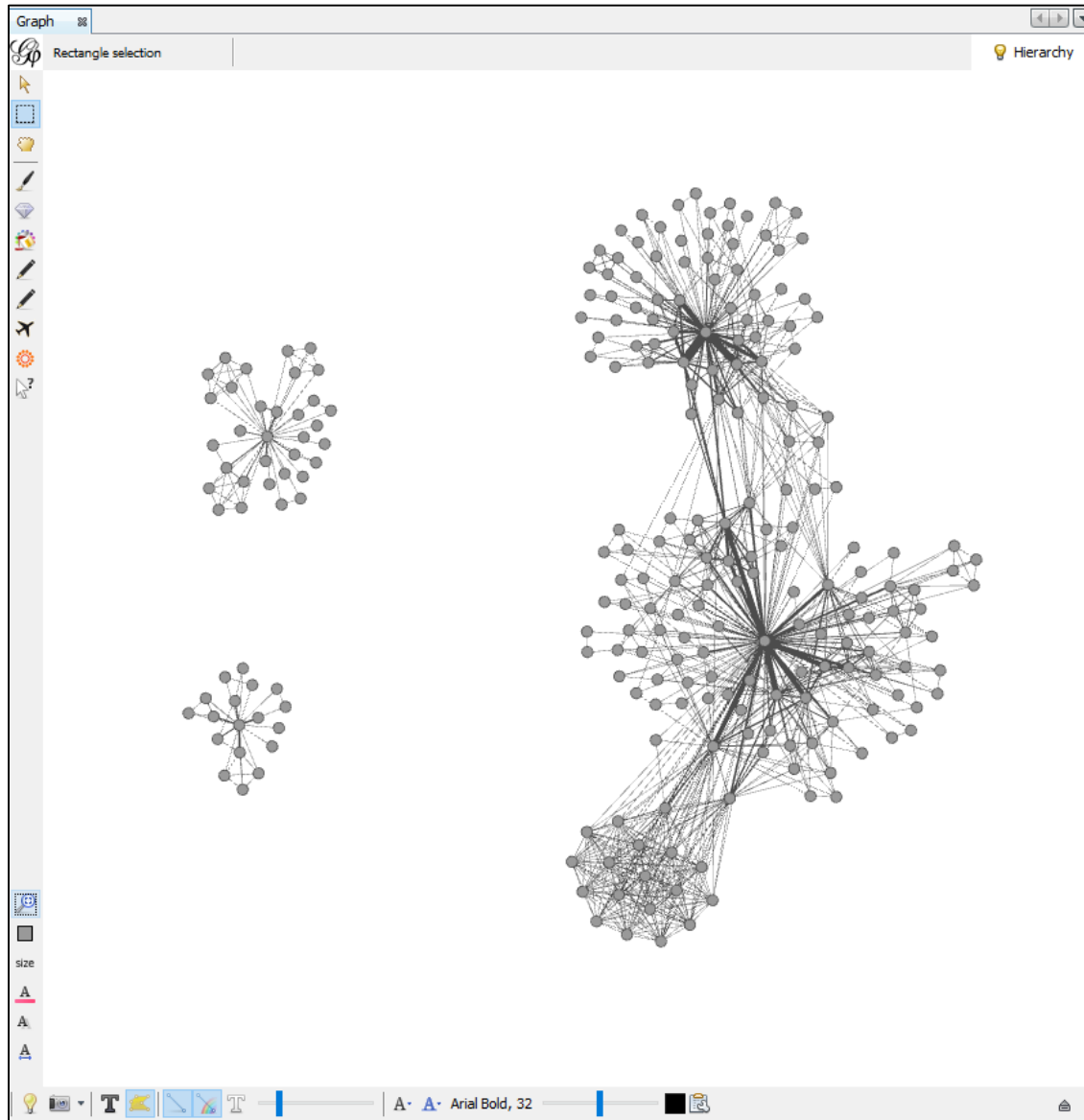
Category	Parameter	Value
Barnes-Hut's properties	Quadtree Max Level	10
	Theta	1.2
Multi-level	Minimum level size	3
	Minimum coarsening rate	0.75
Yifan Hu's properties	Step ratio	0.97
	Optimal Distance	0.97

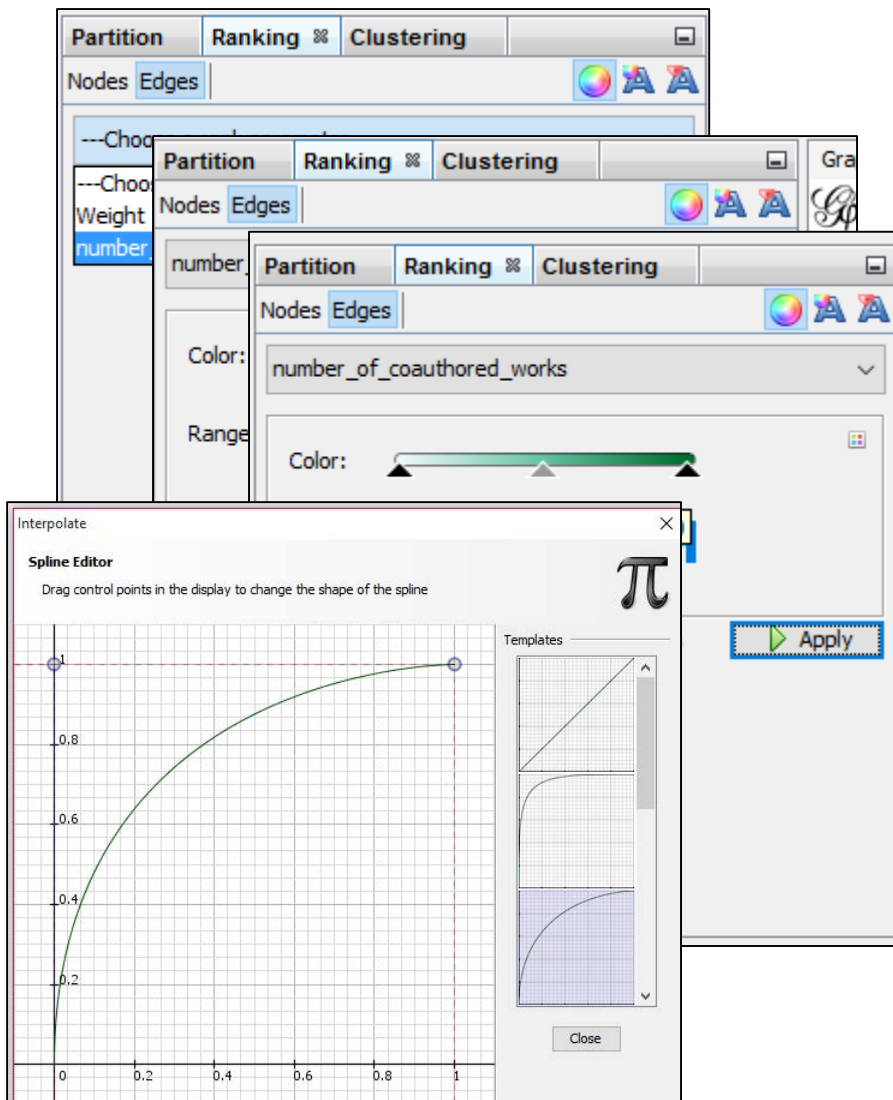
Now we can start an visualization and analysis of the network.

First we will adjust the layout of the network.

From the layout pane, select the “ForceAtlas2” layout algorithm and enter the following parameters, and then select “Run”.

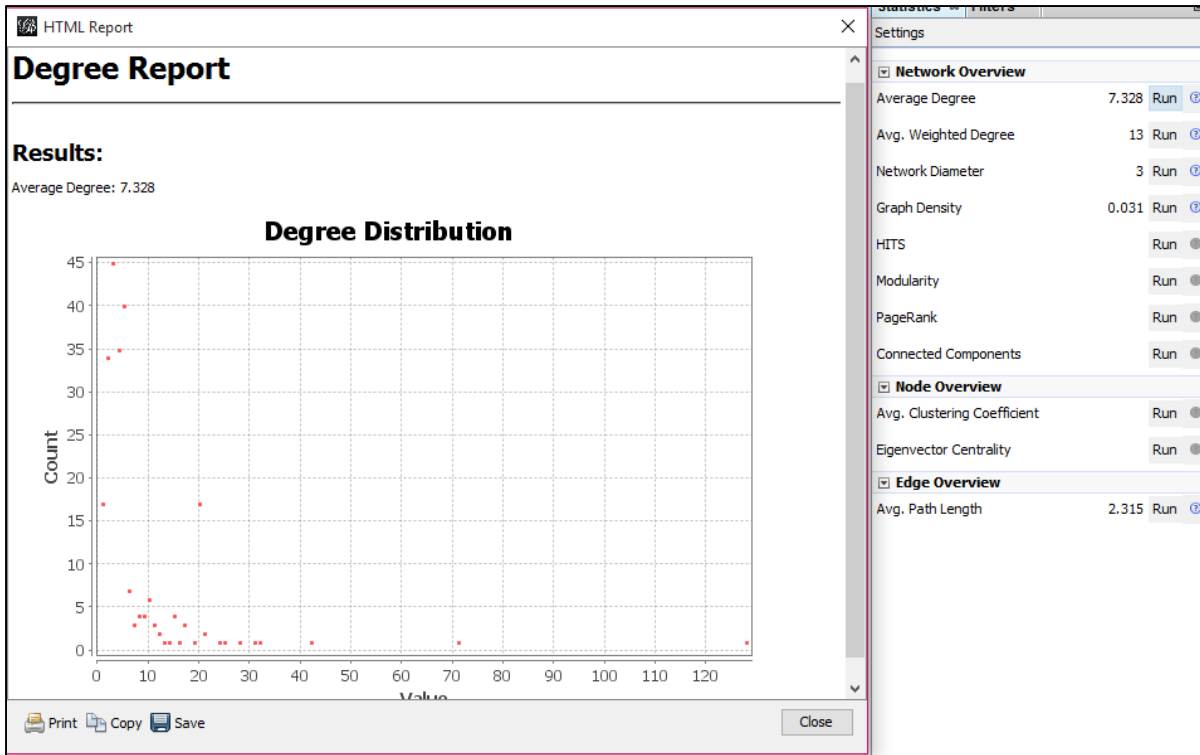
You may also select YifanHu’s Multilevel force network layout.





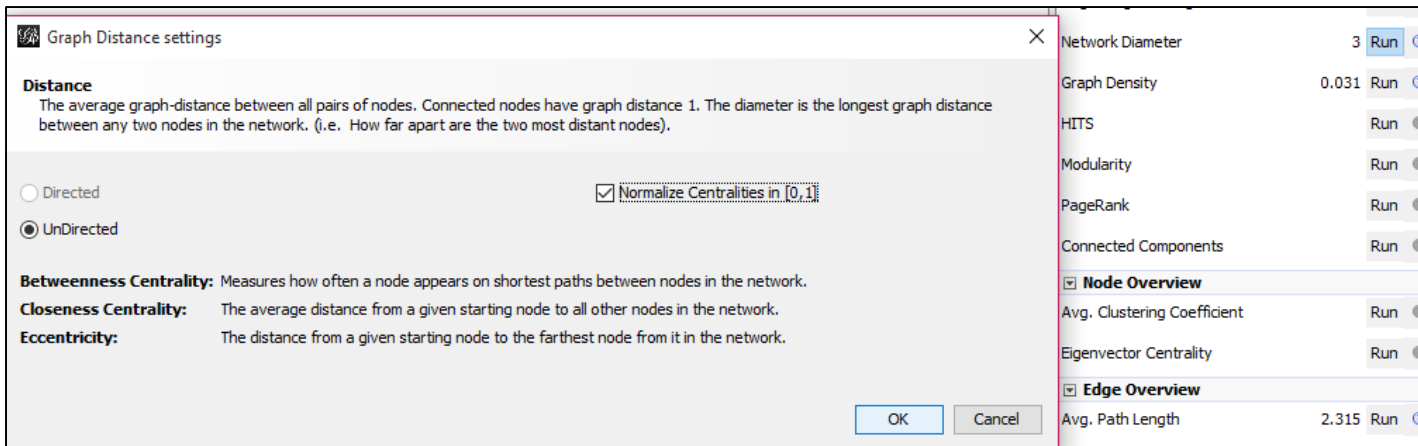
Next, we can adjust the edge color by selecting the Edge tab in the Ranking window.

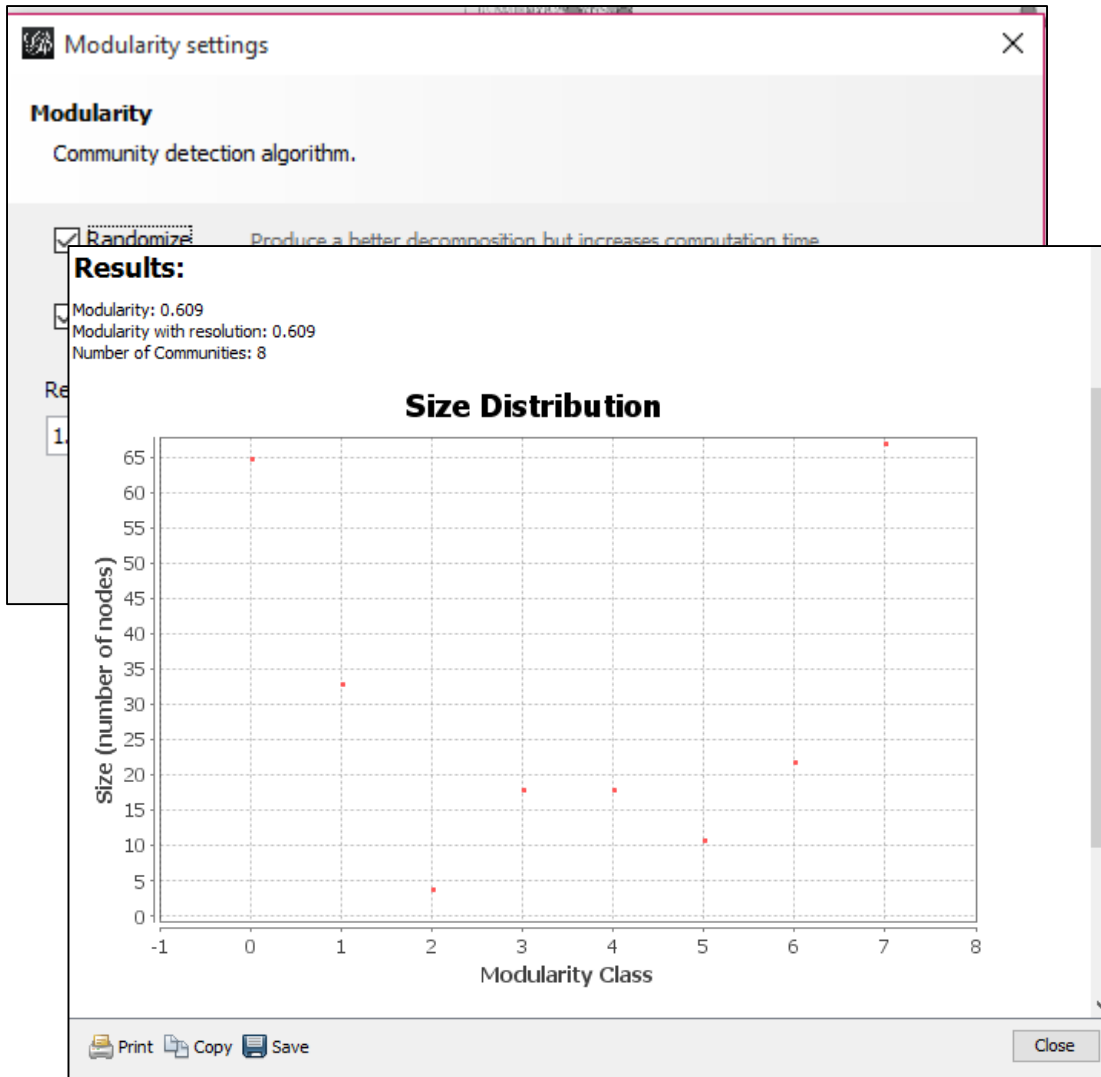
- In the drop down menu, select “number_of_coauthored_works”.
- Select the small square in the right corner of the Color Range box. This lets us choose new color ranges for variables.
- You may also set the color range and values to apply the colors to, or adjust color scaling variables by adjusting the spline.



Gephi provides a variety of node and edge statistics to help understand the relationships, clustering, paths, centrality, and communities within a network.

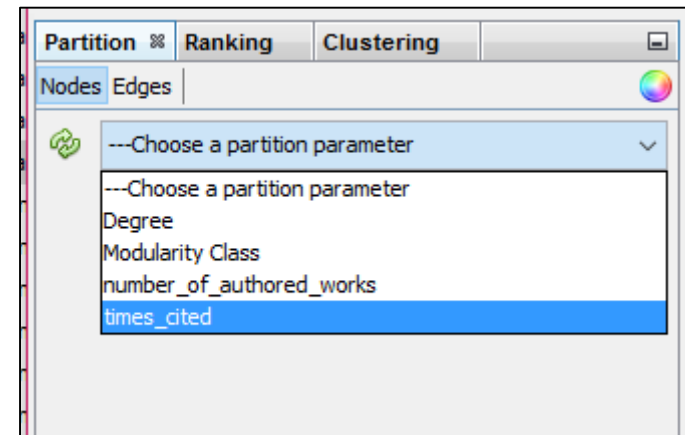
Try implementing the Average Degree, and Network Diameter statistics, which we will next visualize.





The modularity statistical algorithm calculates how the connectedness of a network, and the Blondel Communities that exist in the network. The communities are added as a partition to the nodes.

The modularity categories may be applied to the network from the Partitions window.



Partition Ranking Clustering

Nodes Edges

Degree

---Choose a rank parameter

Betweenness Centrality

Closeness Centrality

Degree

28

Eccentricity

Modularity Class

Weighted Degree

Default

Invert

Recent

Apply

Graph

Rectangle selection

Hierarchy

size

Arial Bold, 32

Close

Partition Ranking Clustering

Nodes Edges

Degree

Color: [Color Scale]

Range: [Range Slider]

1 128

Spline...

Apply

Layout

ForceAtlas 2

Run

Threads

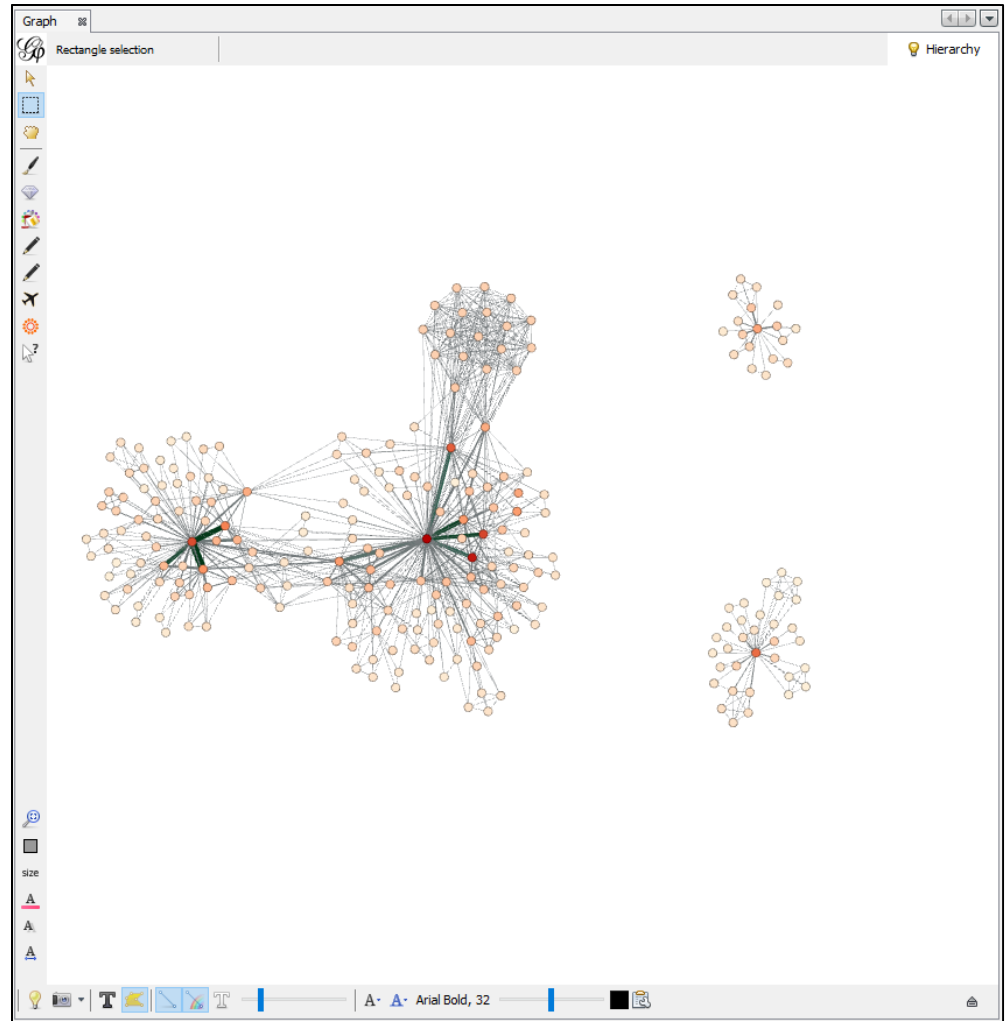
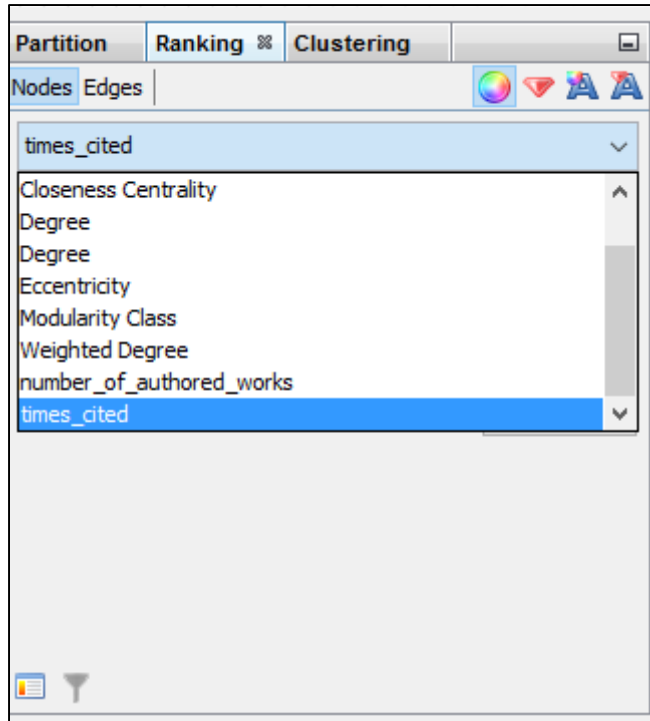
Threads number 2

Behavior Alternatives

Interpolate

Spline Editor

Drag control points in the display to change the shape of the



Gephi 0.8.2 - Project 0

File Workspace View Tools Window Plugins Help

Overview Data Laboratory Preview

Data Table

Nodes Edges Configuration Add node Add edge Search/Replace Import Spreadsheet Export table More actions Filter: Nodes

Nodes	Id	Label	number_of_authored...	times_cited	blondel_community_level_0	blondel_community_level_1
Finelli, Lyn	n1	Finelli, Lyn	9	1389	community_0	community_0
Blanton, Lenee	n3	Blanton, Lenee	4	48	community_0	community_0
Brammer, Lynnette	n4	Brammer, Lynnette	4	48	community_0	community_0
Smith, Sophie	n5	Smith, Sophie	2	17	community_0	community_0
Mustaquim, Desiree	n6	Mustaquim, Desiree	4	48	community_0	community_0
Steffens, Craig	n7	Steffens, Craig	4	48	community_0	community_0
Leon, Michelle	n9	Leon, Michelle	3	41	community_0	community_0
Chaves, Sandra S.	n10	Chaves, Sandra S.	5	365	community_0	community_0
Gubareva, Larisa	n11	Gubareva, Larisa	5	307	community_0	community_0
Hall, Henrietta	n12	Hall, Henrietta				
Wallis, Teresa	n13	Wallis, Teresa				

Add column - Settings

Add column to nodes

Title:

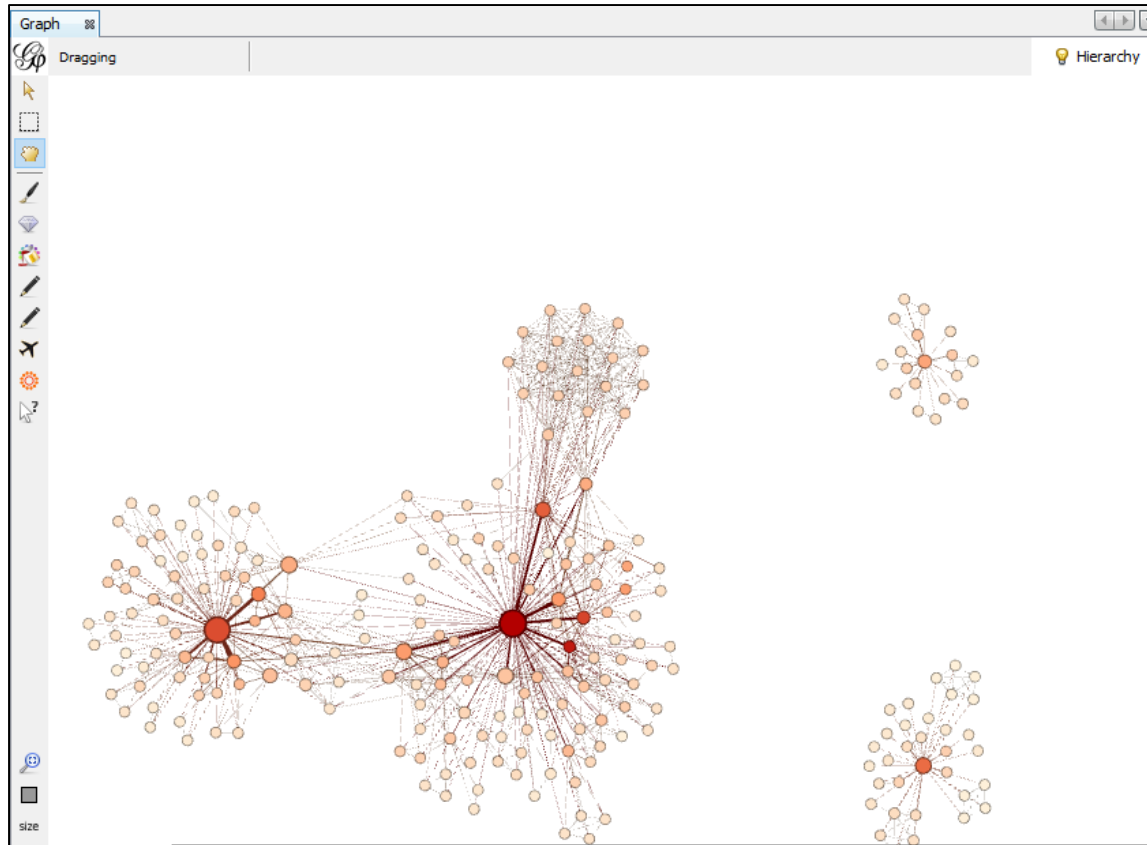
Type:

OK Cancel

Betweenness C...	Modularity Cl...	Label2
1	0.004	3 Wasserman, Ss
7	0.007	0 Vicsek, T
8	0.267	7 Vespignani, A
4	0.02	7 Vazquez, A
3	0.02	6 Stanley, He
1	0.006	7 Pastor-satorras, R
9	0.013	0 Oltvai, Zn
2	0.009	7 Munoz, Ma
5	0.018	0 Kahng, B
1	0.017	1 Garfield, E
2	0.007	7 Barrat, A
2	0.434	0 Barabasi, Al
8	0	1

Add column Merge columns Delete column Clear column Copy data to other column Fill column with a value Duplicate column
 Create a boolean column from regex match Create column with list of regex matching groups Negate boolean values Convert column to dynamic

Workspace 0



Label text settings

Nodes Edges Show properties

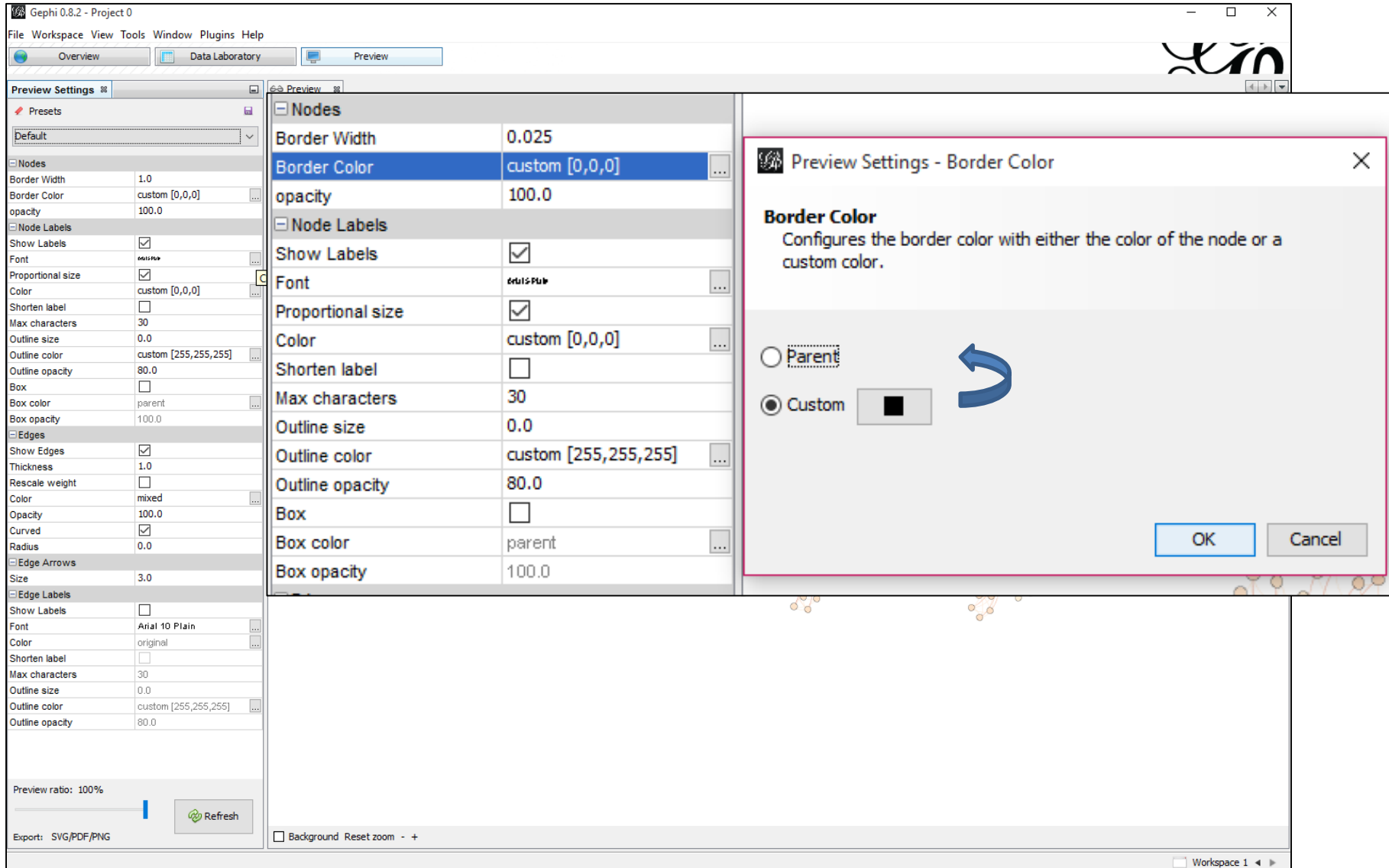
Select attributes to display as labels

- Id
- Label
- number_of_authored_works
- times_cited
- Degree
- Weighted Degree
- Eccentricity
- Closeness Centrality
- Betweenness Centrality
- Modularity Class
- Label2

Global Nodes Edges **Labels**


<p><input checked="" type="checkbox"/> Node</p> <p>Font: <input type="text" value="Arial Bold, 9"/> Color: <input type="color" value="black"/></p> <p>Size: <input type="range" value="9"/></p>	<p><input type="checkbox"/> Edge</p> <p>Font: <input type="text" value="Arial Bold, 32"/> Color: <input type="color" value="gray"/></p> <p>Size: <input type="range" value="32"/></p>	<p>Size: <input type="text" value="Scaled"/></p> <p>Color: <input type="text" value="Unique"/></p> <p>Hide non-selected <input type="checkbox"/></p> <p><input type="button" value="Configure..."/></p>
---	---	---

The screenshot displays the Gephi software interface with a network graph. The graph features several clusters of nodes, with a central hub node labeled 'Barbara, A.' and other prominent nodes like 'Vazquez, A.', 'Pastor-Latorras, R.', 'Vespignani, Daniel, A.', 'Murilo, Ma.', 'Stanley, He.', 'Kahng, B.', 'Victor, T.', and 'Dillon, Zn.'. The interface includes a toolbar on the left with various tools like 'Rectangle selection', 'Hierarchy', and 'Size'. At the bottom, there are configuration panels for 'Global', 'Nodes', 'Edges', and 'Labels'. The 'Labels' panel is currently active, showing settings for 'Node' and 'Edge' labels, including font (Arial Bold, 16), color (black), and size (scaled).



The screenshot shows the Gephi 0.8.2 interface with the Preview Settings dialog open. The dialog is titled "Preview Settings - Border Color" and contains the following information:

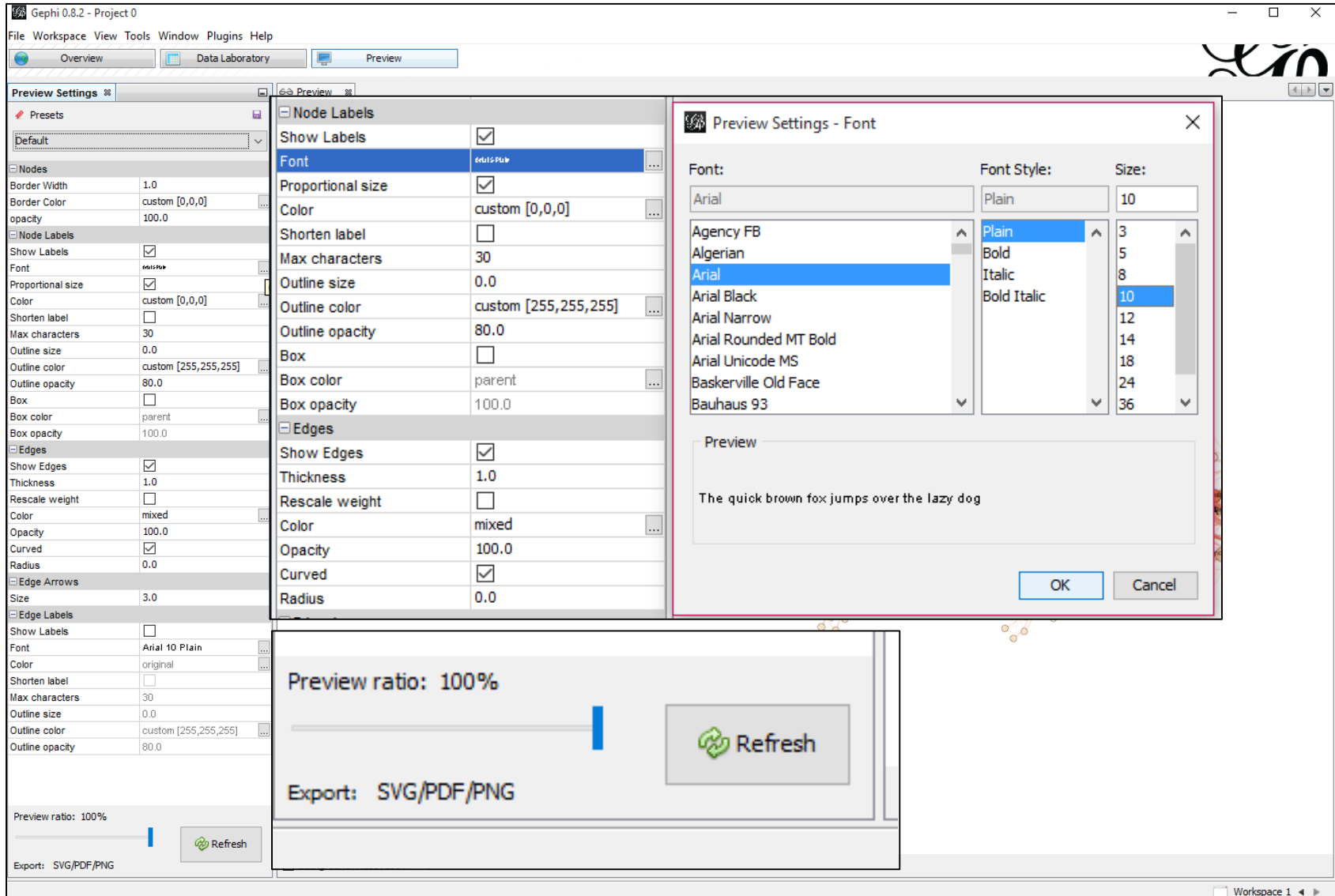
Border Color
Configures the border color with either the color of the node or a custom color.

Parent
 Custom 

OK Cancel

The background interface shows the "Preview Settings" panel with the following parameters:

Category	Parameter	Value
Nodes	Border Width	0.025
	Border Color	custom [0,0,0]
	opacity	100.0
	Node Labels	
Node Labels	Show Labels	<input checked="" type="checkbox"/>
	Font	Arial 10 Plain
	Proportional size	<input checked="" type="checkbox"/>
	Color	custom [0,0,0]
	Shorten label	<input type="checkbox"/>
	Max characters	30
	Outline size	0.0
	Outline color	custom [255,255,255]
	Outline opacity	80.0
	Box	<input type="checkbox"/>
Edges	Show Edges	<input checked="" type="checkbox"/>
	Thickness	1.0
	Rescale weight	<input type="checkbox"/>
	Color	mixed
Edge Arrows	Opacity	100.0
	Curved	<input checked="" type="checkbox"/>
	Radius	0.0
Edge Labels	Size	3.0
	Show Labels	<input type="checkbox"/>
	Font	Arial 10 Plain
	Color	original
	Shorten label	<input type="checkbox"/>



The screenshot shows the Gephi 0.8.2 interface with the 'Preview Settings' panel open. The 'Node Labels' section is expanded, showing various parameters for node labels. A 'Preview Settings - Font' dialog box is overlaid on top, allowing for font selection and preview. The dialog shows 'Arial' as the selected font, 'Plain' as the font style, and '10' as the font size. The preview text reads: 'The quick brown fox jumps over the lazy dog'. Below the dialog, there is a 'Preview ratio: 100%' indicator and an 'Export: SVG/PDF/PNG' button.

Preview Settings - Font

Font:	Font Style:	Size:
Arial	Plain	10
Agency FB	Plain	3
Algerian	Bold	5
Arial	Italic	8
Arial Black	Bold Italic	10
Arial Narrow		12
Arial Rounded MT Bold		14
Arial Unicode MS		18
Baskerville Old Face		24
Bauhaus 93		36

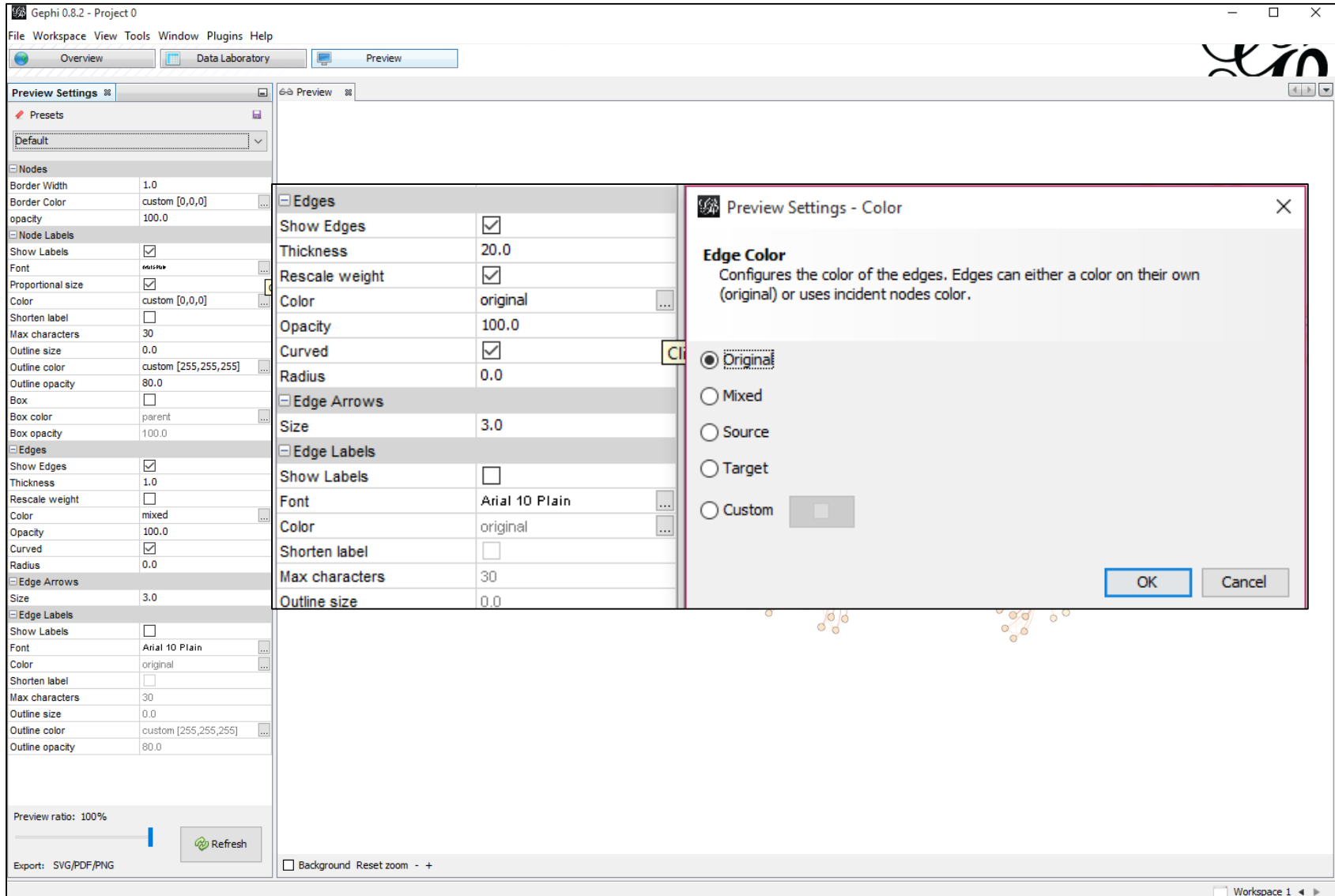
Preview
The quick brown fox jumps over the lazy dog

OK Cancel

Preview ratio: 100%

Export: SVG/PDF/PNG

Refresh



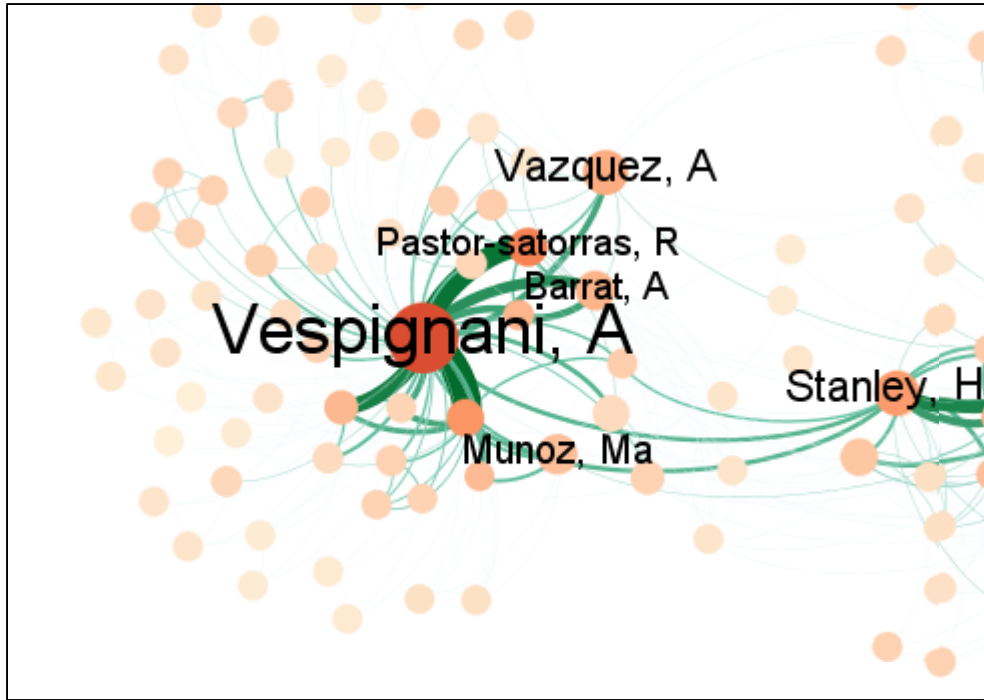
The screenshot shows the Gephi 0.8.2 interface with the Preview Settings dialog open. The dialog is titled "Preview Settings - Color" and is used to configure the color of the edges in the visualization. The "Edge Color" section is active, showing the following options:

- Original
- Mixed
- Source
- Target
- Custom

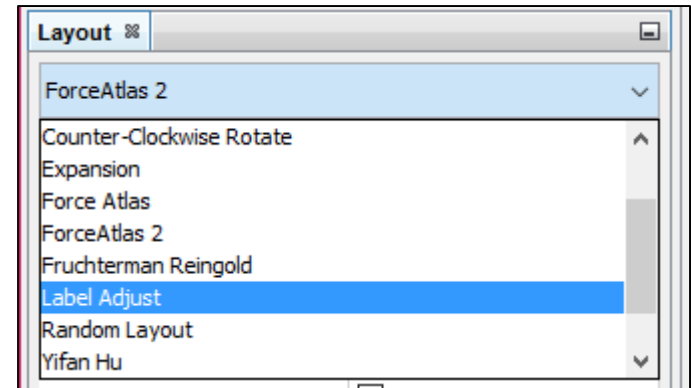
The "Original" option is selected, indicating that the edges will use their own color. The dialog also includes "OK" and "Cancel" buttons at the bottom right.

The background shows the main Gephi interface with the "Preview" tab selected. The "Preview Settings" panel on the left is expanded to show the "Edges" section, which includes the following parameters:

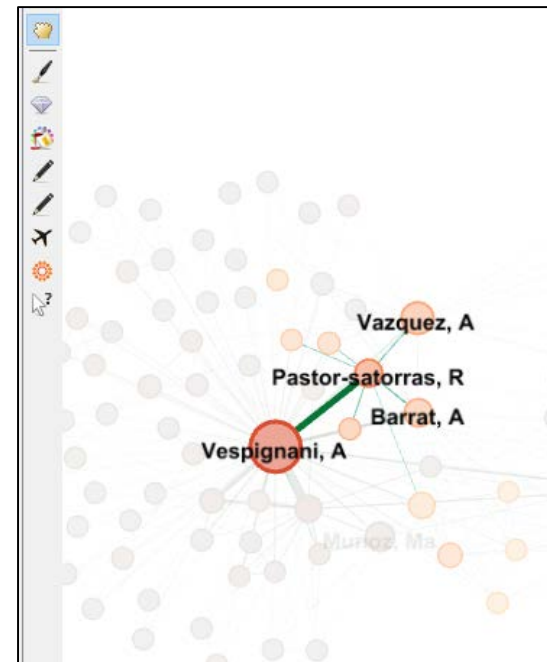
- Show Edges:
- Thickness: 20.0
- Rescale weight:
- Color: original
- Opacity: 100.0
- Curved:
- Radius: 0.0
- Edge Arrows: Size: 3.0
- Edge Labels: Show Labels: Font: Arial 10 Plain Color: original Shorten label: Max characters: 30 Outline size: 0.0



Layout algorithm



Manual adjustments

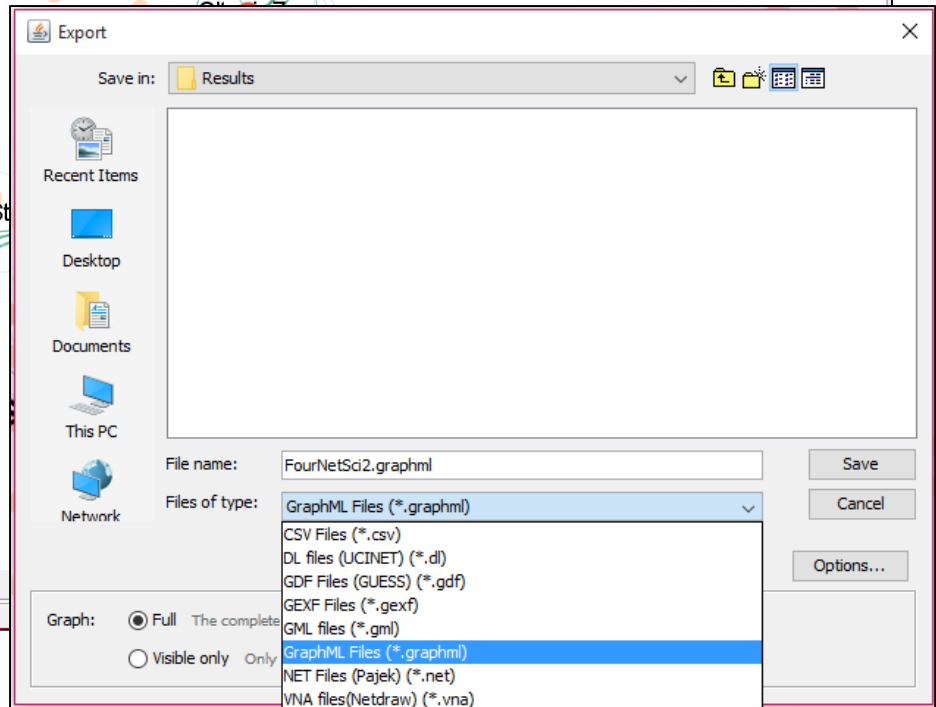
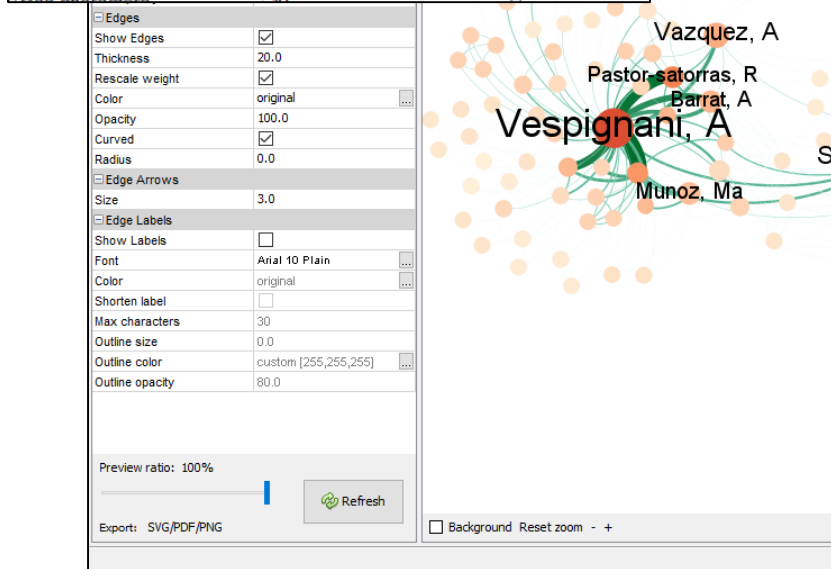
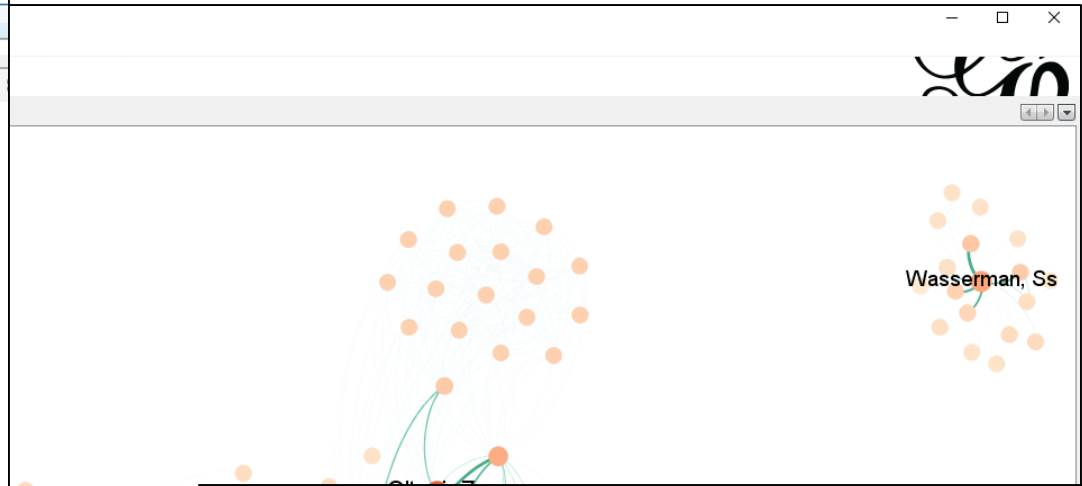
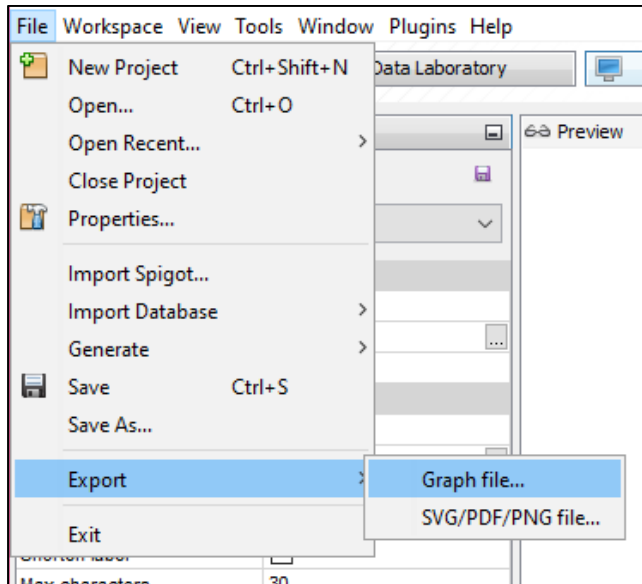


The screenshot displays the Gephi 0.8.2 software interface. The main window shows a network graph with orange nodes and green edges, labeled with 'Oltvai, Zn' and 'Wasserman, Ss'. The interface includes a menu bar (File, Workspace, View, Tools, Window, Plugins, Help) and a toolbar with 'Overview', 'Data Laboratory', and 'Preview' tabs. The 'Preview' tab is active, showing the network visualization.

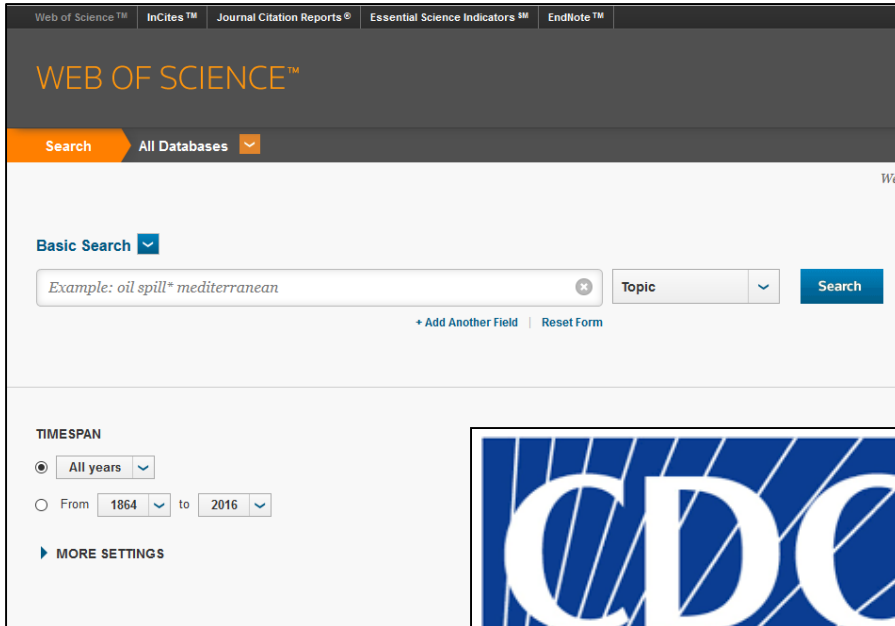
On the left, the 'Preview Settings' panel is open, showing configuration options for Nodes, Node Labels, Edges, and Edge Arrows. The 'Node Labels' section is expanded, showing settings for 'Show Labels', 'Font', 'Proportional size', 'Color', 'Shorten label', 'Max characters', 'Outline size', 'Outline color', and 'Outline opacity'. The 'Edges' section is also expanded, showing settings for 'Show Edges', 'Thickness', 'Rescale weight', 'Color', 'Opacity', 'Curved', 'Radius', 'Edge Arrows', and 'Edge Labels'.

At the bottom of the settings panel, there is a 'Preview ratio: 100%' slider and a 'Refresh' button. The 'Export' dropdown menu is set to 'SVG/PDF/PNG'.

An 'Export' dialog box is open in the foreground, showing the 'Save in:' location as 'Results'. The dialog lists two recent items: 'CDC-Wos-CoAuth-Blondel-CircHeir.pdf' and 'CDC-Wos-CoAuth-Force2-PubsCitations.pdf'. The 'File name:' field contains '4NetSci2Researchers' and the 'Files of type:' dropdown is set to 'PDF Files (*.pdf)'. The 'Save' button is highlighted.



Questions?



Web of Science™ InCites™ Journal Citation Reports® Essential Science Indicators™ EndNote™

WEB OF SCIENCE™

Search All Databases

Basic Search

Example: oil spill* mediterranean

Topic

Search

+ Add Another Field | Reset Form

TIMESPAN

All years

From 1864 to 2016

► MORE SETTINGS



The data for this workflow was collected from **Web of Science**, which is a citation index database produced by Thomson Reuters. Data was collected using an advanced query that looked for all articles that published between 2004 and 2014 that had an author affiliated with CDC.

Only articles with at least 5 citations were downloaded in the ISI format. They were processed in Sci2 to a CSV format.

The workflow demonstrate name/entity disambiguation in OpenRefine, Sci2 network extraction and analysis algorithms, and Gephi for final visualization of the network.

Google refine A power tool for working with messy data.

« Start Over Configure Parsing Options Project name: CDC AIDS Diagnosis CityLocation 8196 Create Project »

Notes	Location	Location Code	Year Diagnosed	Year Diagnosed Code	Race or Ethnicity	Race or Ethnicity Code	Cases
1.	Akron, OH	80	1982	1982	Black (and also not Hispanic)	2054-5	1
2.	Akron, OH	80	1984	1984	White (and also not Hispanic)	2106-3	3
3.	Akron, OH	80	1985	1985	White (and also not Hispanic)	2106-3	4
4.	Akron, OH	80	1986	1986	Black (and also not Hispanic)	2054-5	3
5.	Akron, OH	80	1986	1986	White (and also not Hispanic)	2106-3	9
6.	Akron, OH	80	1987	1987	Black (and also not Hispanic)	2054-5	6
7.	Akron, OH	80	1987	1987	White (and also not Hispanic)	2106-3	14
8.	Akron, OH	80	1988	1988	Black (and also not Hispanic)	2054-5	11
9.	Akron, OH	80	1988	1988	Hispanic	2135-2	1
10.	Akron, OH	80	1988	1988	White (and also not Hispanic)	2106-3	15
11.	Akron, OH	80	1989	1989	Black (and also not Hispanic)	2054-5	7
12.	Akron, OH	80	1989	1989	White (and also not Hispanic)	2106-3	19
13.	Akron, OH	80	1990	1990	Black (and also not Hispanic)	2054-5	6
14.	Akron, OH	80	1990	1990	White (and also not Hispanic)	2106-3	23
15.	Akron, OH	80	1991	1991	Black (and also not Hispanic)	2054-5	11
16.	Akron, OH	80	1991	1991	White (and also not Hispanic)	2106-3	38
17.	Akron, OH	80	1992	1992	Black (and also not Hispanic)	2054-5	17
18.	Akron, OH	80	1992	1992	White (and also not Hispanic)	2106-3	41
19.	Akron, OH	80	1993	1993	Black (and also not Hispanic)	2054-5	28
20.	Akron, OH	80	1993	1993	White (and also not Hispanic)	2106-3	37
21.	Akron, OH	80	1994	1994	Black (and also not Hispanic)	2054-5	15
22.	Akron, OH	80	1994	1994	Hispanic	2135-2	1
23.	Akron, OH	80	1994	1994	White (and also not Hispanic)	2106-3	31
24.	Akron, OH	80	1995	1995	Black (and also not Hispanic)	2054-5	27
25.	Akron, OH	80	1995	1995	Hispanic	2135-2	1
26.	Akron, OH	80	1995	1995	White (and also not Hispanic)	2106-3	32
27.	Akron, OH	80	1996	1996	Black (and also not Hispanic)	2054-5	17

Parse data as

Character encoding Update Preview

CSV / TSV / separator-based files

Line-based text files

Fixed-width field text files

PC-Axis text files

JSON files

RDF/JSON files

XML files

Open Document Format spreadsheets (ods)

RDF/XML files

Columns are separated by

commas (CSV)

tabs (TSV)

custom 't'

Escape special characters with \

Ignore first 0 line(s) at beginning of file

Parse next 1 line(s) as column headers

Discard initial 0 row(s) of data

Load at most 0 row(s) of data

Parse cell text into numbers, dates, ...

Quotation marks are used to enclose cells containing column separators

Store blank rows

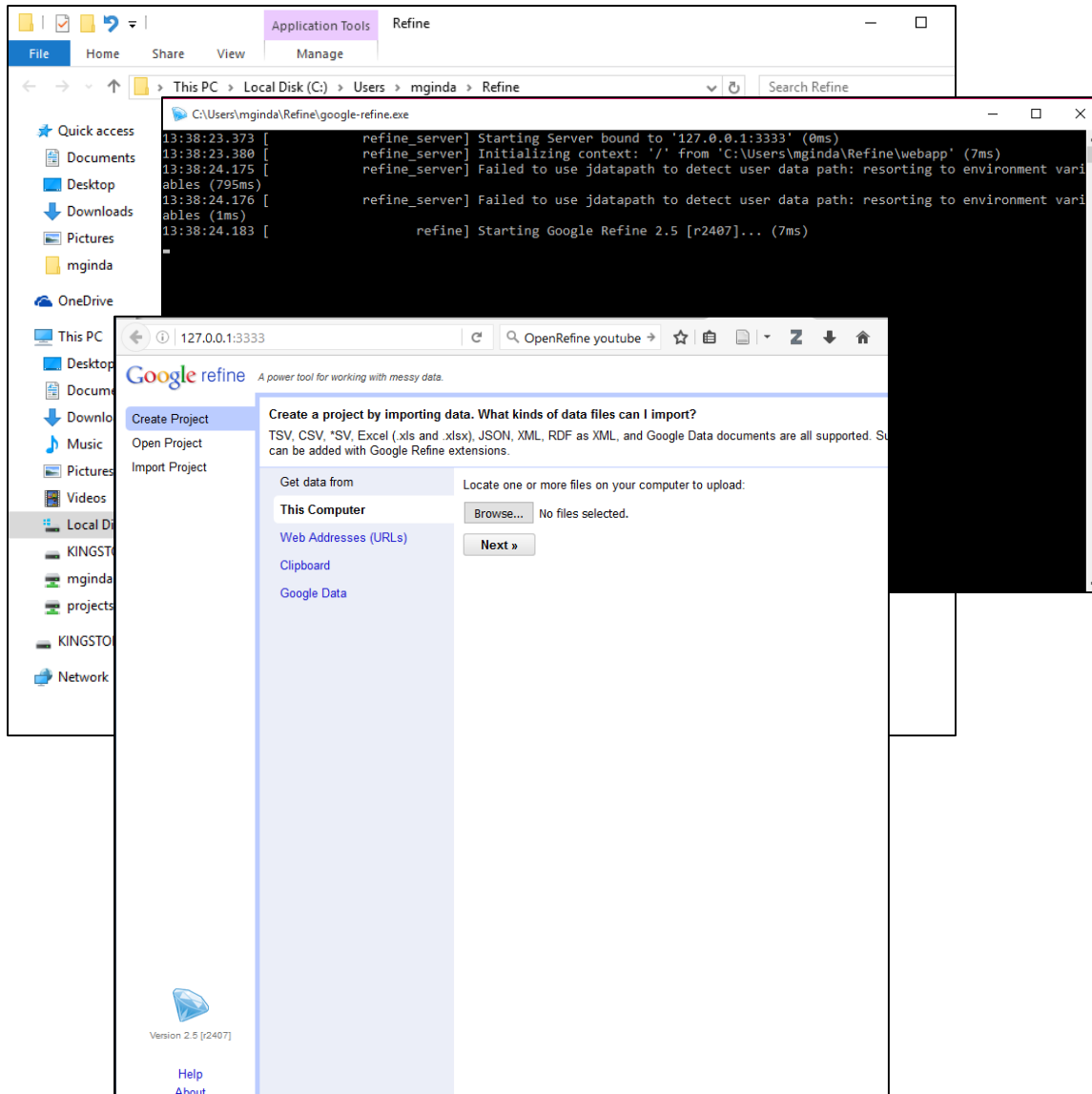
Store blank cells as nulls

Store file source (file names, URLs) in each row

For this portion of the workshop, we will use OpenRefine to review the data set, filter and facet data, and geocode the locations.

The process of geolocation uses Google's Geocoding API to identify latitude and longitude positions for address, city or state, country or place names.

Other open geocoding APIs may be used in place of the Google geocoder.

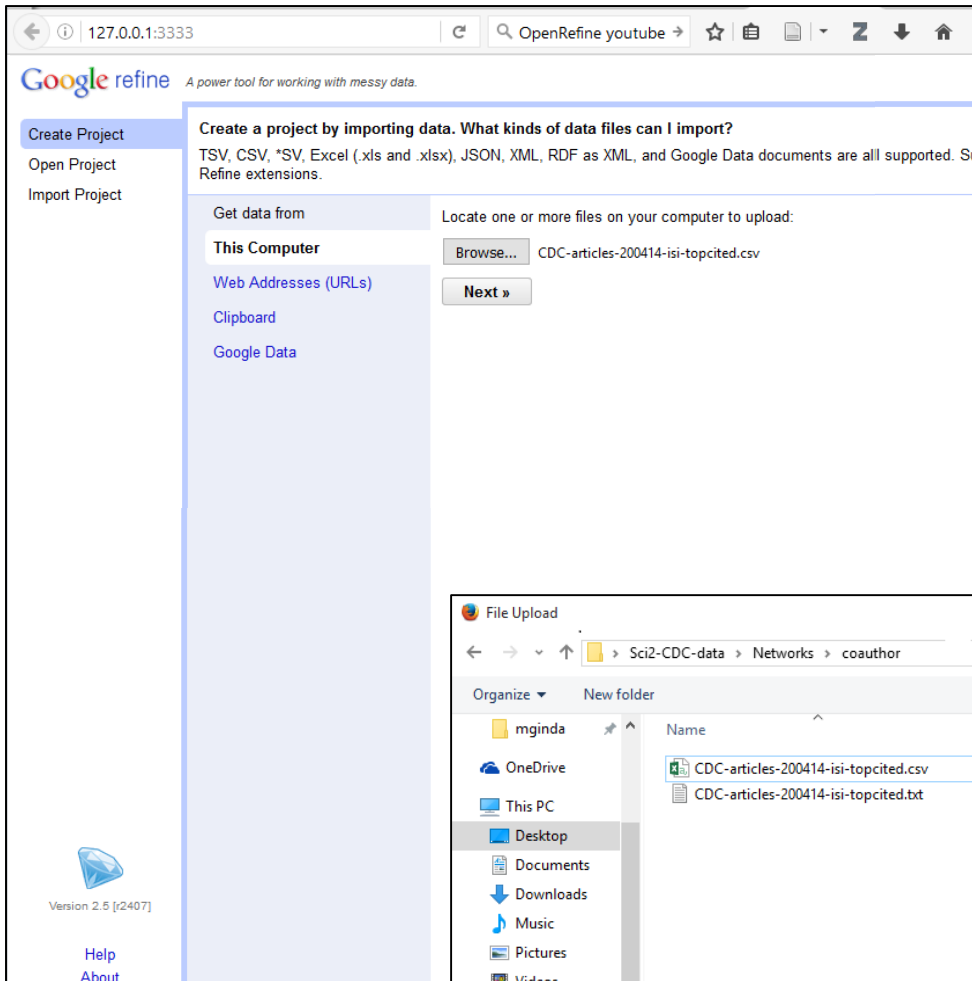


First we need to start OpenRefine. Navigate to the Desktop folder Refine.

In the directory, select the executable file “google-refine.exe”.

A console window will appear, this may be ignored for now, but can indicate error messages and process log.

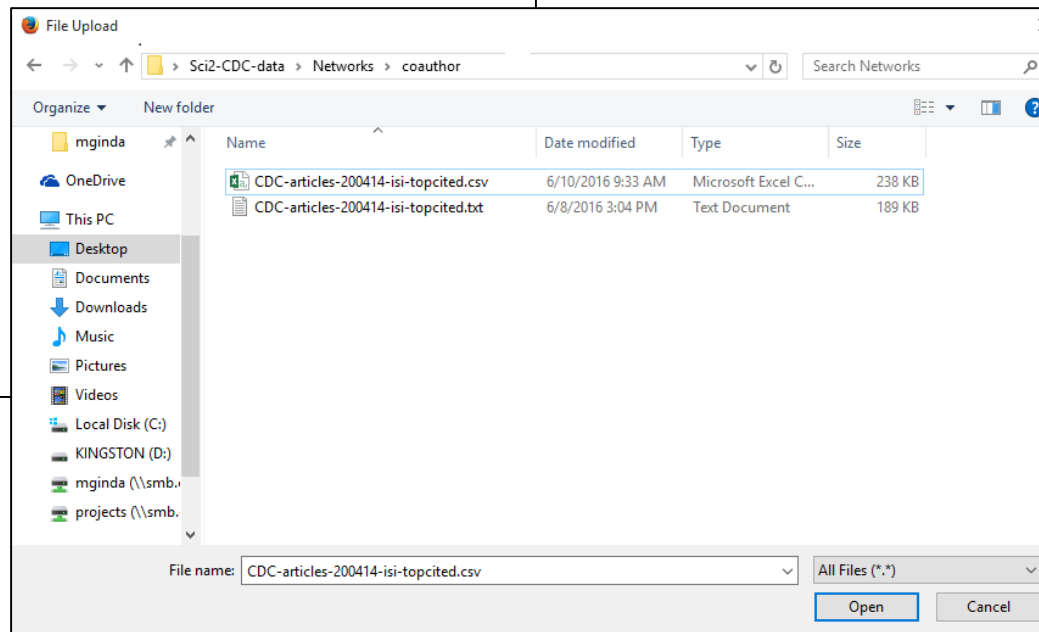
A new browser window will appear at the default domain 127.0.0.1:3333/3334



Once OpenRefine has started. You will want to create a project.

First, you need to select your data source, which in this case is a sample data file on this computer,
C:/.../Sci2-CDC-data > Networks > coauthor > “CDC-articles-200414-isi-topcited.csv”

Once the data file is uploaded, hit next.



Google refine *A power tool for working with messy data.*

Create Project « Start Over Configure Parsing Options Project name Create Project »

Open Project
Import Project

	Abstract	Associated Group	Author Identifiers	Authors	Authors (Full Names)	Beginning Page	BIOSIS Citation Index	Book Author	Book Authc
1.				Roffes, Melissa Blanton, Lenee Brammer, Lynnette Smith, Sophie Mustaquim, Desiree Steffens, Craig Cohen, Jessica Leon, Michelle Chaves, Sandra S. Abd Elal, Anwar Isaa Gubareva, Larisa Hall, Henrietta Wallis, Teresa Villanueva, Julie Xu, Xiyao Bresee, Joseph Cox, Nancy Finelli, Lyn		1189	9		
2.				Ridpath, Alison Driver, Cynthia R. Nolan, Michelle L. Karpati, Adam Kass, Daniel Paone, Denise Jakubowski, Andrea Hoffman, Robert S. Nelson, Lewis S. Kunin, Hillary V.		1195	2		
3.				Summers, Aimee Nyenswah, Tolbert G. Montgomery, Joel M. Neatherlin, John Tappero, Jordan W.		1202	4		

Line-based text files
Fixed-width field text files
PC-Axis text files
JSON files
RDF/N3 files
XML files
Open Document Format spreadsheets (.ods)

commas (CSV)
 tabs (TSV)
 custom , _____
Escape special characters with \

Parse next 1 line(s) as column headers
 Discard initial 0 row(s) of data
 Load at most 0 row(s) of data

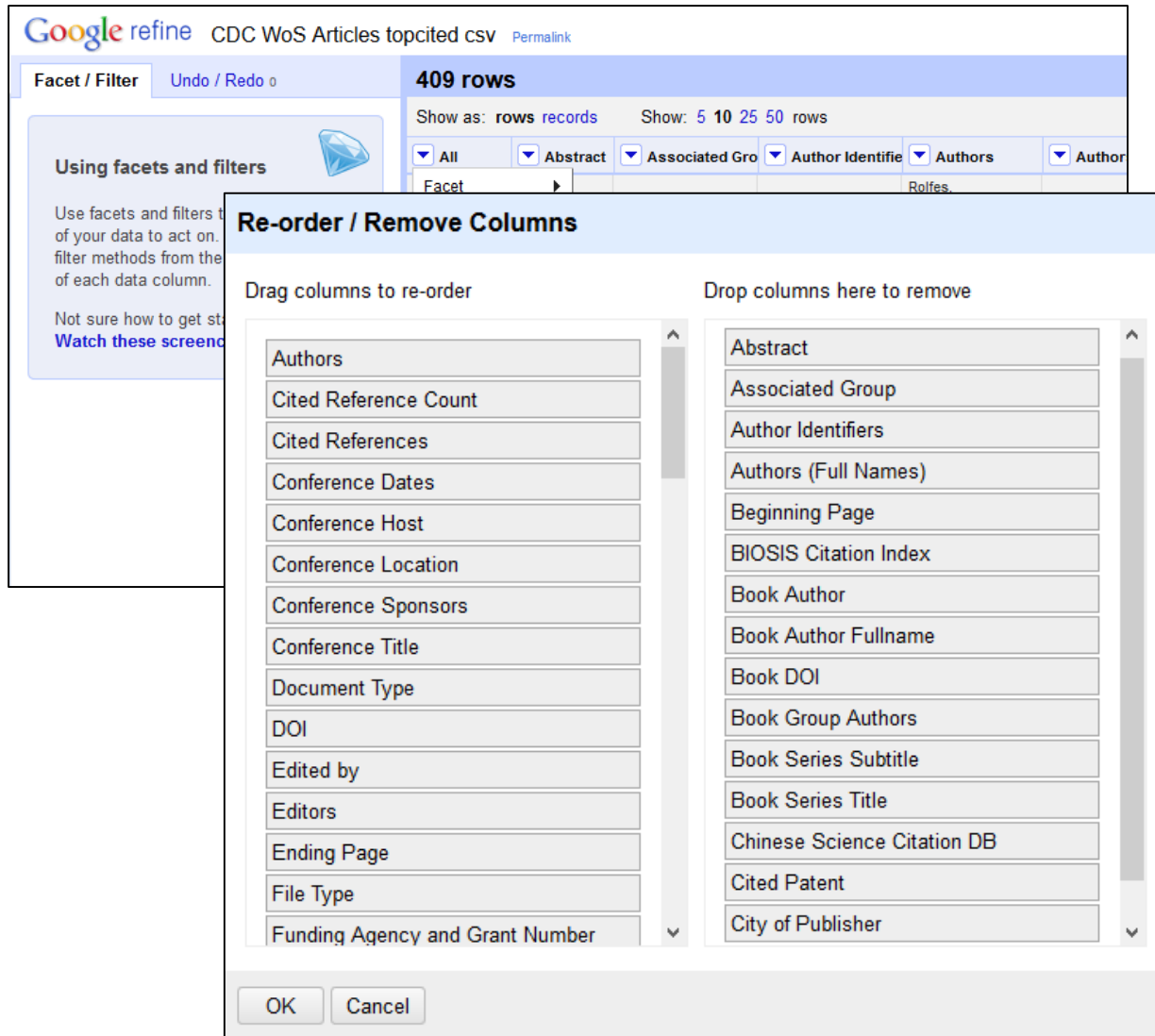
Parse cell text into numbers, dates, ...
 Quotation marks are used to enclose cells containing column separators

Store blank rows
 Store blank cells as nulls
 Store file source (file names, URLs) in each row

Once the data is loaded, it has to be parsed by OpenRefine. In this case, the data is in TSV format, and has column headers for the data.

The base setting for a TSV file are going to be all that is required. In the right hand corner, select **“Create Project”**

There are a number of parsing options beyond TSV/CSV.



Google refine CDC WoS Articles topcited csv Permalink

Facet / Filter Undo / Redo 0

409 rows

Show as: rows records Show: 5 10 25 50 rows

Facet: All Abstract Associated Gro Author Identifie Authors Author

Roles.

Re-order / Remove Columns

Drag columns to re-order

- Authors
- Cited Reference Count
- Cited References
- Conference Dates
- Conference Host
- Conference Location
- Conference Sponsors
- Conference Title
- Document Type
- DOI
- Edited by
- Editors
- Ending Page
- File Type
- Funding Agency and Grant Number

Drop columns here to remove

- Abstract
- Associated Group
- Author Identifiers
- Authors (Full Names)
- Beginning Page
- BIOSIS Citation Index
- Book Author
- Book Author Fullname
- Book DOI
- Book Group Authors
- Book Series Subtitle
- Book Series Title
- Chinese Science Citation DB
- Cited Patent
- City of Publisher

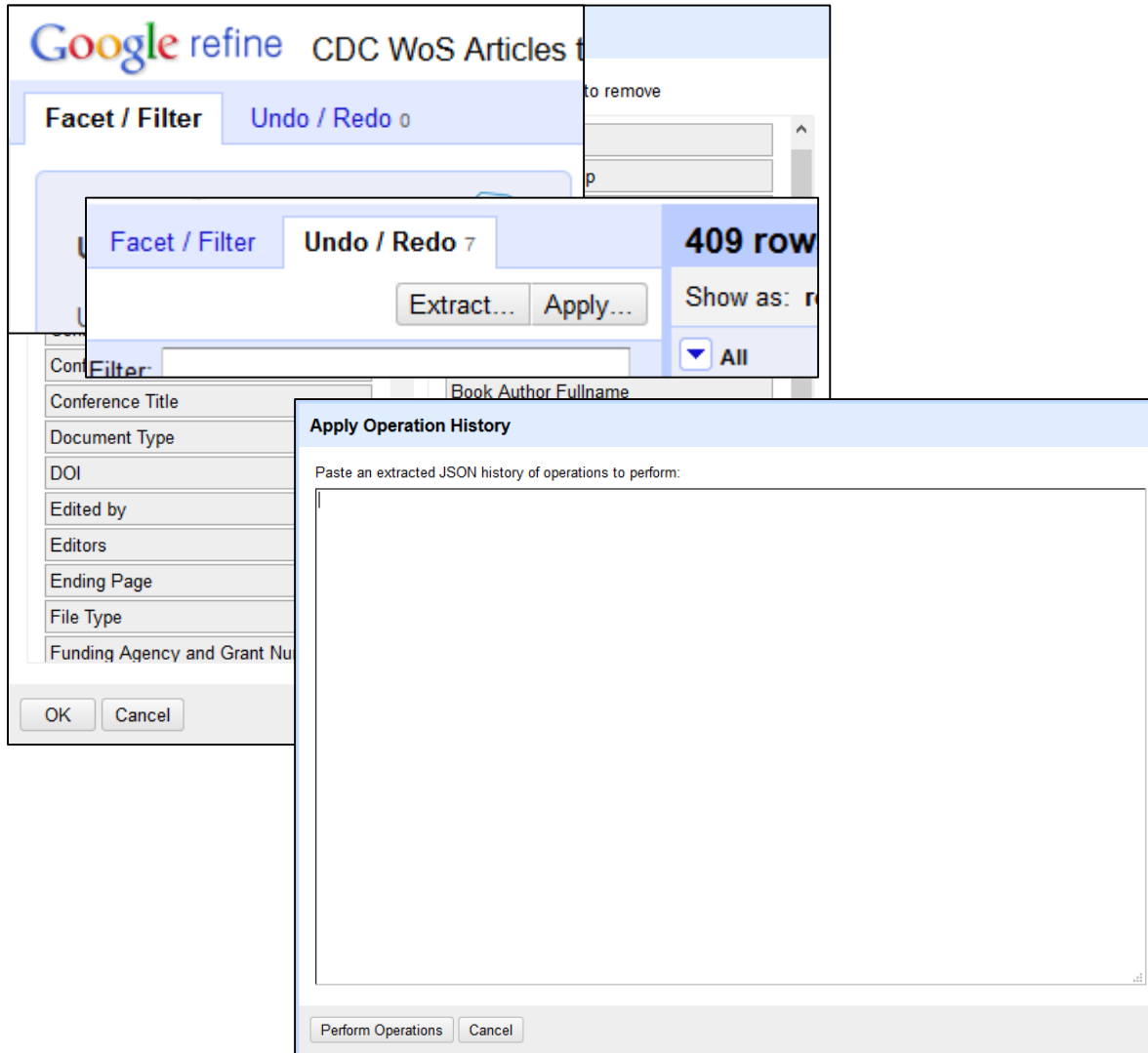
OK Cancel

Data files downloaded from ISI in the plain text format, for which this CSV file was derived, contain many irrelevant columns to our analysis.

To quickly remove columns, navigate to the **All** column's drop down menu and navigate to **Edit columns> Re-order/remove columns...**

This allows you to quickly drag and drop columns that need to be removed. The left box also allows repositioning of columns.

In this case, we will want to move the UniqueID field to the top of the first column for cluster analysis of author names.



The screenshot displays the Open Refine web interface. The main window shows a list of columns for 'CDC WoS Articles' with a 'Facet / Filter' and 'Undo / Redo' tab. A dialog box titled 'Apply Operation History' is open, prompting the user to 'Paste an extracted JSON history of operations to perform:'. The dialog has 'Perform Operations' and 'Cancel' buttons at the bottom. The background interface shows a table with columns like 'Conference Title', 'Document Type', 'DOI', etc., and a '409 row' indicator.

To speed up the process, in advance of the workshop, I have identified the all columns that need to be removed, and saved a JSON script that will allow you to duplicate this data preprocessing task.

First in Open Refine, in the left tab navigate to **Undo/Redo**.

In the **Undo/Redo** tab, select the **Apply...** button. A box will appear in the main data screen for you to copy a JSON text to duplicate data operations performed for similar data.

Apply Operation History

Paste an extracted JSON history of operations to perform:

```

1  [
2  {
3  "op": "
4  "descri
5  "column
6  "Auth
7  "Cite
8  "Cite
9  "Conf
10 "Docu
11 "DOI"
12 "Edit
13 "Edit
14 "Fund
15 "Fund
16 "ISI
17 "ISBN
18 "ISSN
19 "Jour
20 "Jour
21 "Jour
22 "Lang
23 "New
24 "New
25 "Orig
26 "Publ
27 "Publication type",
28 "Publication Year",
29 "Publisher",
30 "Research Field",
31 "Researcher ID",
32 "Subject Category",
33 "Times Cited",
34 "Title",
35 "Total Times Cited",
36 "Unique ID",
37 "File Name",
38 "Cite Me As"
39 }
40 },
41 {
42 "op": "core/column-reorder",

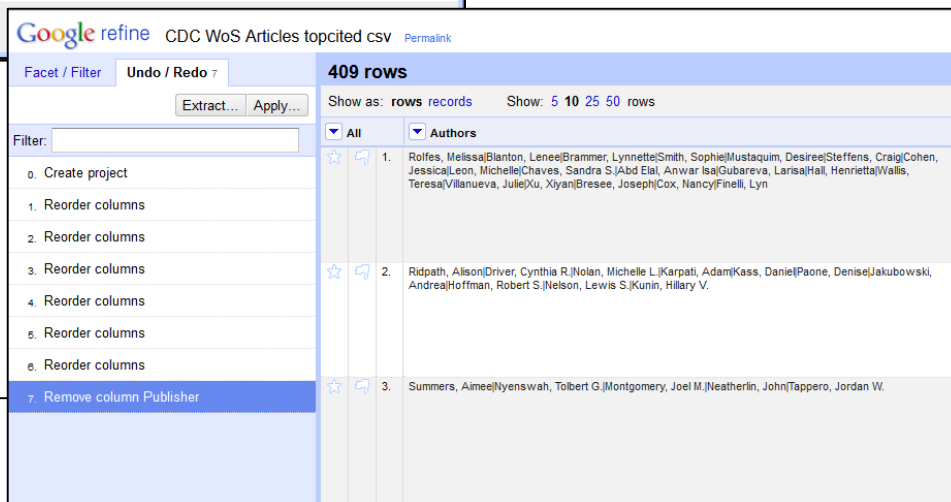
```

Perform Operations Cancel

To apply the re-order tasks, navigate to **C:/.../Sci2-CDC-data > Networks > coauthor > refine > CDC-WoS-JSON-1-Preprocessing.txt**, and open the file in Notepad.

Copy the text (Ctrl-c) and then Past (Ctrl-V) in the **Apply Operation History** window in OpenRefine.

Then select the **Perform Operations** button.



Google refine CDC WoS Articles topcited csv Permalink

Facet / Filter Undo / Redo 7 **409 rows**

Show as: rows records Show: 5 10 25 50 rows

Filter:

- 0 Create project
- 1 Reorder columns
- 2 Reorder columns
- 3 Reorder columns
- 4 Reorder columns
- 5 Reorder columns
- 6 Reorder columns
- 7 Remove column Publisher

All	Authors
1.	Rolfes, Melissa Blanton, Lenee Brammer, Lynnette Smith, Sophie Mustaquim, Desiree Steffens, Craig Cohen, Jessica Leon, Michelle Chaves, Sandra S. Abd Elal, Anwar Isa Gubareva, Larisa Hall, Henrietta Wallis, Teresa Vilanova, Jule Xu, Xiyao Bressee, Joseph Cox, Nancy Finelli, Lyn
2.	Ridpath, Alison Driver, Cynthia R. Nolan, Michelle L. Karpati, Adam Kass, Danie Paone, Denise Jakubowski, Andrea Hoffman, Robert S. Nelson, Lewis S. Kunin, Hillary V.
3.	Summers, Aimee lyenswah, Tolbert G. Montgomery, Joel M. Neatherlin, John Tappero, Jordan W.

OpenRefine provides a number of clustering algorithms that are designed to "finding groups of different values that might be alternative representations of the same thing."

In other words, the clustering algorithms can identify potential duplicate entities in a data set, including names, places, keywords and across controlled vocabularies.

OpenRefine notes in its documentation:

“...Clustering in OpenRefine works only at the syntactic level (the character composition of the cell value) and while very useful to spot errors, typos, and inconsistencies it's by no means enough to perform effective semantically-aware reconciliation.”

Cluster & Edit column "Authors"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

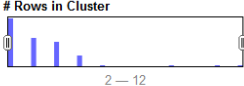
Method: nearest neighbor Distance Function: levenshtein Radius: 1.0 Block Chars: 6 67 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	2	<ul style="list-style-type: none"> Jiles, Ruth (1 rows) Jiles, Ruch (1 rows) 	<input type="checkbox"/>	Jiles, Ruth
2	6	<ul style="list-style-type: none"> Hlavsa, Michele C. (4 rows) Hlavsa, Michele C (2 rows) 	<input type="checkbox"/>	Hlavsa, Michele C.
2	2	<ul style="list-style-type: none"> Onwen, Diana H (1 rows) Onweh, Diana H (1 rows) 	<input type="checkbox"/>	Onwen, Diana H
2	2	<ul style="list-style-type: none"> Blau, Dianna M. (1 rows) Blau, Dianna M (1 rows) 	<input type="checkbox"/>	Blau, Dianna M.
2	4	<ul style="list-style-type: none"> Richardson, Lisa C (2 rows) Richardson, Lisa C. (2 rows) 	<input type="checkbox"/>	Richardson, Lisa C
2	3	<ul style="list-style-type: none"> Meltzer, Martin I. (2 rows) Meltzer, Martin I (1 rows) 	<input type="checkbox"/>	Meltzer, Martin I.
2	4	<ul style="list-style-type: none"> Frieden, Thomas R (3 rows) Frieden, Thomas R. (1 rows) 	<input type="checkbox"/>	Frieden, Thomas R

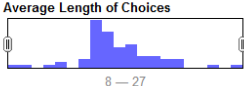
Select All Deselect All

Merge Selected & Re-Cluster Merge Selected & Close Close

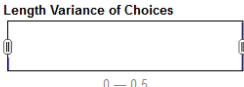
Rows in Cluster



Average Length of Choices



Length Variance of Choices



OpenRefine provides a number of clustering algorithms that are designed to "finding groups of different values that might be alternative representations of the same thing."

In other words, the clustering algorithms can identify potential duplicate entities in a data set, including names, places, keywords and across controlled vocabularies.

OpenRefine notes in its documentation:

“...Clustering in OpenRefine works only at the syntactic level (the character composition of the cell value) and while very useful to spot errors, typos, and inconsistencies it's by no means enough to perform effective semantically-aware reconciliation.”

The next, we'll walk through preparing the data for clustering and editing a text field.

Then we'll discuss the various clustering algorithms and their strengths.

Last, we will try out the algorithms, and bulk apply author disambiguation for extracting a Co-Author network in Sci2.

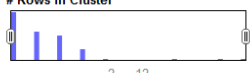
Cluster & Edit column "Authors"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method: nearest neighbor Distance Function: levenshtein Radius: 1.0 Block Chars: 6 67 clusters found

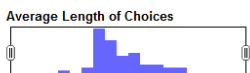
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	2	<ul style="list-style-type: none"> Jiles, Ruth (1 rows) Jiles, Ruch (1 rows) 	<input type="checkbox"/>	Jiles, Ruth
2	6	<ul style="list-style-type: none"> Hlavsa, Michele C. (4 rows) Hlavsa, Michele C (2 rows) 	<input type="checkbox"/>	Hlavsa, Michele C.
2	2	<ul style="list-style-type: none"> Onwen, Diana H (1 rows) Onweh, Diana H (1 rows) 	<input type="checkbox"/>	Onwen, Diana H
2	2	<ul style="list-style-type: none"> Blau, Dianna M. (1 rows) Blau, Dianna M (1 rows) 	<input type="checkbox"/>	Blau, Dianna M.
2	4	<ul style="list-style-type: none"> Richardson, Lisa C (2 rows) Richardson, Lisa C. (2 rows) 	<input type="checkbox"/>	Richardson, Lisa C
2	3	<ul style="list-style-type: none"> Meltzer, Martin I. (2 rows) Meltzer, Martin I (1 rows) 	<input type="checkbox"/>	Meltzer, Martin I.
2	4	<ul style="list-style-type: none"> Frieden, Thomas R (3 rows) Frieden, Thomas R. (1 rows) 	<input type="checkbox"/>	Frieden, Thomas R

Rows in Cluster



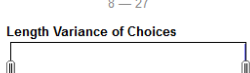
2 — 12

Average Length of Choices



8 — 27

Length Variance of Choices



0 — 0.5

Select All Deselect All Merge Selected & Re-Cluster Merge Selected & Close Close

409 rows

Show as: rows records Show: 5 10 25 50 rows

All	Unique ID	Authors
☆	1. WOS:000346946700002	Facet Text filter Edit cells Edit column Transpose Sort... View Reconcile
☆	2. WOS:000346946700003	Facet Text filter Edit cells Edit column Transpose Sort... View Reconcile
☆	3. WOS:000346946700005	Facet Text filter Edit cells Edit column Transpose Sort... View Reconcile

Facet menu options:
 Facet
 Text filter
 Edit cells
 Edit column
 Transpose
 Sort...
 View
 Reconcile

Edit cells sub-menu:
 Transform...
 Common transforms
 Fill down
 Blank down
 Split multi-valued cells...
 Join multi-valued cells...
 Cluster and edit...

In OpenRefine, first navigate to the **Authors** column drop down menu **Edit cells > Split multi-valued cells...**

In the window, change the comma (,) to a pipe (|) character. Then select **OK**.

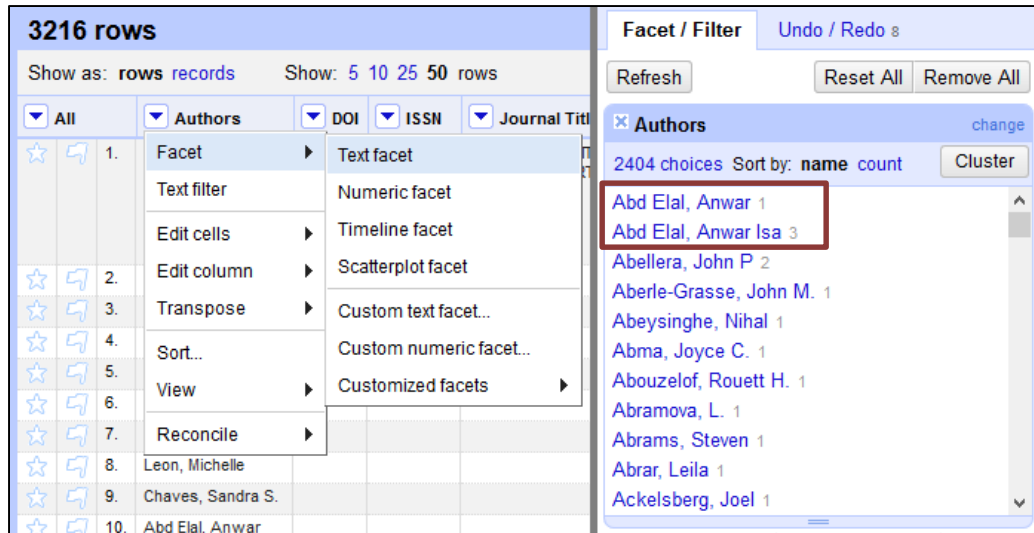
What separator currently separates the values?

OK Cancel

409 records

Show as: rows records Show: 5 10 25 50 records

All	Unique ID	Authors	DOI	ISSN	Journal
☆	1. WOS:000346946700002	Rolfes, Melissa		0149-2195	MMWR-MOR MORTALITY
☆		Blanton, Lenee			
☆		Brammer, Lynnette			
☆		Smith, Sophie			
☆		Mustaquim, Desiree			
☆		Steffens, Craig			
☆		Cohen, Jessica			
☆		Leon, Michelle			
☆		Chaves, Sandra S.			
☆		Abd Elal, Anwar Isa			
☆		Gubareva, Larisa			
☆		Hall, Henrietta			
☆		Wallis, Teresa			



3216 rows

Show as: rows records Show: 5 10 25 50 rows

Facet / Filter Undo / Redo 8

Refresh Reset All Remove All

Authors 2404 choices Sort by: name count Cluster

- Abd Elal, Anwar 1
- Abd Elal, Anwar Isa 3**
- Abellera, John P 2
- Aberle-Grasse, John M. 1
- Abeyasinghe, Nihal 1
- Abma, Joyce C. 1
- Abouzelof, Rouett H. 1
- Abramova, L. 1
- Abrams, Steven 1
- Abrar, Leila 1
- Ackelsberg, Joel 1

1. Facet Text facet

2. Text filter Numeric facet

3. Edit cells Timeline facet

4. Edit column Scatterplot facet

5. Transpose Custom text facet...

6. Sort... Custom numeric facet...

7. View Customized facets

8. Reconcile

8. Leon, Michelle

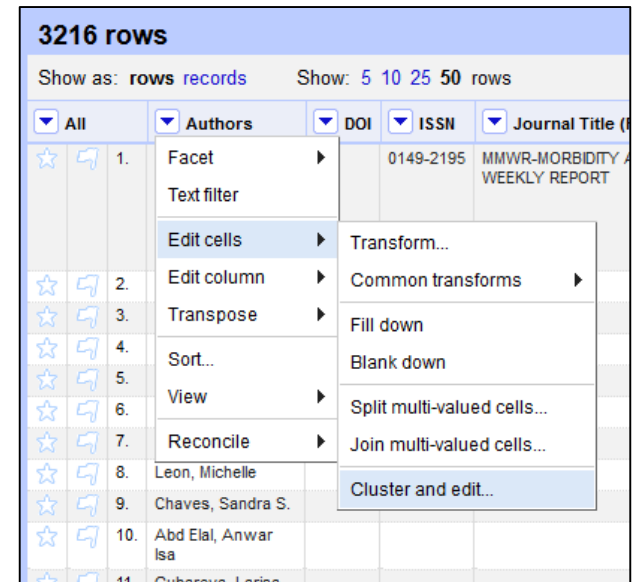
9. Chaves, Sandra S.

10. Abd Elal, Anwar

To track the effects of your merges, it is nice to have a Text Facet of the field you are editing.

The facet can help identify if there are potential names to be disambiguated.

To perform a cluster and edit on a cell, navigate to **Edit cells > Cluster and edit...**



3216 rows

Show as: rows records Show: 5 10 25 50 rows

All Authors DOI ISSN Journal Title

1. Facet Text filter

2. Edit cells Transform...

3. Edit column Common transforms

4. Transpose Fill down

5. Sort... Blank down

6. View Split multi-valued cells...

7. Reconcile Join multi-valued cells...

8. Leon, Michelle

9. Chaves, Sandra S.

10. Abd Elal, Anwar Isa

11. Gubareva, Larisa

Cluster and edit...

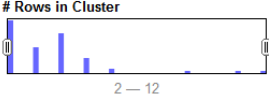
Cluster & Edit column "Authors"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method: key collision Keying Function: fingerprint 56 clusters found

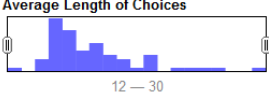
Cluster Size	Row Count	Values in Cluster	Keying Function	Cell Value
2	4	<ul style="list-style-type: none"> George, Mary G (2 rows) George, Mary G. (2 rows) 	fingerprint	ge, Mary G
2	4	<ul style="list-style-type: none"> Frieden, Thomas R (3 rows) Frieden, Thomas R. (1 rows) 	<input type="checkbox"/>	Frieden, Thomas R
2	4	<ul style="list-style-type: none"> Roy, Sharon L (2 rows) Roy, Sharon L. (2 rows) 	<input type="checkbox"/>	Roy, Sharon L
2	4	<ul style="list-style-type: none"> Kilmarx, Peter H. (3 rows) Kilmarx, Peter H (1 rows) 	<input type="checkbox"/>	Kilmarx, Peter H.
2	2	<ul style="list-style-type: none"> McGuire, Lisa C (1 rows) McGuire, Lisa C. (1 rows) 	<input type="checkbox"/>	McGuire, Lisa C
2	3	<ul style="list-style-type: none"> Hicks, Lauri A (2 rows) Hicks, Lauri A. (1 rows) 	<input type="checkbox"/>	Hicks, Lauri A
2	2	<ul style="list-style-type: none"> Yu, Patricia A (1 rows) Yu, Patricia A. (1 rows) 	<input type="checkbox"/>	Yu, Patricia A

Rows in Cluster



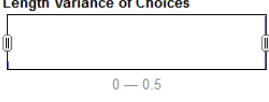
2 — 12

Average Length of Choices



12 — 30

Length Variance of Choices



0 — 0.5

Select All Deselect All Merge Selected & Re-Cluster Merge Selected & Close Close

Key Collision

Key collision methods are based on the idea of creating an alternative representation of a value (a "key") that contains only the most valuable or meaningful part of the string.

Keys are placed in 'bucket's (or 'bin' as it's described inside OpenRefine's code) together different strings based on the fact that their key is the same (hence the name "key collision").

This class of methods is the fastest in OpenRefine because its computational complexity is linear in the number of values processed and can produce results in seconds even with millions of values to cluster.

Cluster & Edit column "Authors"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method: key collision Keying Function: fingerprint 56 clusters found

Cluster Size	Row Count	Values in Cluster	Keying Function	Cell Value
2	4	<ul style="list-style-type: none"> George, Mary G (2 rows) George, Mary G. (2 rows) 	fingerprint	ge, Mary G
2	4	<ul style="list-style-type: none"> Frieden, Thomas R (3 rows) Frieden, Thomas R. (1 rows) 	<input type="checkbox"/>	Frieden, Thomas R
2	4	<ul style="list-style-type: none"> Roy, Sharon L (2 rows) Roy, Sharon L. (2 rows) 	<input type="checkbox"/>	Roy, Sharon L
2	4	<ul style="list-style-type: none"> Kilmarx, Peter H. (3 rows) Kilmarx, Peter H (1 rows) 	<input type="checkbox"/>	Kilmarx, Peter H.
2	2	<ul style="list-style-type: none"> McGuire, Lisa C (1 rows) McGuire, Lisa C. (1 rows) 	<input type="checkbox"/>	McGuire, Lisa C
2	3	<ul style="list-style-type: none"> Hicks, Lauri A (2 rows) Hicks, Lauri A. (1 rows) 	<input type="checkbox"/>	Hicks, Lauri A
2	2	<ul style="list-style-type: none"> Yu, Patricia A (1 rows) Yu, Patricia A. (1 rows) 	<input type="checkbox"/>	Yu, Patricia A

Select All Deselect All

Merge Selected & Re-Cluster Merge Selected & Close Close

Rows in Cluster

Average Length of Choices

Length Variance of Choices

Fingerprint

Fingerprinting key collision is a fast and simple method for clustering.

The process that generates the key from a string value is the following (note that the order of these operations is significant):

- remove leading and trailing whitespace
- change all characters to their lowercase representation
- remove all punctuation and control characters
- split the string into whitespace-separated tokens
- sort the tokens and remove duplicates
- join the tokens back together
- normalize extended western characters to their ASCII representation (for example "gödel" → "godel")

NGRAM-Fingerprint

The n-gram fingerprint method is similar to the fingerprint method described above but instead of using whitespace separated tokens, it uses n-grams, where the n (or the size in chars of the token) can be specified by the user.

The algorithm:

- change all characters to their lowercase representation
- remove all punctuation, whitespace, and control characters
- obtain all the string n-grams
- sort the n-grams and remove duplicates
- join the sorted n-grams back together
- normalize extended western characters to their ASCII representation

Cluster & Edit column "Authors"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Godel" and "Gödel" probably refer to the same person. [Find out more ...](#)

Method: key collision Keying Function: ngram-fingerprint Ngram Size: 1 31 clusters found

Cluster	Members	Key
2	• Miles, Iss (1 rows)	
2	• Biermann, Janis (1 rows) • Miner, James B (1 rows)	Biermann, Janis
2	• Kallen, A. (1 rows) • Keenan, N. L. (1 rows)	Kallen, A.
2	• Rosenman, Kenneth (1 rows) • Thomas, Karen E. (1 rows)	Rosenman, Kenneth
2	• Kahn, Katherine E. (1 rows) • Tan, Kathrine R. (1 rows)	Kahn, Katherine E.
2	• Carter, Marion (1 rows)	Carter, Marion

Choices in Cluster (Range: 2 - 4)

Rows in Cluster (Range: 2 - 6)

Average Length of Choices (Range: 7 - 20)

So, for example, the 2-gram fingerprint of "Paris" is "arispari" and the 1-gram fingerprint is "aiprs".

Why is this useful? In practice, using big values for n-grams doesn't yield any advantage over the previous fingerprint method, but using 2-grams and 1-grams, while yielding many false positives, can find clusters that the previous method didn't find even with strings that have small differences, with a very small performance price.

Phonetic fingerprint

A third keying method uses a phonetic fingerprinting (specifically, [Metaphone3](#) method for English and the [Cologne phonetic keyer](#) for German), which is a way to transform tokens into the way they are pronounced.

This is useful to spot errors that are due to people misunderstanding or not knowing the spelling of a word after only hearing it. The idea being that similar sounding words will end up sharing the same key and thus being binned in the same cluster.

Cluster & Edit column "Authors"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example "New York" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person.

Method: key collision Keying Function: metaphone3

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
3	3	<ul style="list-style-type: none"> Freedman, David (1 rows) Freedman, David O. (1 rows) Freedman, David S. (1 rows) 	<input type="checkbox"/>	Freedman, David

Cluster & Edit column "Authors"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example "New York" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person.

Method: key collision Keying Function: cologne-phonetic

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	3	<ul style="list-style-type: none"> Lu, Hua (2 rows) Luo, Yao-Hua (1 rows) 	<input type="checkbox"/>	Lu, Hua

Cluster & Edit column "Authors"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method: nearest neighbor Distance Function: levenshtein Radius: 1.0 Block Chars 6 67 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	2	<ul style="list-style-type: none"> Jiles, Ruth (1 rows) Jiles, Ruch (1 rows) 	<input type="checkbox"/>	Jiles, Ruth
2	6	<ul style="list-style-type: none"> Hlavsa, Michele C. (4 rows) Hlavsa, Michele C (2 rows) 	<input type="checkbox"/>	Hlavsa, Michele C.
2	2	<ul style="list-style-type: none"> Onwen, Diana H (1 rows) Onweh, Diana H (1 rows) 	<input type="checkbox"/>	Onwen, Diana H
2	2	<ul style="list-style-type: none"> Blau, Dianna M. (1 rows) Blau, Dianna M (1 rows) 	<input type="checkbox"/>	Blau, Dianna M.
2	4	<ul style="list-style-type: none"> Richardson, Lisa C (2 rows) Richardson, Lisa C. (2 rows) 	<input type="checkbox"/>	Richardson, Lisa C
2	3	<ul style="list-style-type: none"> Meltzer, Martin I. (2 rows) Meltzer, Martin I (1 rows) 	<input type="checkbox"/>	Meltzer, Martin I.
2	4	<ul style="list-style-type: none"> Frieden, Thomas R (3 rows) Frieden, Thomas R. (1 rows) 	<input type="checkbox"/>	Frieden, Thomas R

Rows in Cluster
2 — 12

Average Length of Choices
8 — 27

Length Variance of Choices
0 — 0.5

Select All Deselect All Merge Selected & Re-Cluster Merge Selected & Close Close

Nearest Neighbor

While key collisions methods are very fast, they tend to be either too strict or too lax with no way to fine tune how much difference between strings we are willing to tolerate.

The Nearest Neighbor methods (also known as kNN), on the other hand, provide a parameter (the radius, or k) which represents a distance threshold: any pair of strings that is closer than a certain value will be binned together.

To speed up processing, NN methods first implement blocking, a hybrid key collision and KNN method. Blocks are sequences of stings that share common substring of a certain size (e.g. 6 characters.)

Cluster & Edit column "Authors"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method: nearest neighbor Distance Function: levenshtein Radius: 3 Block Chars: 6 89 clusters filtered from 95 total

Cluster Size	Distance	Members	Selected	Representative
2	5	<ul style="list-style-type: none"> Seward, Jane F. (4 rows) Seward, Jane (1 rows) 	<input type="checkbox"/>	Seward, Jane F.
2	2	<ul style="list-style-type: none"> Fagan, Ryan (1 rows) Fagan, R (1 rows) 	<input type="checkbox"/>	Fagan, Ryan
2	6	<ul style="list-style-type: none"> Khan, Yosef (4 rows) Khan, Yosef M. (2 rows) 	<input checked="" type="checkbox"/>	Khan, Yosef
2	4	<ul style="list-style-type: none"> Keenan, Nora L. (3 rows) Keenan, N. L. (1 rows) 	<input checked="" type="checkbox"/>	Keenan, Nora L.
2	3	<ul style="list-style-type: none"> Mahoney, Frank (2 rows) Mahoney, Frank J. (1 rows) 	<input checked="" type="checkbox"/>	Mahoney, Frank
2	2	<ul style="list-style-type: none"> Watson, John (1 rows) Watson, J (1 rows) 	<input type="checkbox"/>	Watson, John
2	2	<ul style="list-style-type: none"> Faul, Mark (1 rows) Faul, Mark D. (1 rows) 	<input checked="" type="checkbox"/>	Faul, Mark

Choices in Cluster

Rows in Cluster

Average Length of Choices

Length Variance of Choices

89 clusters filtered from 95 total

Buttons: Select All Deselect All Merge Selected & Re-Cluster Merge Selected & Close Close

Levenshtein Distance

The Levenshtein distance (also known as "edit distance") is probably the simplest and most intuitive distance function between strings and is often still very effective due to its general applicability.

It measures the minimal number of 'edit operations' that are required to change one string into the other.

For example, "Paris" and "paris" have an edit distance of 1 as changing P into p is the only operation required.

"New York" and "newyork" has edit distance 3: 2 substitutions and 1 removal.

"Al Pacino" and "Albert Pacino" have an edit distance of 4 because it requires 4 insertions.

PPM

Prediction by Partial Matching is an implementation of the Kolmogorov complexity estimating similarity between strings, initially used for comparing DNA sequences.

The algorithm text compressors estimate the information content of two strings to tell if they are identical.

OpenRefine implements a normalized version:

$$\mathbf{distance(A,B) = comp(A+B) + comp(B+A) / (comp(A+A) + comp(B+B))}$$

where $comp(s)$ is the length of bytes of the compressed sequence of the string s and $+$ is the append operator. This is used to account for deviation in optimality of the given compressors.

409 records

Show as: [rows](#) [records](#) Show: [5](#) [10](#) [25](#) [50](#) records

All	Unique ID	Authors	DOI	ISSN	Journal
☆	1.	WOS:000346946700002	Rolfes, Melissa		0149-2195 MMWR-MOR MORTALITY
☆			Blanton, Lenee		
☆			Brammer, Lynnette		
☆			Smith, Sophie		
☆			Mustaquim, Desiree		
☆			Steffens, Craig		
☆			Cohen, Jessica		
☆			Leon, Michelle		
☆			Chaves, Sandra S.		
☆			Abd Elal, Anwar Isa		
☆			Gubareva, Larisa		
☆			Hall, Henrietta		
☆			Wallis, Teresa		



3216 rows

Show as: [rows](#) [records](#) Show: [5](#) [10](#) [25](#) [50](#) rows

All	Unique ID	Authors	DOI	ISSN	Journal Title (Full)
☆	1.	WOS:000346946700002		0149-2195	MMWR-MORBIDITY AND MORTALITY WEEKLY REPORT
☆	2.				
☆	3.				
☆	4.				
☆	5.				
☆	6.				
☆	7.				
☆	8.	Leon, Michelle			
☆	9.	Chaves, Sandra S.			
☆	10.	Abd			
☆	11.	Gub			
☆	12.	Hall			

What separator currently separates the values?

OK Cancel

opocited csv [Permalink](#) **Join multi-valued cells in column Authors** [Undo](#)

409 records

Show as: [rows](#) [records](#) Show: [5](#) [10](#) [25](#) [50](#) records

All	Unique ID	Authors	DOI	
☆	1.	WOS:000346946700002	Rolfes, Melissa Blanton, Lenee Brammer, Lynnette Smith, Sophie Mustaquim, Desiree Steffens, Craig Cohen, Jessica Leon, Michelle Chaves, Sandra S. Abd Elal, Anwar Isa Gubareva, Larisa Hall, Henrietta Wallis, Teresa Villanueva, Julie Xiyang, Xiyang Bresee, Joseph Cox, Nancy Finelli, Lyn	0149
☆	2.	WOS:000346946700003	Ridpath, Alison Driver, Cynthia R. Nolan, Michelle L. Karpati, Adam Kass, Daniel Paone, Denise Jakubowski, Andrea Hoffman, Robert S. Nelson, Lewis S. Kunin, Hillary V.	0149



▼ All	▼ Unique ID	▼ Authors
Facet ▶	0346946700002	Rolfes, Melissa Blanton, L Desiree Steffens, Craig Co Isa Gubareva, Larisa Hall, JosephiCox, Nancy Finelli,
Edit rows ▶		
Edit columns ▶	Re-order / remove columns...	
View ▶		
☆ ↶	2. WOS:000346946700003	Ridpath, Alison Driver, Cyn Denise Jakubowski, Andri


Next, reorder the remaining columns by moving **Unique ID** to the end of the list.

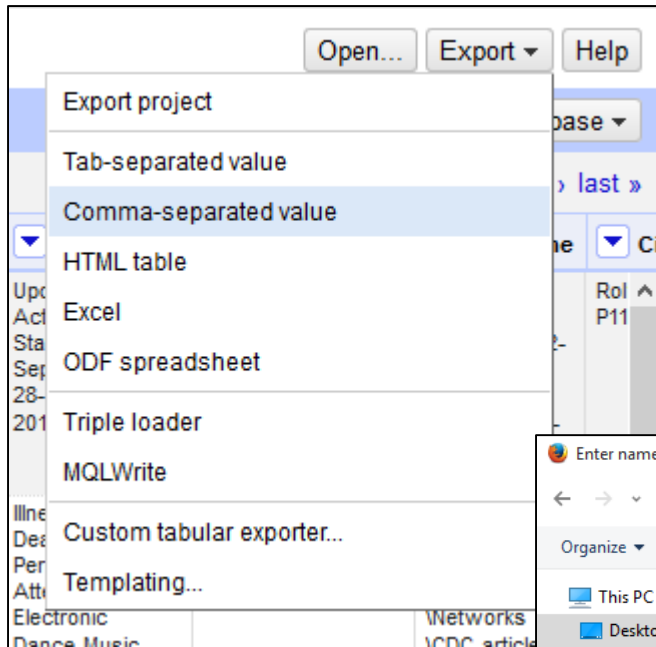
Re-order / Remove Columns

Drag columns to re-order

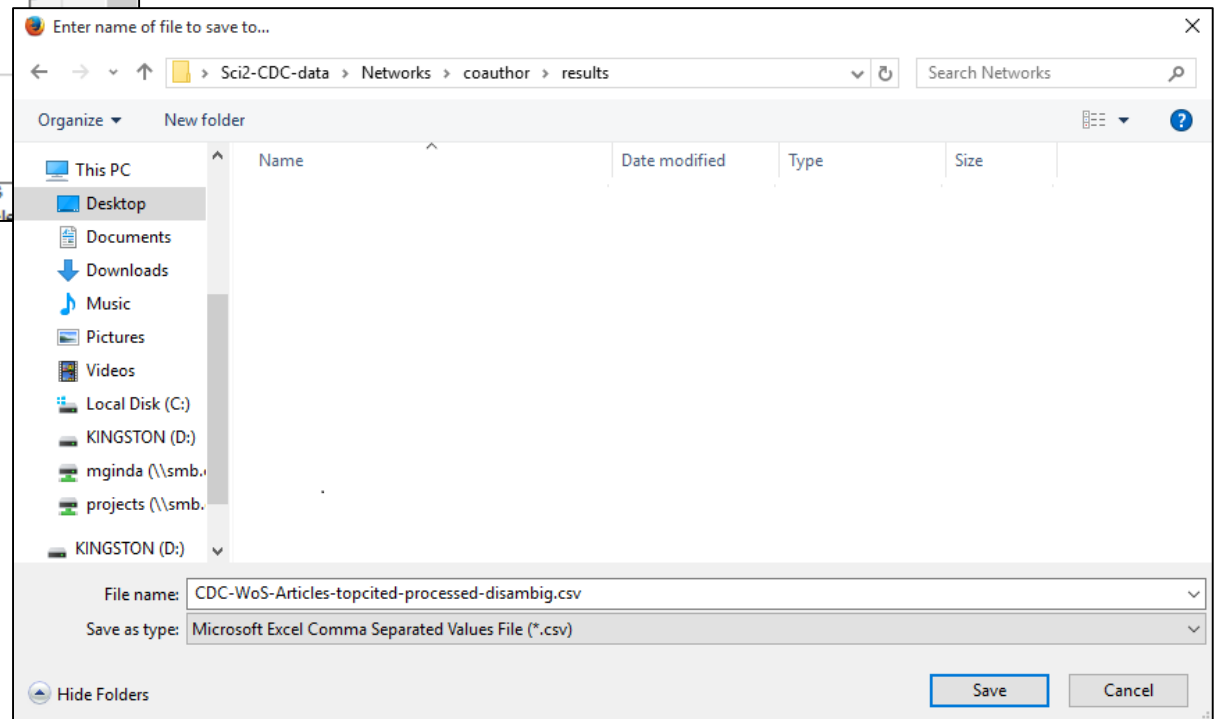
- Authors
- DOI
- ISSN
- Journal Title (Full)
- Publication Date
- Publication Type
- Publication Year
- Times Cited
- Title
- Total Times Cited
- File Name
- Cite Me As
- Unique ID

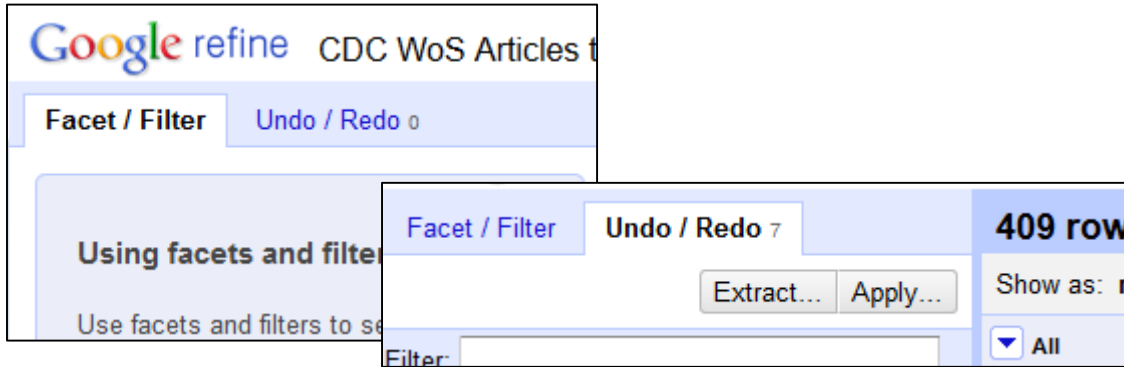
Drop columns here to remove





In OpenRefine, navigate to the upper right corner of the screen, and select the **Export** drop down menu, and select **Comma-separated value**. Save the file as below.

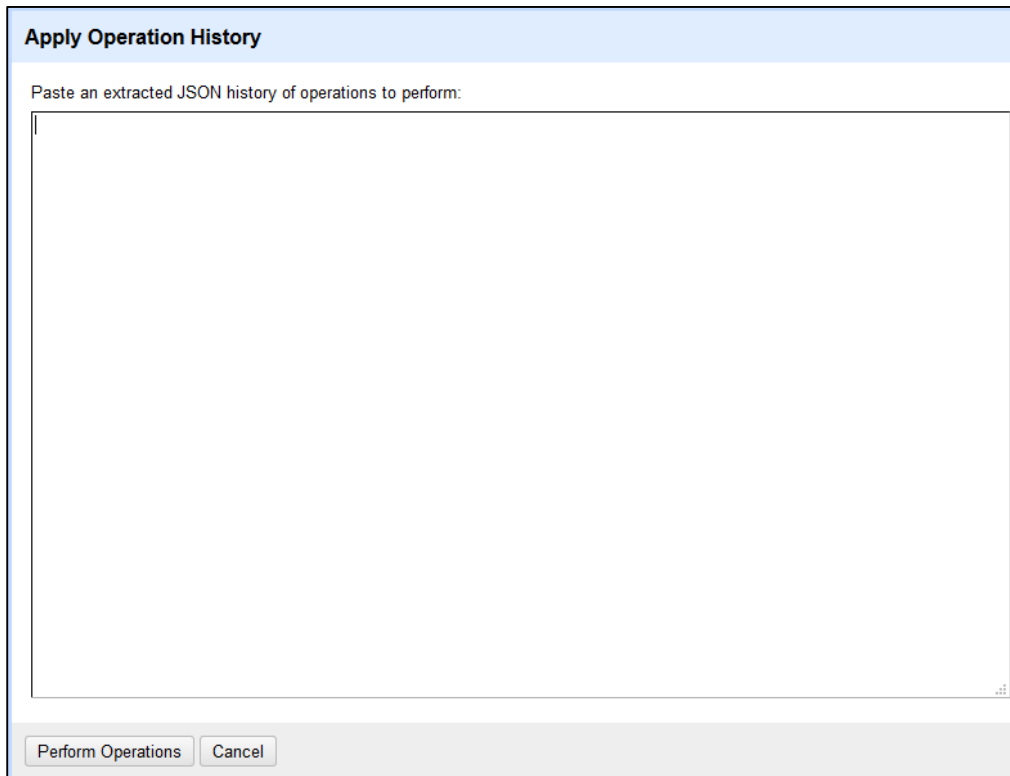




To speed up the process, in advance of the workshop, I have identified the all columns that need to be removed, and saved a JSON script that will allow you to duplicate this data preprocessing task.

First in Open Refine, in the left tab navigate to **Undo/Redo**.

In the **Undo/Redo** tab, select the **Apply...** button. A box will appear in the main data screen for you to copy a JSON text to duplicate data operations performed for similar data.



The screenshot shows the OpenRefine interface with a JSON file open in the editor. The file content is as follows:

```

1758 ],
1759   "to": "Ahluwalia, Indu B."
1760 }
1761 ],
1762 },
1763 {
1764   "op": "core/mass-edit",
1765   "description": "Mass edit cells in column Authors",
1766   "engineConfig": {
1767     "mode": "row-based",
1768     "facets": []
1769   },
1770   "columnName": "Authors",
1771   "expression": "value",
1772   "edit": {
1773     "expression": "value",
1774     "from": [
1775       "Thomas, Ann R.",
1776       "Thomas, Ann"
1777     ],
1778     "to": "Thomas, Ann R."
1779   },
1780 },
1781 ],
1782 },
1783 },
1784 {
1785   "op": "core/mass-edit",
1786   "description": "Mass edit cells in column Authors",
1787   "engineConfig": {
1788     "mode": "row-based",
1789     "facets": []
1790   },
1791   "columnName": "Authors",
1792   "expression": "value",
1793   "edit": {
1794     "expression": "value",
1795     "from": [
1796       "Moran, John",
1797       "Moran, John S."
1798     ],
1799     "to": "Moran, John"
1800   },
1801 },
1802 ],
1803 },
1804 {
1805   "op": "core/multivalued-cell-join",
1806   "description": "Join multi-valued cells in column Authors",
1807   "columnName": "Authors",
1808   "keyColumnName": "Unique ID",
1809   "separator": "|"
1810 },
1811 ],
1812 }

```

An "Apply Operation History" dialog box is overlaid on the editor. It contains the following JSON text:

```

{
  "expression": "value",
  "edits": [
    {
      "fromBlank": false,
      "fromError": false,
      "from": [
        "Thomas, Ann R.",
        "Thomas, Ann"
      ],
      "to": "Thomas, Ann R."
    },
    {
      "fromBlank": false,
      "fromError": false,
      "from": [
        "Moran, John",
        "Moran, John S."
      ],
      "to": "Moran, John"
    }
  ]
}

```

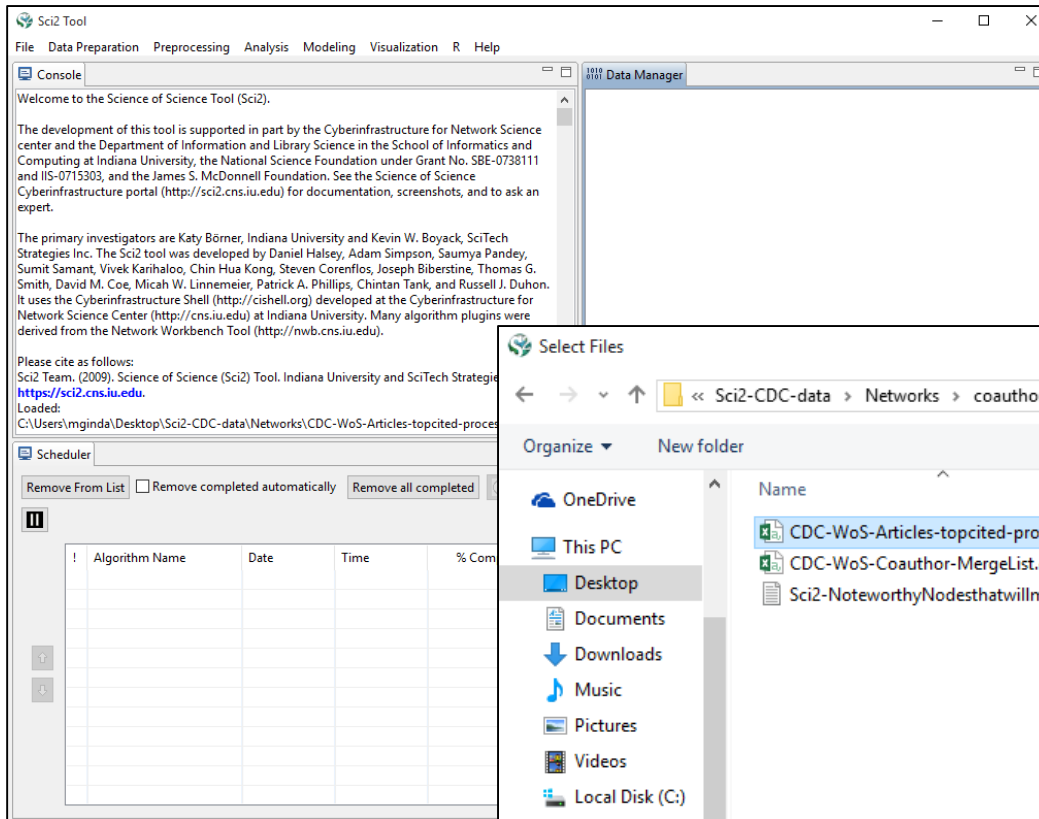
At the bottom of the dialog box, there are two buttons: "Perform Operations" and "Cancel".

To apply the re-order tasks, navigate to **C:/.../Sci2-CDC-data > Networks > coauthor > refine > CDC-WoS-JSON-2-ClusterEdit.txt**, and open the file in Notepad.

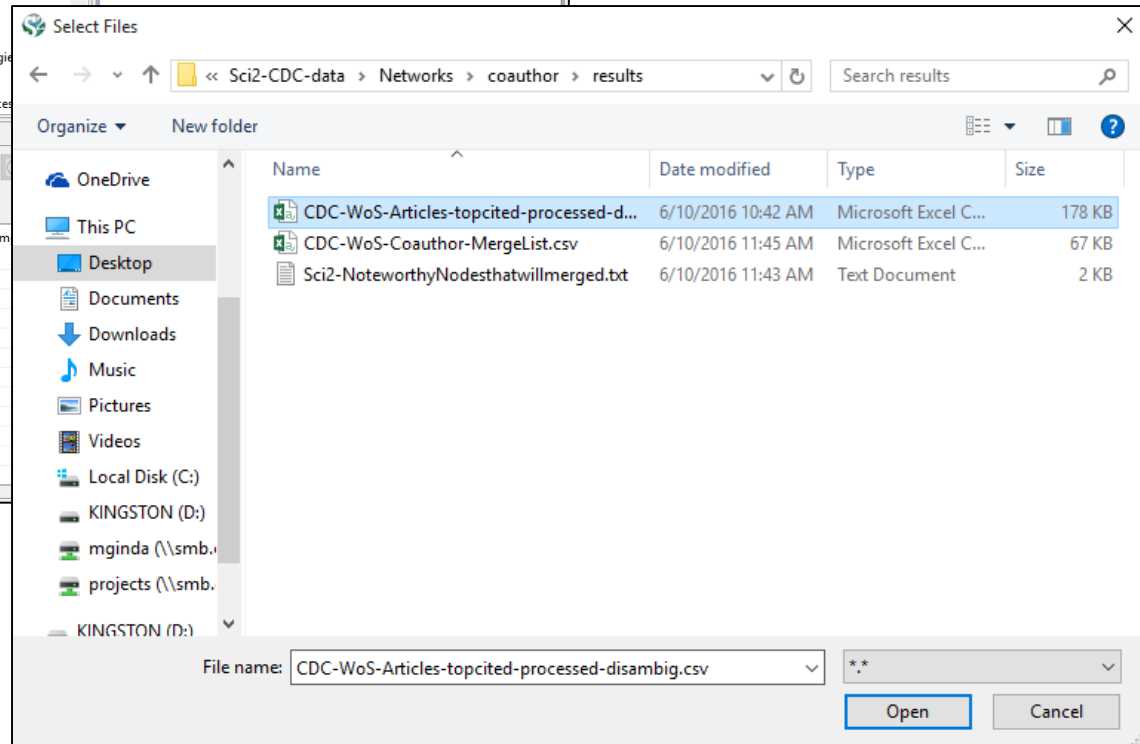
Copy the text (Ctrl-c) and then Past (Ctrl-V) in the **Apply Operation History** window in OpenRefine.

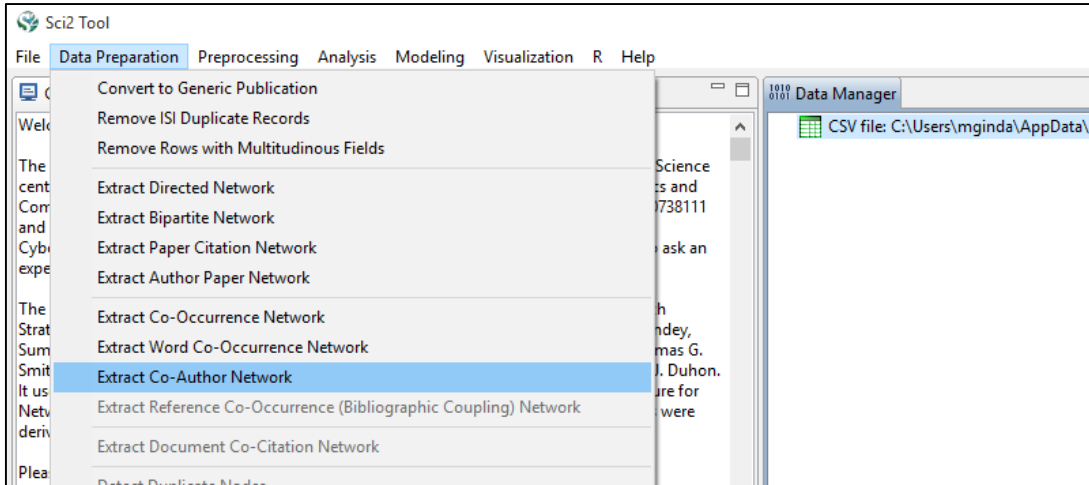
Then select the **Perform Operations** button.

Questions?



To extract a co-author network, first navigate to **C:/.../Sci2-CDC-data > Networks > coauthor > results > CDC-WoS-Articles-topcited-processed-disambig.csv**

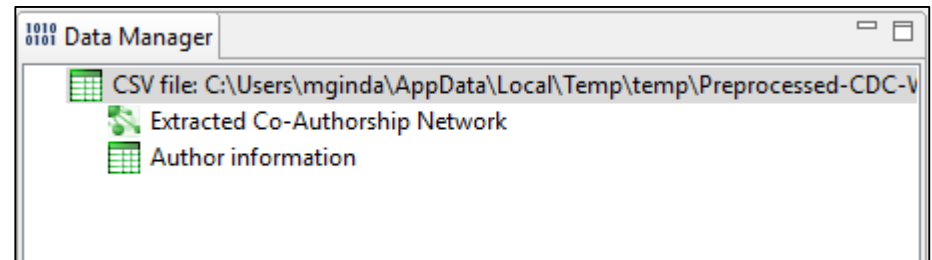
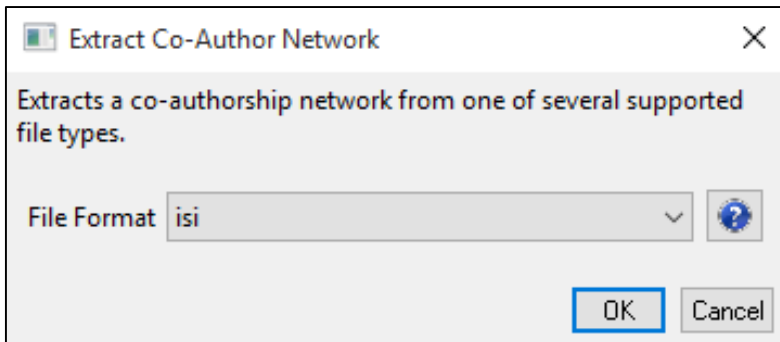


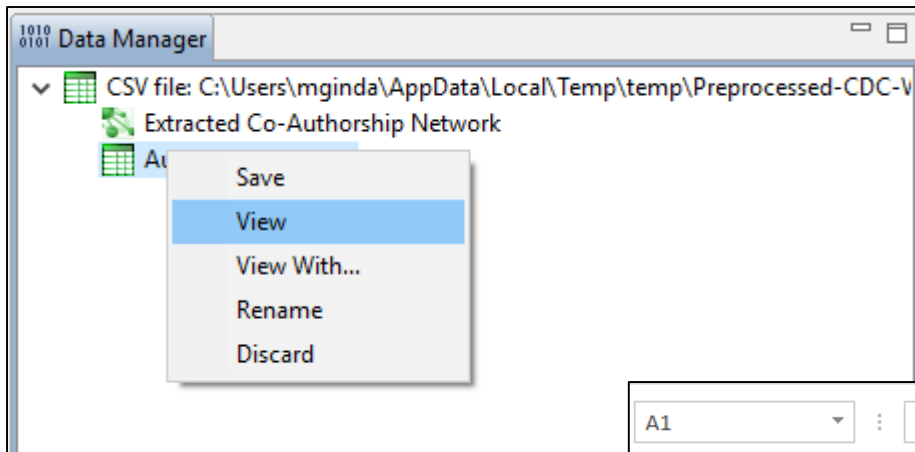


Select the data CDC articles in the Data manager and the then navigate in the menu **Data Preparation > Extract Co-Author Network**.

A pop-up window will appear; select the format ISI.

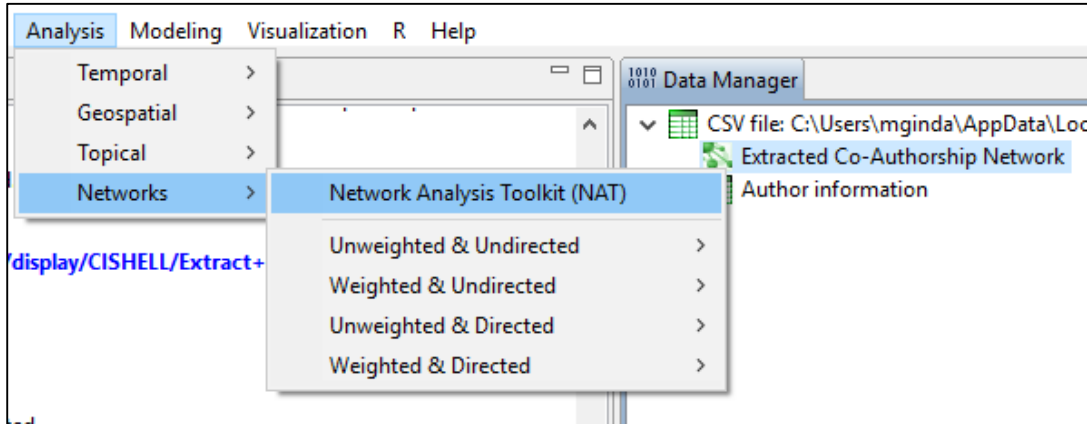
The output network and author list files will appear in the data manager below the original CSV.





To see the resulting node variables calculated for each author in the data set, right click on the Author list data table in the Data Manager, and select **View** or **View with...**

	A	B	C	D	E
1	label	number_of_authored_works	times_cited	uniqueIndex	combineVal
2	Finelli, Lyn	9	1389	1	*
3	Rolfes, Melissa	1	10	2	*
4	Blanton, Lenee	4	48	3	*
5	Brammer, Lynnette	4	48	4	*
6	Smith, Sophie	2	17	5	*
7	Mustaquim, Desiree	4	48	6	*
8	Steffens, Craig	4	48	7	*
9	Cohen, Jessica	2	15	8	*
10	Leon, Michelle	3	41	9	*
11	Chaves, Sandra S.	5	365	10	*
12	Abd Elal, Anwar Isa	3	30	11	*
13	Gubareva, Larisa	5	307	12	*
14	Hall, Henrietta	2	17	13	*
15	Wallis, Teresa	4	48	14	*
16	Villanueva, Julie	4	48	15	*
17	Xu, Xiyan	4	48	16	*
18	Bresee, Joseph	10	1712	17	*
19	Cox, Nancy	10	1995	18	*
20	Kunin, Hillary V.	1	5	19	*



After extracting a network, it is good practice to get the initial statistics for the network, before further processing and analysis.

Navigate to **Analysis > Networks > Network Analysis Toolkit (NAT)**.

To view the results, select the output file in the data manager and select **View** or **View with...**

```

This graph claims to be undirected.

Nodes: 2220
Isolated nodes: 4
Node attributes present: label, number_of_authored_works, times_cited

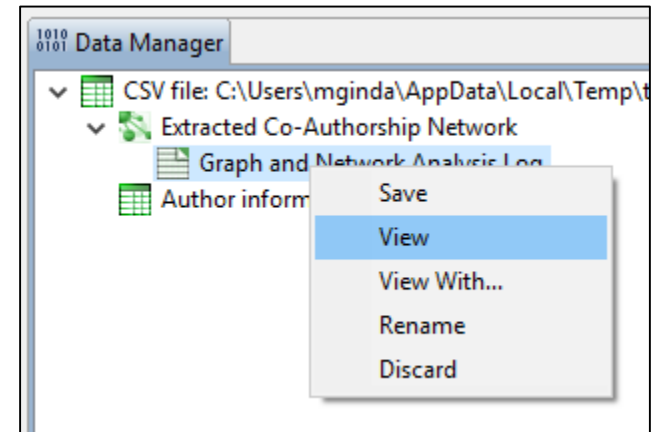
Edges: 16122
No self loops were discovered.
No parallel edges were discovered.

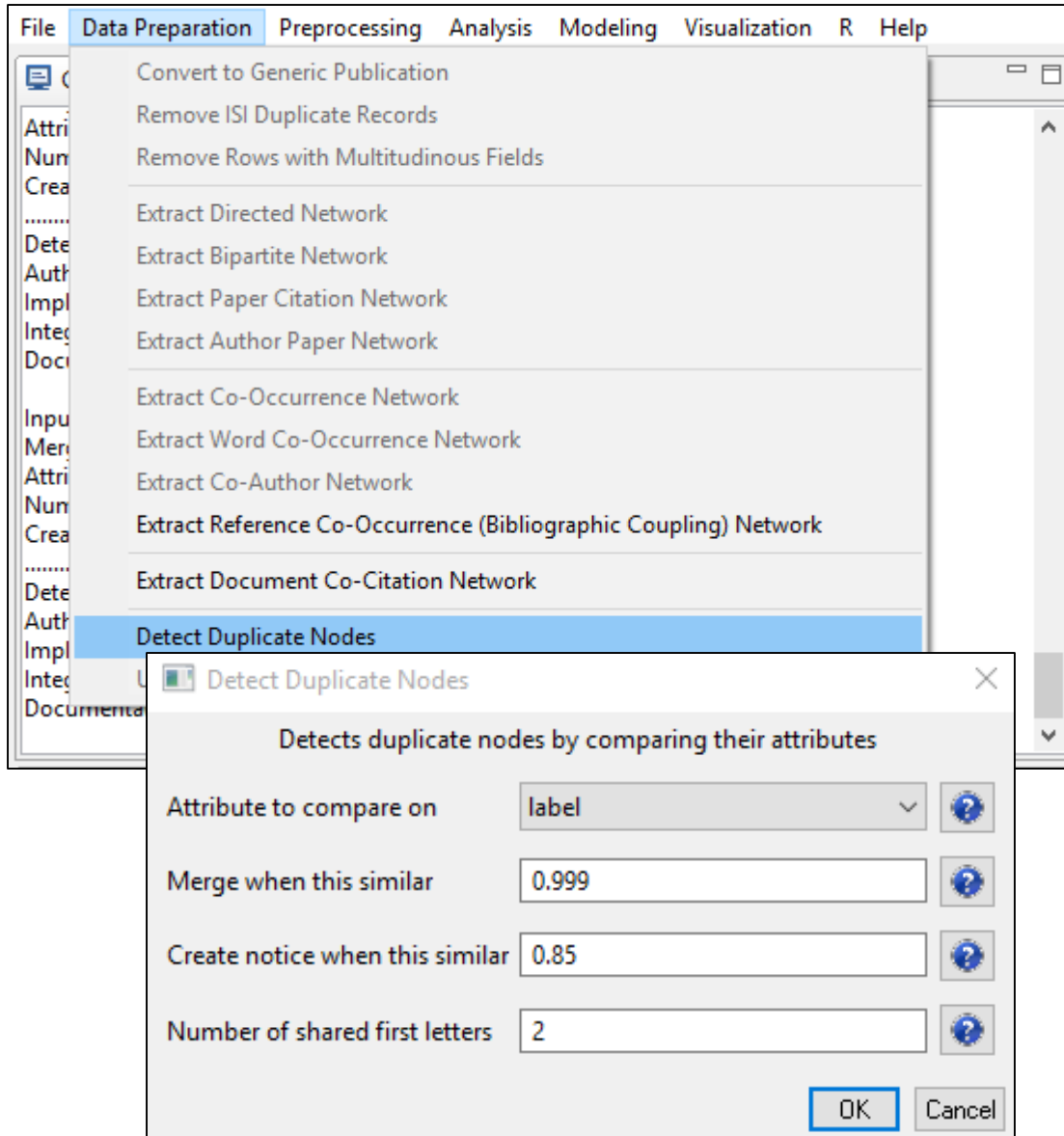
Edge attributes:
  Did not detect any nonnumeric attributes.
  Numeric attributes:
      number_...      min    max    mean
      weight         1     10    1.10811

      This network seems to be valued.

Average degree: 14.5243
This graph is not weakly connected.
There are 70 weakly connected components. (4 isolates)
The largest connected component consists of 1660 nodes.
Did not calculate strong connectedness because this graph was not directed.

Density (disregarding weights): 0.0065
Additional Densities by Numeric Attribute
    
```



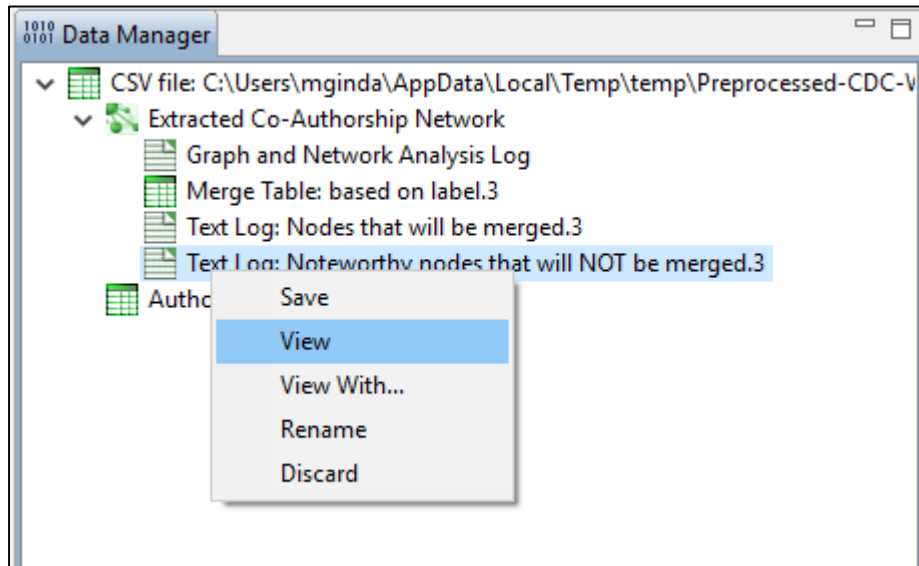


OpenRefine found a majority of the duplicate nodes in the network quickly. However, Sci2 also permits a simple duplicate node detection algorithm.

Select the network file in the data manager, and then navigate to **Data Preparation > Detect Duplicate Nodes** algorithm.

A pop-up box will let you set initial parameters. These are set to create three files: a merge list .txt list and a CSV that has updated duplicate node IDs for you when their match is above a similarity threshold of N%.

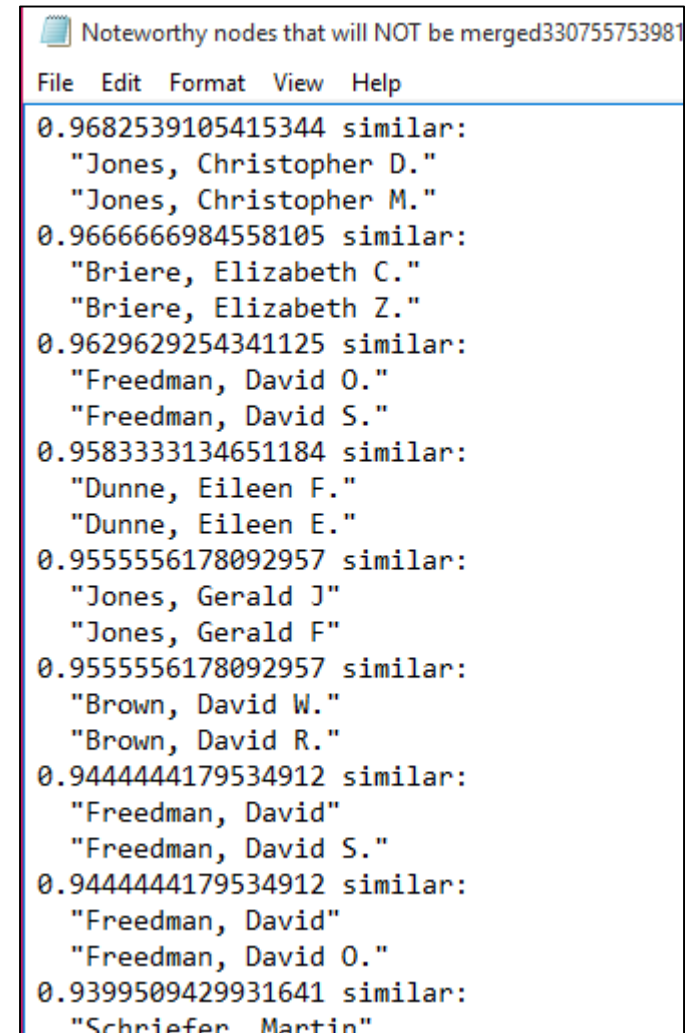
The second file lists potential node pairs for merging, and allows for a manual update to the automatically created merge list.



The initial parameters are set to allow for no merges in the automatically created file, which allows for a manual entry.

View the Not-merged list to review potential merges. In this case, I have already completed this process. You can view a list of merges that will be made by looking at file: **C:\...\Sci2-CDC-data\Networks\coauthor\results\Sci2-NoteworthyNodesthatwillmerged.txt.**

The updated merge list is also available here as: **CDC-WoS-Coauthor-MergeList.csv**



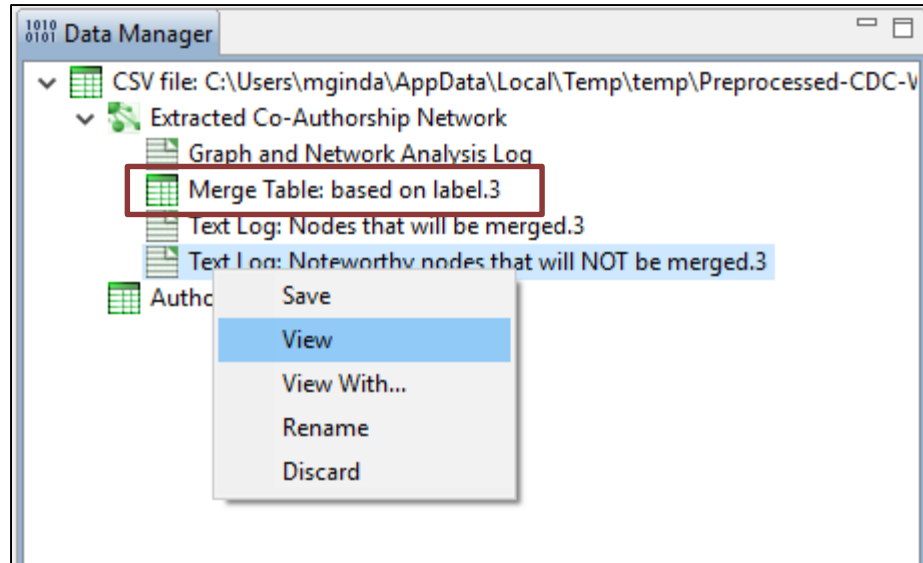
```

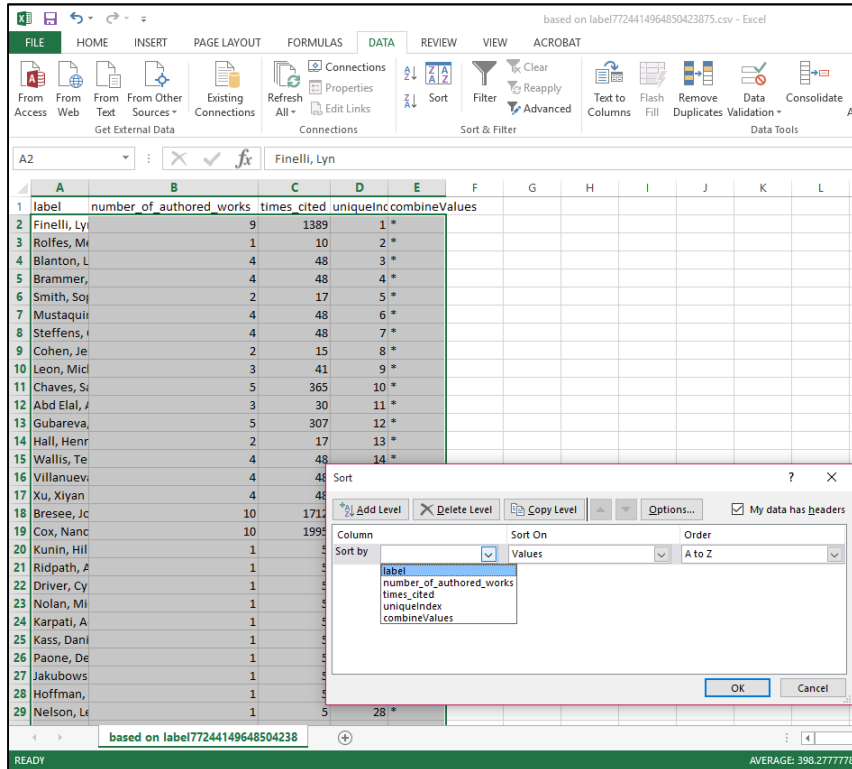
Noteworthy nodes that will NOT be merged330755753981
File Edit Format View Help
0.9682539105415344 similar:
"Jones, Christopher D."
"Jones, Christopher M."
0.9666666984558105 similar:
"Briere, Elizabeth C."
"Briere, Elizabeth Z."
0.9629629254341125 similar:
"Freedman, David O."
"Freedman, David S."
0.9583333134651184 similar:
"Dunne, Eileen F."
"Dunne, Eileen E."
0.9555556178092957 similar:
"Jones, Gerald J"
"Jones, Gerald F"
0.9555556178092957 similar:
"Brown, David W."
"Brown, David R."
0.9444444179534912 similar:
"Freedman, David"
"Freedman, David S."
0.9444444179534912 similar:
"Freedman, David"
"Freedman, David O."
0.9399509429931641 similar:
"Schriefer, Martin"

```

Using the the Not-merged list to review potential merges. The output from Sci2 can be edited as a text file, and allows you to first review and remove all potential merges that are not accurate.

With those that duplicate node pairs that remain, you now are able to update the Merge Table from Sci2. Right click the file in the data manager and save it as a CSV file.





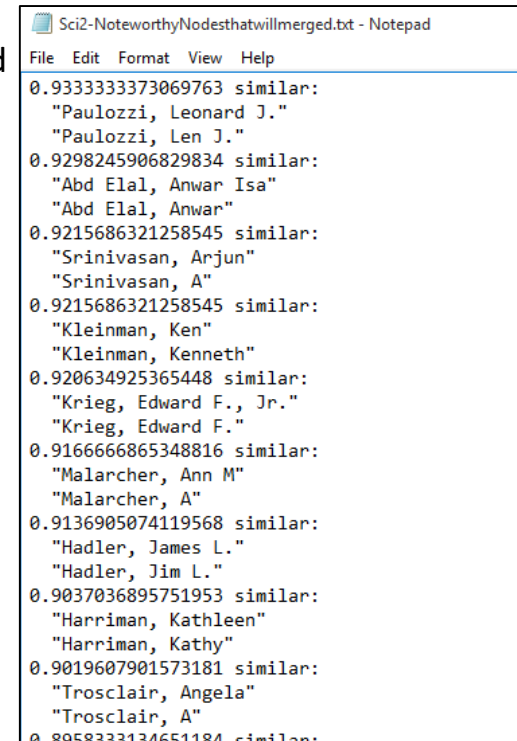
label	number_of_authored_works	times_cited	uniqueindex	combineValues
Finelli, Lyn	9	1389	1 *	
Rolfes, M	1	10	2 *	
Blanton, L	4	48	3 *	
Brammer,	4	48	4 *	
Smith, Soj	2	17	5 *	
Mustaquit	4	48	6 *	
Steffens, i	4	48	7 *	
Cohen, Je	2	15	8 *	
Leon, Micl	3	41	9 *	
Chaves, St	5	365	10 *	
Abd Elal, A	3	30	11 *	
Gubareva,	5	307	12 *	
Hall, Henr	2	17	13 *	
Wallis, Te	4	48	14 *	
Villanuev	4	48		
Xu, Xiyan	4	48		
Bressee, Jc	10	171		
Cox, Nanc	10	199		
Kunin, Hil	1			
Ridpath, A	1			
Driver, Cy	1			
Nolan, Mi	1			
Karpati, A	1			
Kass, Dani	1			
Paone, De	1			
Jakubows	1			
Hoffman,	1			
Nelson, Lt	1	5	28 *	

Open the merge list saved from Sci2 in a tabular data editor, such as Excel.

Sort the table by the **label** column from A to Z.

Then using the updated “Not-merged list” to update this the merge list.

In Excel, search for the authors listed by using the Find tool (Ctrl-F), and copy (Ctrl-C) the name from the list and select (Ctrl-C) **OK**.



```

Sci2-NoteworthyNodesThatwillmerged.txt - Notepad
File Edit Format View Help
0.9333333373069763 similar:
  "Paulozzi, Leonard J."
  "Paulozzi, Len J."
0.9298245906829834 similar:
  "Abd Elal, Anwar Isa"
  "Abd Elal, Anwar"
0.9215686321258545 similar:
  "Srinivasan, Arjun"
  "Srinivasan, A"
0.9215686321258545 similar:
  "Kleinman, Ken"
  "Kleinman, Kenneth"
0.920634925365448 similar:
  "Krieg, Edward F., Jr."
  "Krieg, Edward F."
0.9166666865348816 similar:
  "Malarcher, Ann M"
  "Malarcher, A"
0.9136905074119568 similar:
  "Hadler, James L."
  "Hadler, Jim L."
0.9037036895751953 similar:
  "Harriman, Kathleen"
  "Harriman, Kathy"
0.9019607901573181 similar:
  "Trosclair, Angela"
  "Trosclair, A"
0.8958333134651184 similar:
  
```

When you find a duplicate pair of nodes from your list “Not-merged list”, you will need to update the **uniqueIndex** and **combinedValues** columns.

1	label	number of authored works	times cited	uniqueIndex	combinedValue
2	Finelli, J	0	1200	1	*

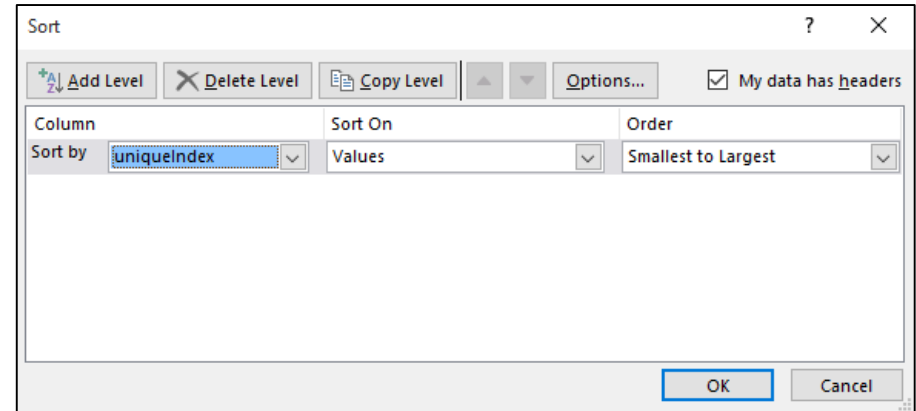
1) First copy (Ctrl-C) and paste (Ctrl-V) the **uniqueIndex** value of the node label that you would like to preserve. Then for the **uniqueIndex** value that you updated, erase the **combinedValue** *.

1063	Kitimbo, D		1	6	2024 *
1064	Kleinman, Ken		4	307	1538
1065	Kleinman, Kenneth		1	79	1538 *
1066	Klompas, Michael		5	196	560 *
1067	Knight, Nancy		1	42	420 *

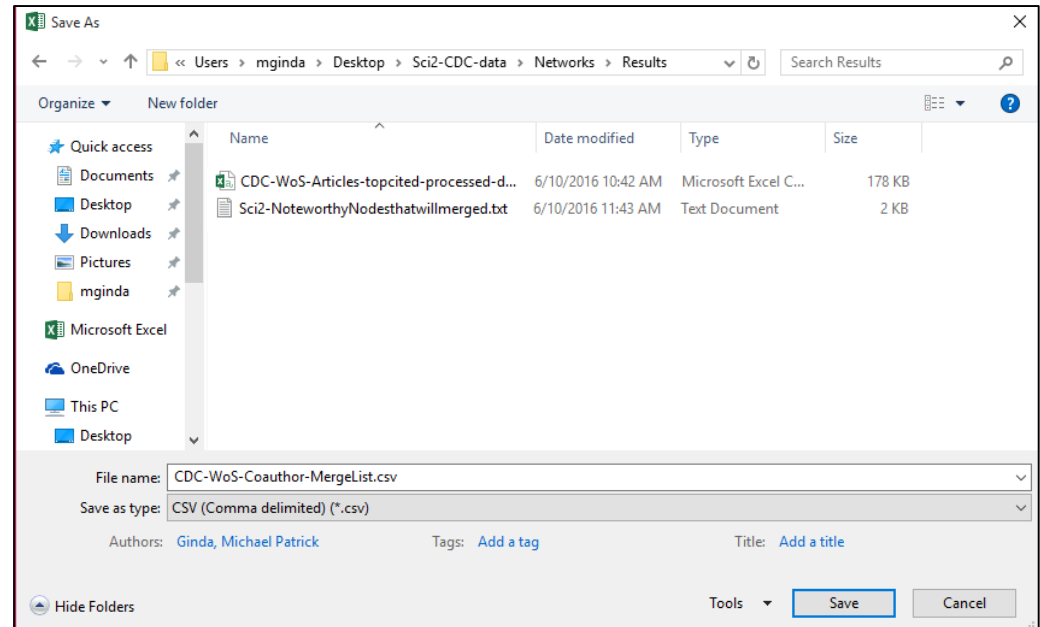
After updating a pair of **uniqueIndex** & **combinedValue** columns, search for the next author listed by using the Find tool (Ctrl-F), and copy (Ctrl-C) the name from the list and select (Ctrl-C) **OK**, and repeat the process listed above in 1). Remember, not every pair is necessarily good, sometimes, you end up with pairs such as below where review of a copy of a publication is needed to confirm a merge (which is too detailed for most analysis).

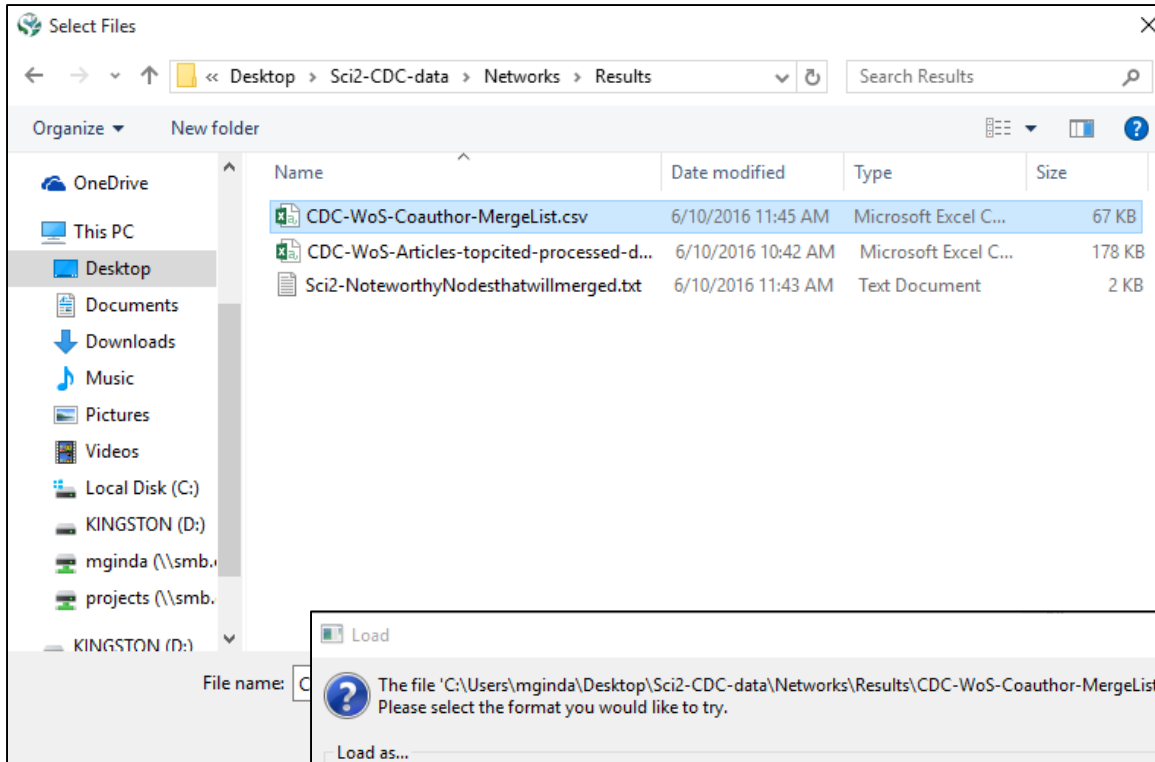
542	Fagan, Jennifer L.		1	21	627 *
543	Fagan, R		1	7	2015 *
544	Fagan, Robert F		2	76	2097 *
545	Fagan, Ryan		1	43	1038 *
546	Fahnbulleh, Miatta		1	11	199 *

Last, selecting all of the data, resort the Merge list by the **uniqueIndex** values from smallest to largest.

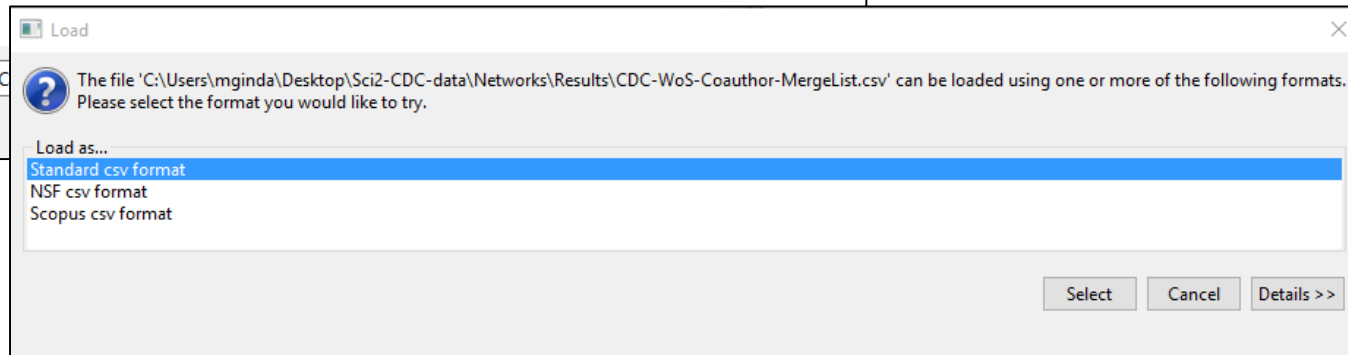


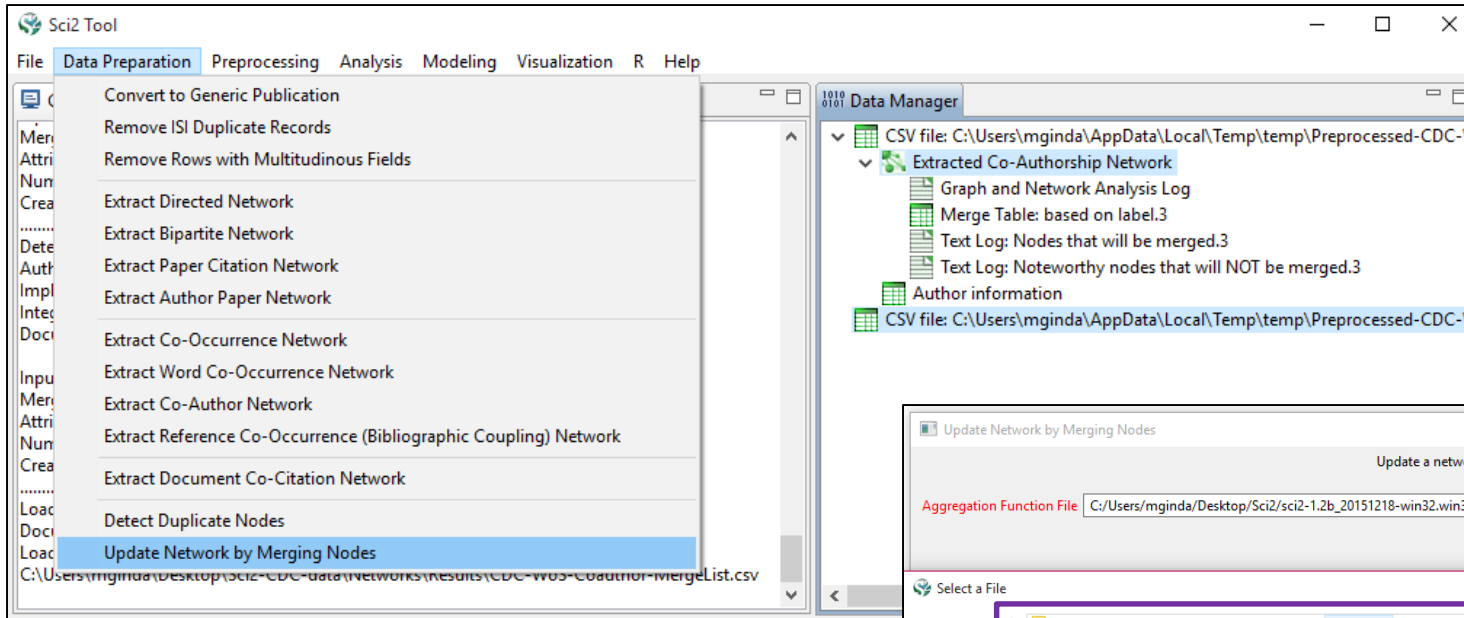
Then save the resulting file in
**C:\...\Sci2-CDC-data\Networks\coauthor\results\
 CDC-WoS-Coauthor-MergeList.csv**





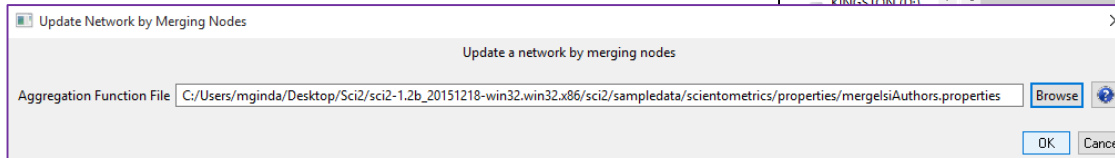
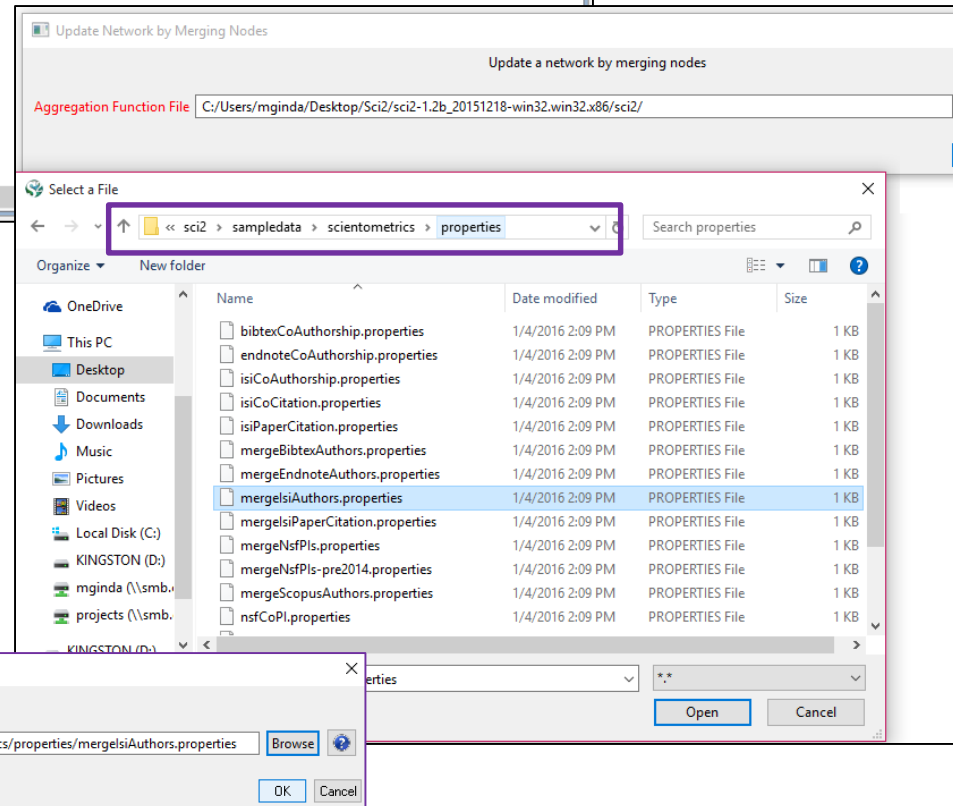
To update the co-author network with the duplicate node merges, first load the Merge list as the Standard CSV format.





Selecting both the original network and the loaded merge list CSV, navigate to **Data Preparation > Update Network by Merging Nodes**.

Browse for the aggregate function file in the Sci2 sample data directory. Load the **“mergelsiAuthors.properties”** file, and then select **OK**.



This graph claims to be undirected.

Nodes: 2191

Isolated nodes: 4

Node attributes present: label, number_of_authored_works, times_cited

Edges: 16042

No self loops were discovered.

No parallel edges were discovered.

Edge attributes:

Did not detect any nonnumeric attributes.

Numeric attributes:

	min	max	mean
number_...	1	10	1.11364
weight	1	10	1.10865

This network seems to be valued.

Average degree: 14.6435

This graph is not weakly connected.

There are 60 weakly connected components. (4 isolates)

The largest connected component consists of 1706 nodes.

Did not calculate strong connectedness because this graph was not directed.

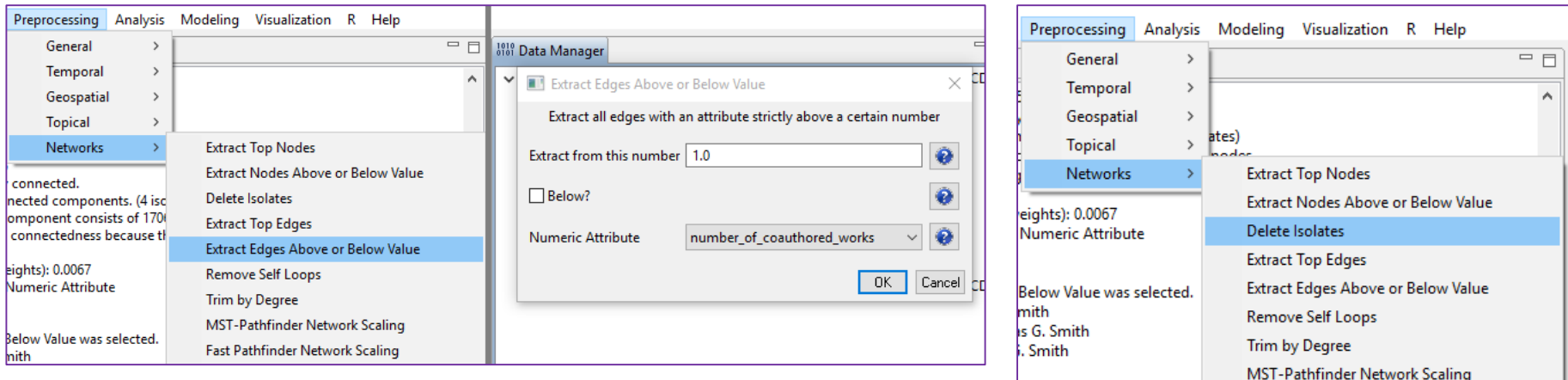
Density (disregarding weights): 0.0067

Additional Densities by Numeric Attribute

After updating the network by merging nodes, we can review the effects of our update on the network by using the NAT algorithm.

Navigate to **Analysis > Networks > Network Analysis Toolkit (NAT)**.

To view the results, select the output file in the data manager and select **View** or **View with...**

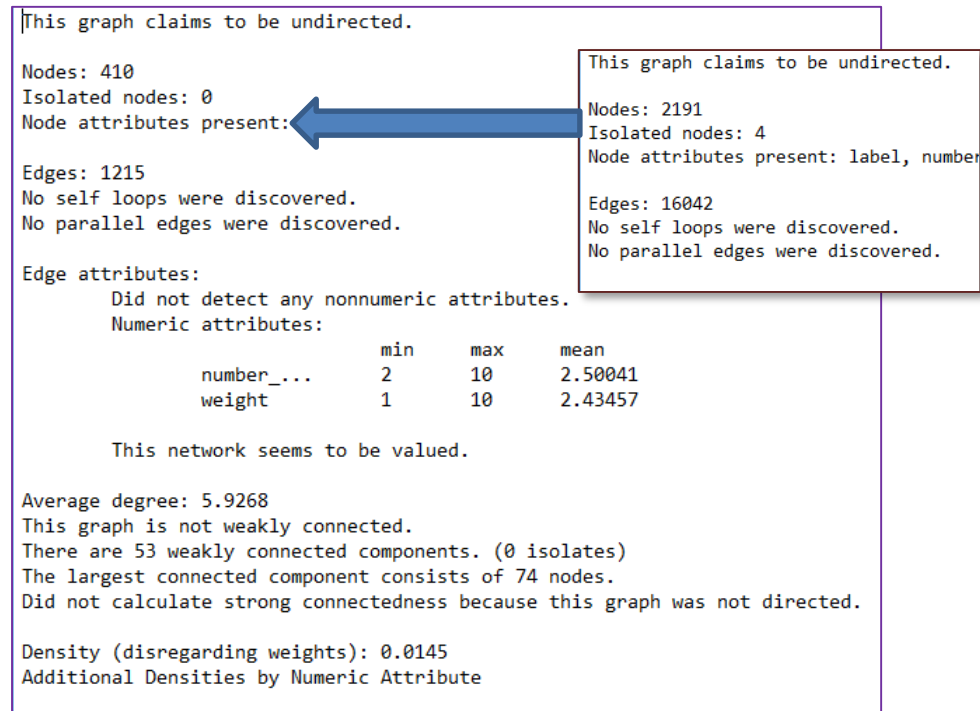


The image shows two screenshots of the Sci2 software interface. The left screenshot shows the 'Preprocessing' menu with 'Networks' selected, and the 'Extract Edges Above or Below Value' dialog box open. The dialog box has 'Extract from this number' set to 1.0 and 'Numeric Attribute' set to 'number_of_coauthored_works'. The right screenshot shows the 'Delete Isolates' option selected in the 'Networks' menu.

A majority of the 16,000+ edges are weak connections. In order to show repeated collaborations, we will extract only edges that are > 1 .

To do this, navigate to **Preprocessing > Networks > Extract Edges Above or Below Value**, and use the standard parameters and select the **numeric attribute** as the **number_of_coauthored_works**, and select **OK**.

Last run **Preprocessing > Networks > Delete Isolates** to remove any authors without a relationship in the network. A NAT report shows the changes made.



This graph claims to be undirected.

Nodes: 410
Isolated nodes: 0
Node attributes present:

Edges: 1215
No self loops were discovered.
No parallel edges were discovered.

Edge attributes:
Did not detect any nonnumeric attributes.
Numeric attributes:

	min	max	mean
number_...	2	10	2.50041
weight	1	10	2.43457

This network seems to be valued.

Average degree: 5.9268
This graph is not weakly connected.
There are 53 weakly connected components. (0 isolates)
The largest connected component consists of 74 nodes.
Did not calculate strong connectedness because this graph was not directed.

Density (disregarding weights): 0.0145
Additional Densities by Numeric Attribute

This graph claims to be undirected.

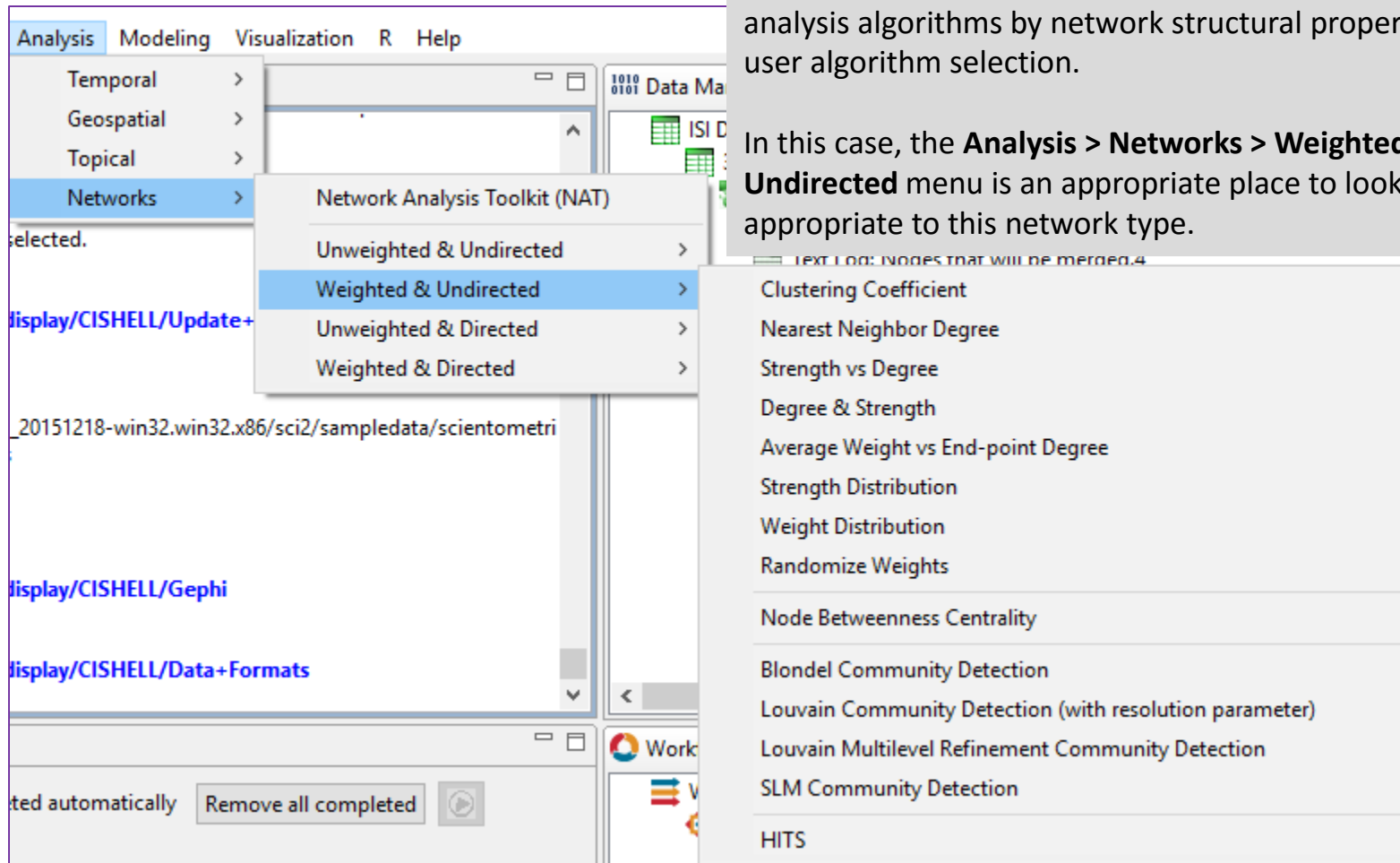
Nodes: 2191
Isolated nodes: 4
Node attributes present: label, number

Edges: 16042
No self loops were discovered.
No parallel edges were discovered.

The co-authorship network extracted from the CDC Web of Science publication citations, are an example of a Weighted & Undirected network.

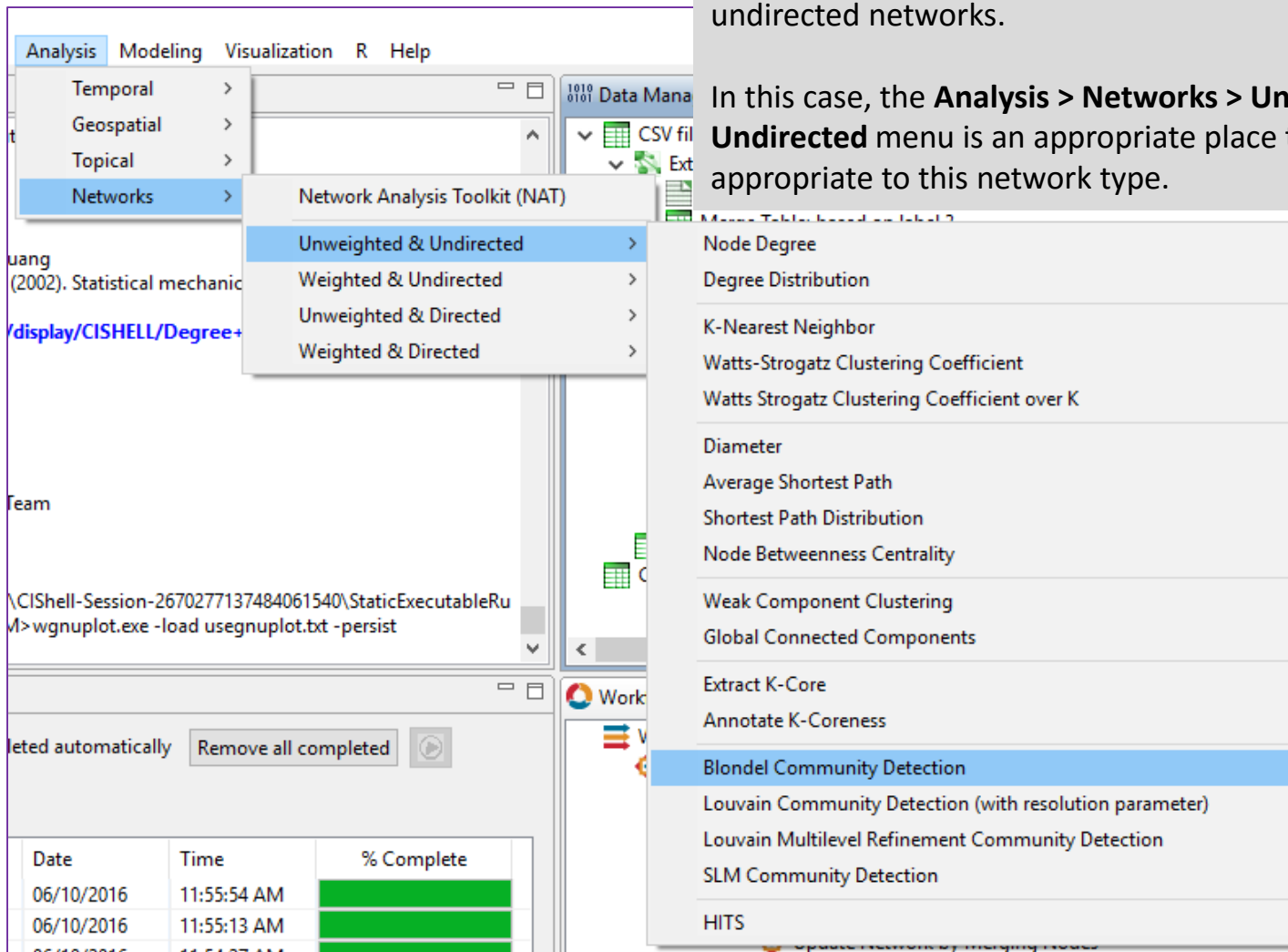
Sci2's Analysis menu has broken out appropriate network analysis algorithms by network structural properties to aid user algorithm selection.

In this case, the **Analysis > Networks > Weighted & Undirected** menu is an appropriate place to look for analysis appropriate to this network type.



Additionally, co-authorship network may be treated as an unweighted network (node and edge weights). These algorithms may still provide insights into weighted & undirected networks.

In this case, the **Analysis > Networks > Unweighted & Undirected** menu is an appropriate place to look for analysis appropriate to this network type.



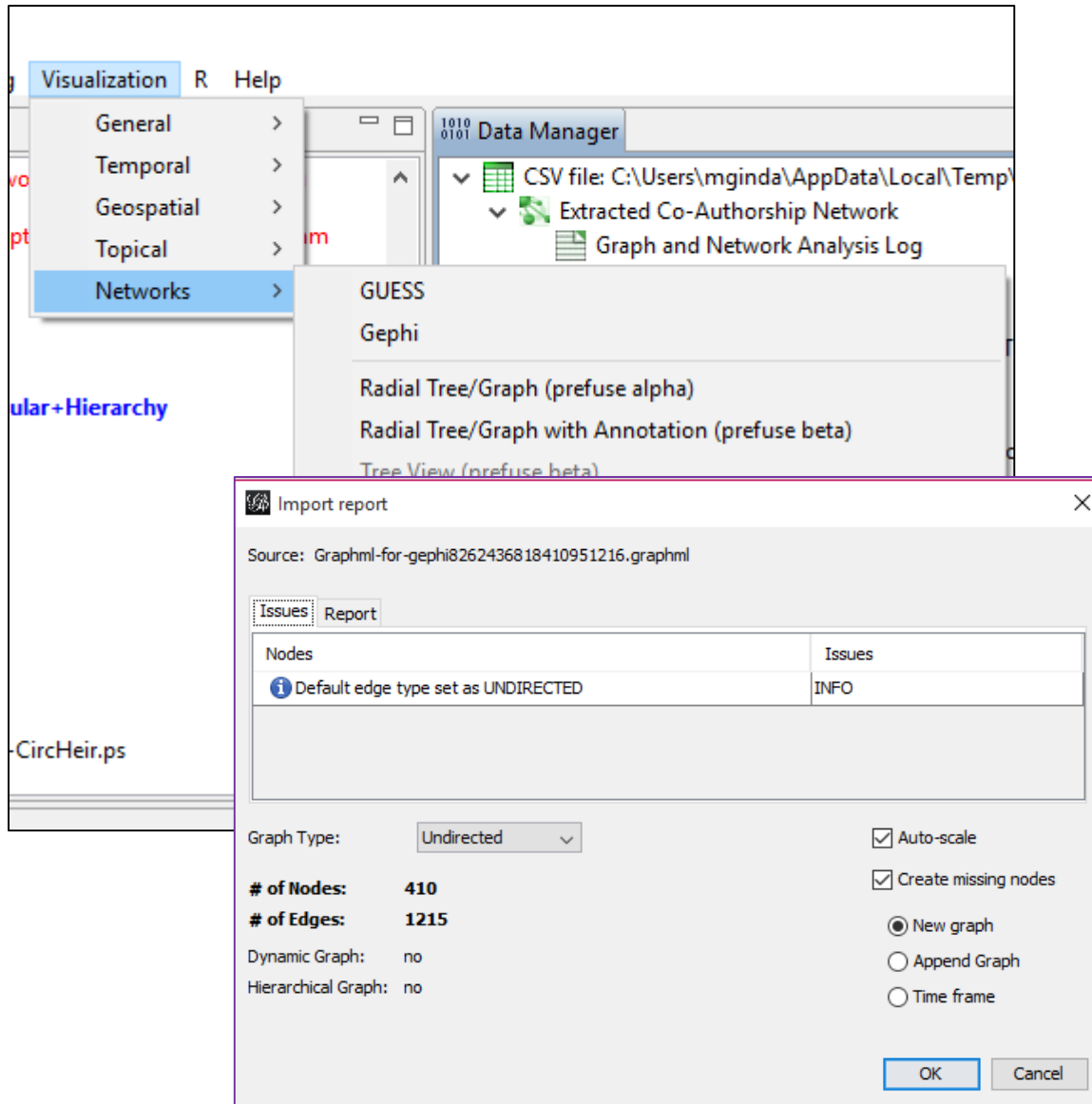
The screenshot shows the Sci2 software interface with the following menu structure:

- Analysis
 - Temporal
 - Geospatial
 - Topical
 - Networks**
 - Network Analysis Toolkit (NAT)
 - Unweighted & Undirected**
 - Node Degree
 - Degree Distribution
 - K-Nearest Neighbor
 - Watts-Strogatz Clustering Coefficient
 - Watts Strogatz Clustering Coefficient over K
 - Diameter
 - Average Shortest Path
 - Shortest Path Distribution
 - Node Betweenness Centrality
 - Weak Component Clustering
 - Global Connected Components
 - Extract K-Core
 - Annotate K-Coreness
 - Blondel Community Detection**
 - Louvain Community Detection (with resolution parameter)
 - Louvain Multilevel Refinement Community Detection
 - SLM Community Detection
 - HITS
 - Weighted & Undirected
 - Unweighted & Directed
 - Weighted & Directed

In the background, a terminal window shows the command: `wgnuplot.exe -load usegnuplot.txt -persist`. A taskbar at the bottom shows a table with the following data:

Date	Time	% Complete
06/10/2016	11:55:54 AM	100%
06/10/2016	11:55:13 AM	100%
06/10/2016	11:54:27 AM	100%

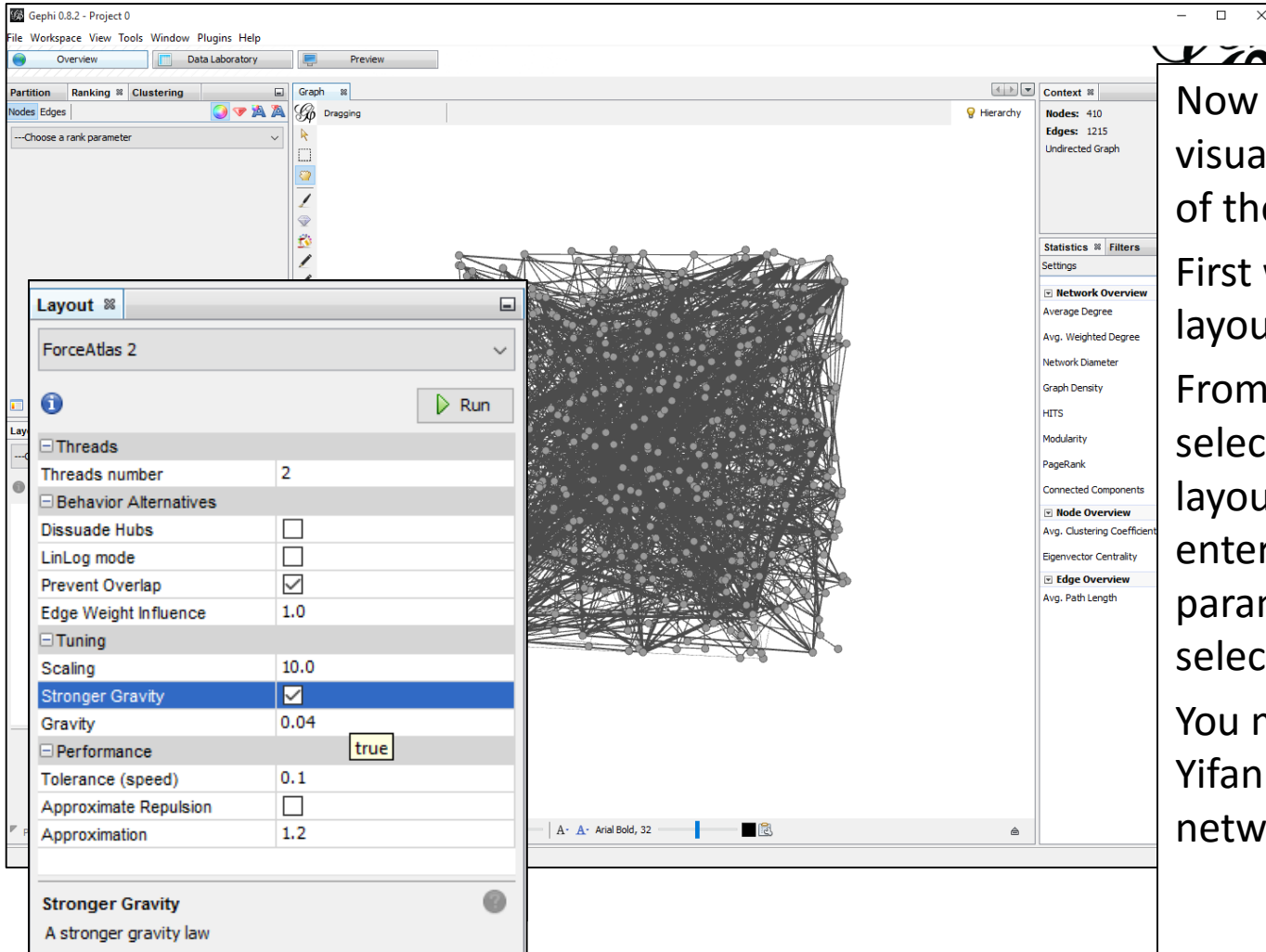
Questions?



Next we will visualize the network in Gephi.

Navigate to **Visualization > Networks > Gephi**.

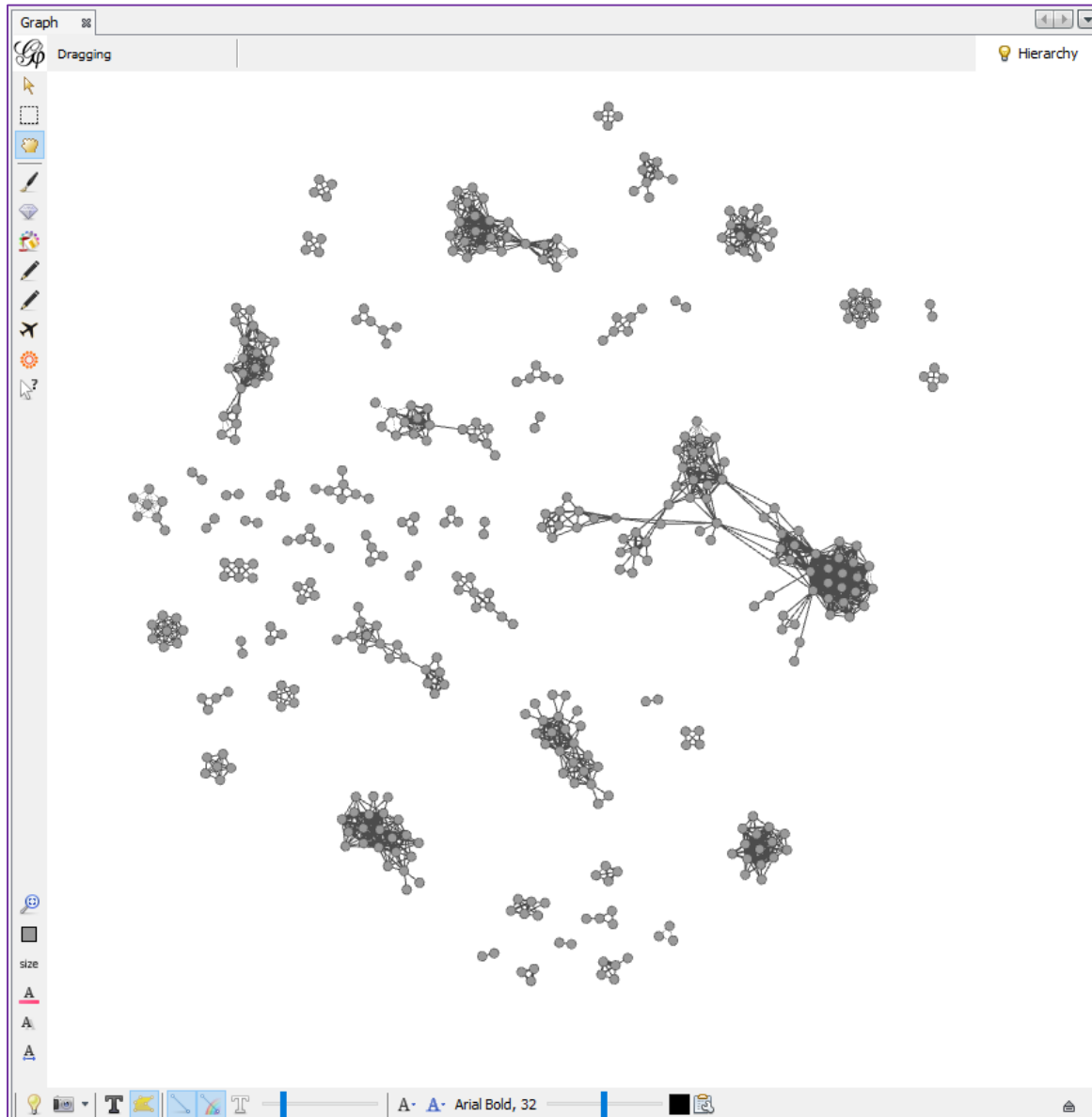
The algorithm is a bridge that passes the network data to Gephi. The program will automatically start. The tool produces an Import Report. It lets you select the network type, gives load errors, etc.



Now we can start an visualization and analysis of the network.

First we will adjust the layout of the network. From the layout pane, select the “ForceAtlas2” layout algorithm and enter the following parameters, and then select “Run”.


You may also select YifanHu’s Multilevel force network layout.

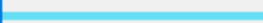


Partition Ranking Clustering

Nodes Edges

times_cited

Color:  Default Invert Recent

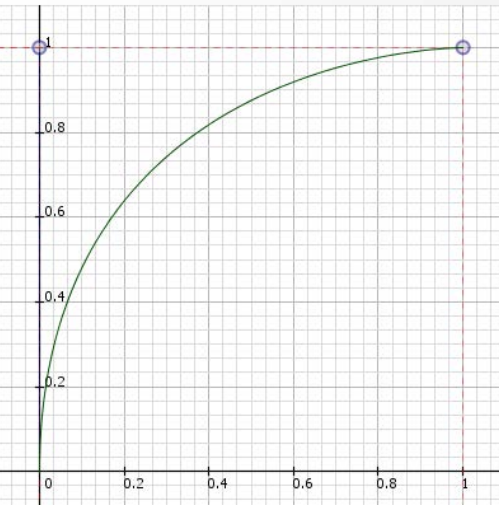
Range:  13 2198

Spline...

Interpolate

Spline Editor

Drag control points in the display to change the shape of the spline

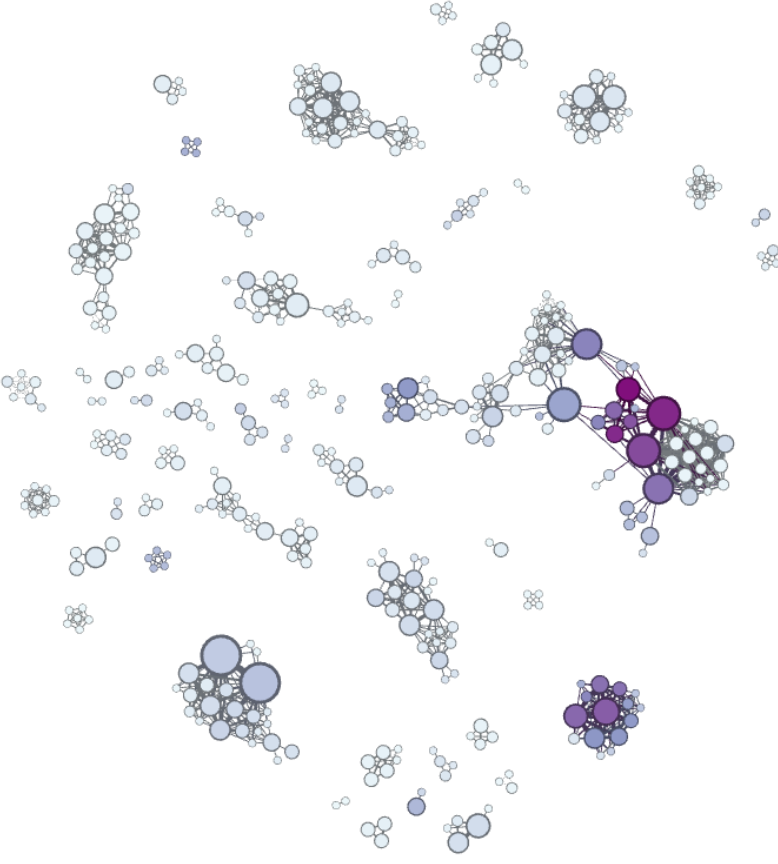


Templates

Close

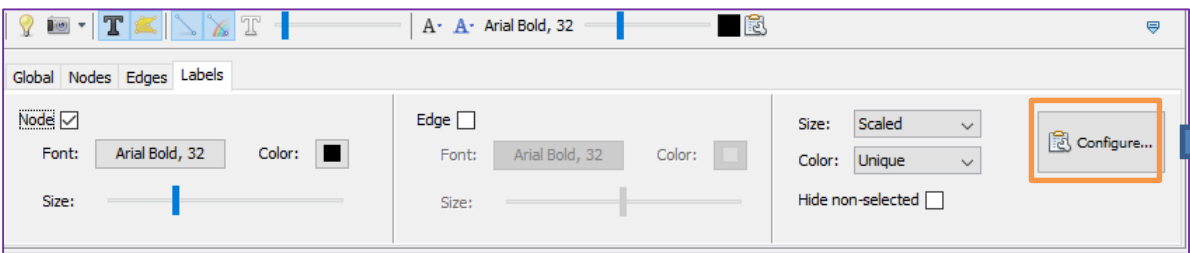
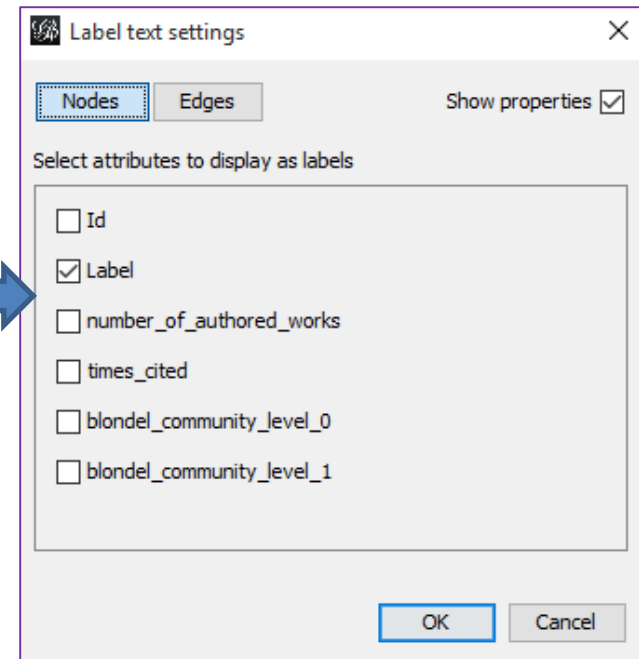
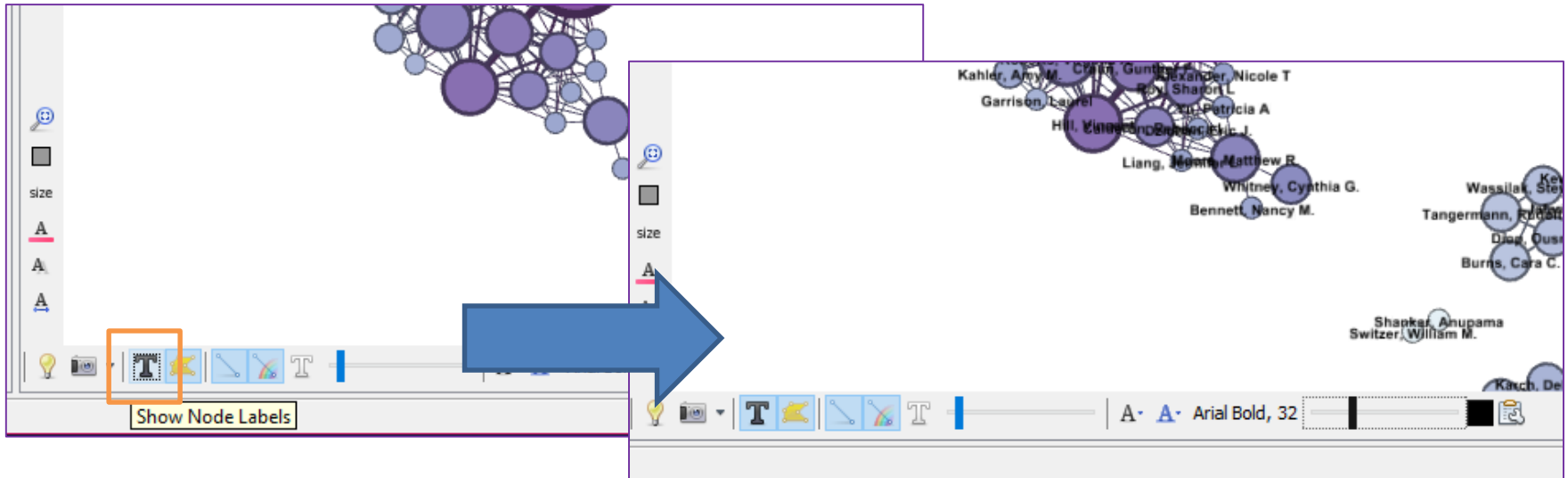
Dragging

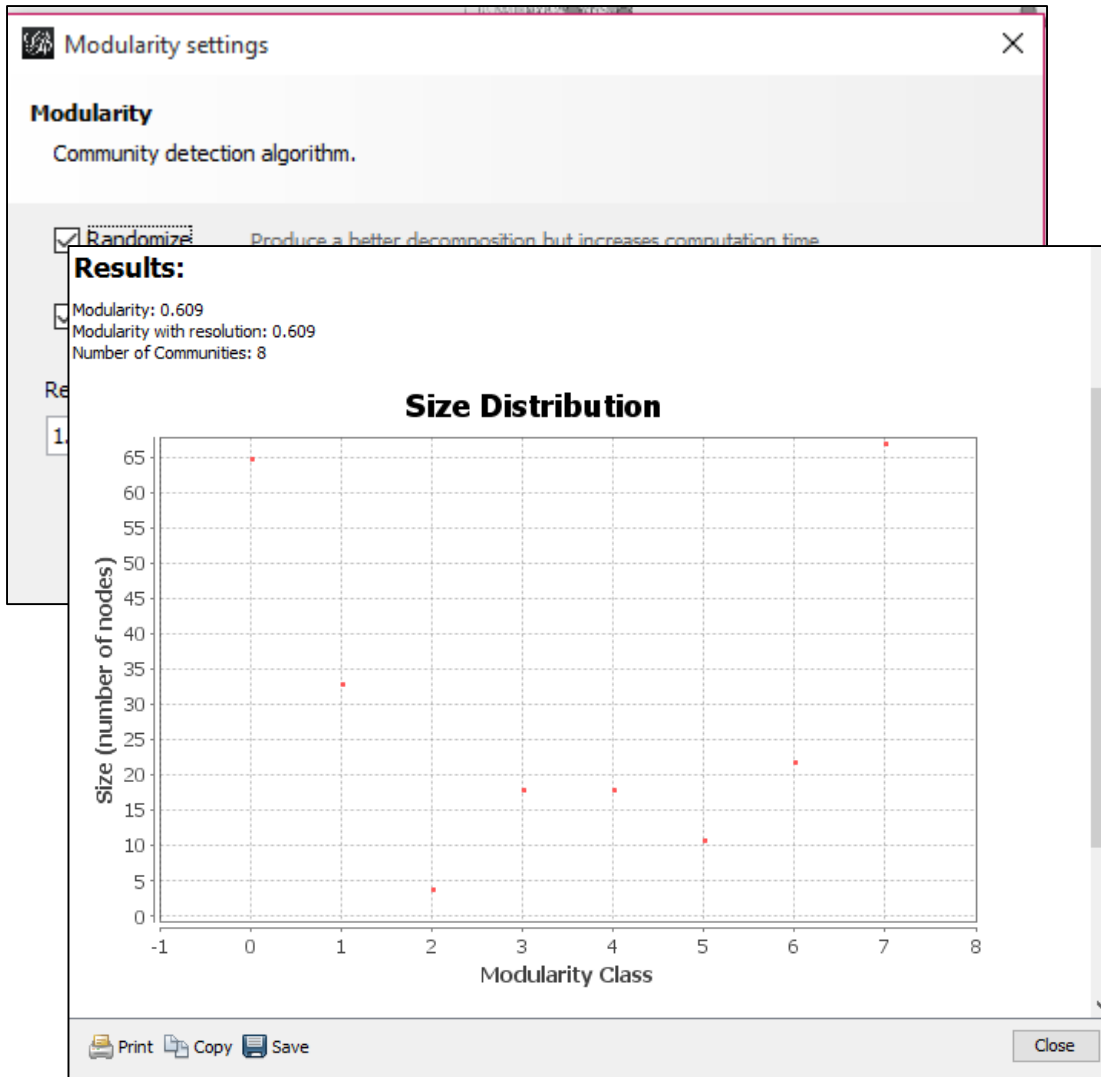
Hierarchy



size

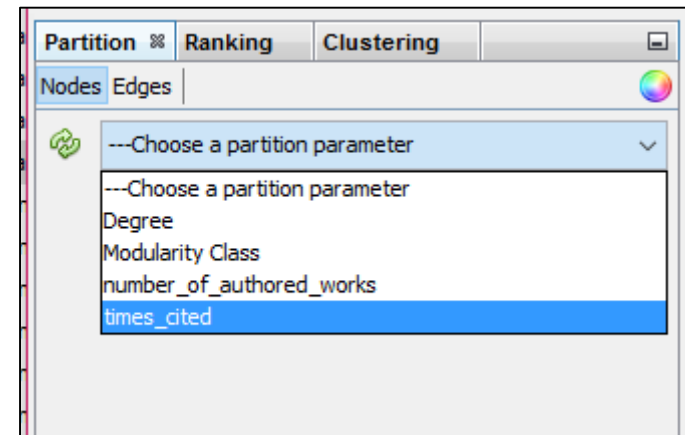
Arial Bold, 32





The modularity statistical algorithm calculates how the connectedness of a network, and the Blondel Communities that exist in the network. The communities are added as a partition to the nodes.

The modularity categories may be applied to the network from the Partitions window.



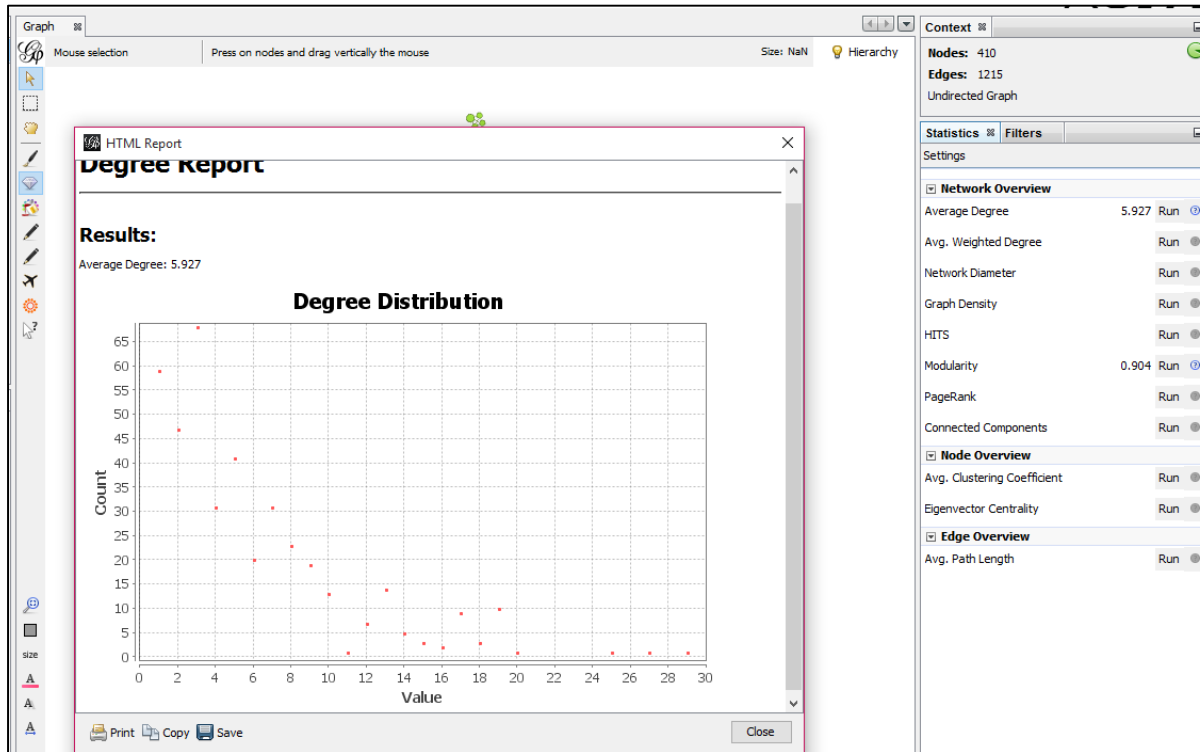
The screenshot displays the Gephi software interface. The main window shows a network graph with nodes colored by community. The left sidebar is divided into several panels:

- Partition**: Shows the 'Modularity Class' and a list of communities with their respective percentages.

Community	Percentage
community_44	5.61%
community_0	4.88%
community_16	4.39%
community_15	4.15%
community_1	3.9%
community_23	3.66%
community_19	3.41%
community_25	2.93%
community_42	2.68%
community_17	2.44%
community_41	2.2%
community_3	1.95%
community_10	1.95%
- Layout**: Shows the 'ForceAtlas 2' layout algorithm selected. It includes a 'Run' button and various configuration options:

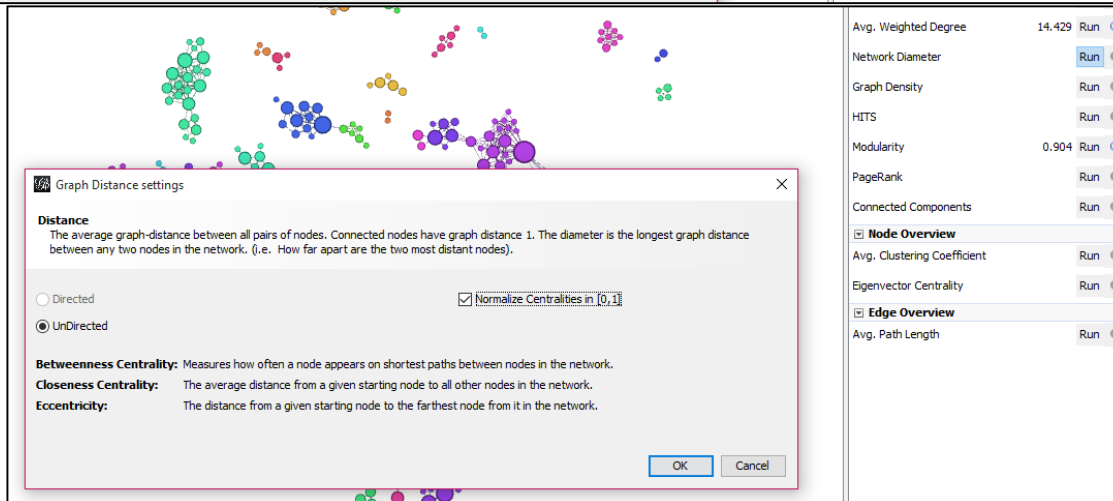
Category	Option	Value
Threads	Threads number	2
	Behavior Alternatives	
Dissuade Hubs	Dissuade Hubs	<input type="checkbox"/>
	LinLog mode	<input type="checkbox"/>
	Prevent Overlap	<input checked="" type="checkbox"/>
	Edge Weight Influence	1.0
Tuning	Scaling	10.0
	Stronger Gravity	<input checked="" type="checkbox"/>
	Gravity	0.03
Performance	Tolerance (speed)	0.1
	Approximate Repulsion	<input type="checkbox"/>
	Approximation	1.2

The central graph area shows a complex network of nodes and edges, with nodes colored according to the community detection results. The bottom toolbar contains various icons for navigation and editing.



Gephi provides a variety of node and edge statistics to help understand the relationships, clustering, paths, centrality, and communities within a network.

Try implementing the Average Degree, and Network Diameter statistics, which we will next visualize.



Context

Nodes: 410

Edges: 1215

Undirected Graph

Statistics **Filters**

Reset A>

Library

- Attributes
- Equal
- Inter Edges
- Intra Edges
- Non-null
- Partition
- Partition Count
- Range
- Betweenness Centrality *Double (Node)*
- Closeness Centrality *Double (Node)*
- Clustering Coefficient *Double (Node)*
- Degree *Integer (Node)***
- Eccentricity *Double (Node)*

Queries

Range (Degree)

Queries

Range (Degree)

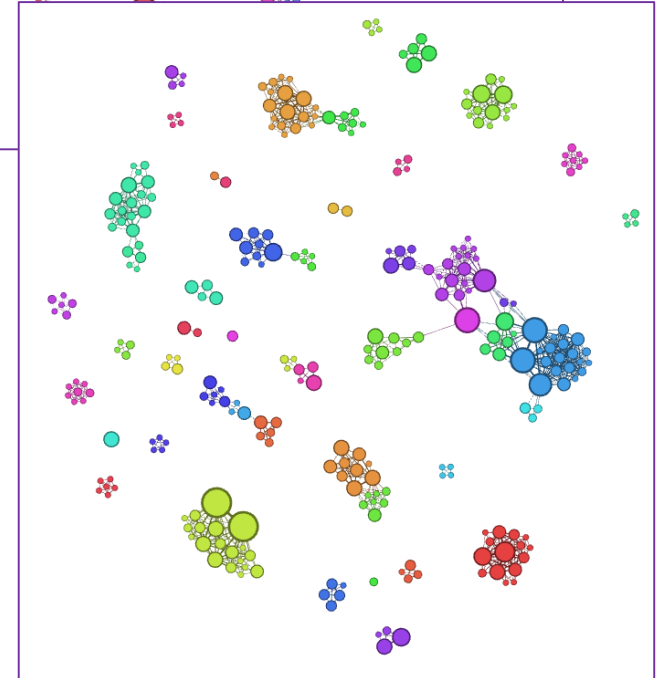
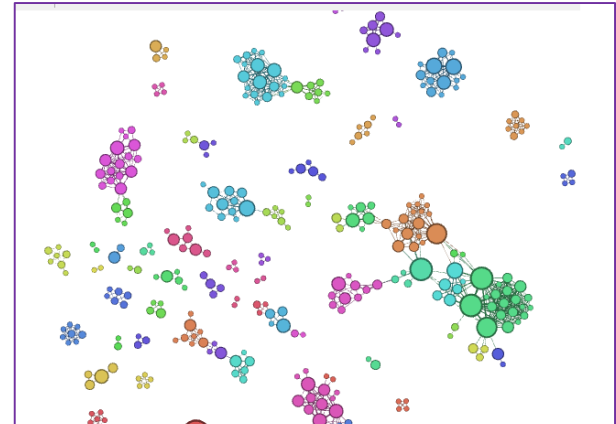
Range

1

Range (Degree) Settings

3 29

Workspace 0



Gephi 0.8.2 - Project 0

File Workspace View Tools Window Plugins Help

Overview Data Laboratory Preview

Preview Settings

Presets: Default

Nodes

Border Width: 1.0
 Border Color: custom [0,0,0]
 opacity: 100.0

Node Labels

Show Labels:
 Font: Arial 12 Plain
 Proportional size:
 Color: custom [0,0,0]
 Shorten label:
 Max characters: 30
 Outline size: 0.0
 Outline color: custom [255,255,255]
 Outline opacity: 80.0
 Box:
 Box color: parent
 Box opacity: 100.0

Edges

Show Edges:
 Thickness: 1.0
 Rescale weight:
 Color: mixed
 Opacity: 100.0
 Curved:
 Radius: 0.0

Edge Arrows

Size: 3.0

Edge Labels

Show Labels:
 Font: Arial 10 Plain
 Color: original
 Shorten label:
 Max characters: 30
 Outline size: 0.0
 Outline color: custom [255,255,255]
 Outline opacity: 80.0

Preview ratio: 100%

Export: SVG/PDF/PNG

Refresh

Background Reset zoom - +

Background Reset zoom - +

Workspace 0

The image shows the 'Preview Settings' window on the left and the 'Preview - Font' dialog box in the center. The 'Preview Settings' window has several sections: 'Nodes' (Border Width: 0.25, Border Color: custom [0,0,0], opacity: 100.0), 'Node Labels' (Show Labels: checked, Font: Arial, Proportional size: checked, Color: custom [0,0,0], Shorten label: unchecked, Max characters: 30, Outline size: 0.0, Outline color: custom [255,255,255], Outline opacity: 80.0, Box: unchecked, Box color: parent, Box opacity: 100.0), 'Edges' (Show Edges: checked, Thickness: 1.0, Rescale weight: unchecked, Color: mixed, Opacity: 100.0, Curved: checked, Radius: 0.0), 'Edge Arrows' (Size: 3.0), and 'Edge Labels'.

The 'Preview - Font' dialog box has three columns: 'Font', 'Font Style', and 'Size'. The 'Font' column lists various fonts, with 'Arial' selected. The 'Font Style' column lists 'Plain', 'Bold', 'Italic', and 'Bold Italic', with 'Plain' selected. The 'Size' column lists sizes from 3 to 36, with '3' selected. There is a 'Preview' section at the bottom of the dialog and 'OK' and 'Cancel' buttons.

Preview ratio: 100%

Export: SVG/PDF/PNG

Refresh

Background Reset zoom - +

Preview ratio: 100%

Export: SVG/PDF/PNG

Refresh

Background Reset zoom - +

Edges

- Show Edges
- Thickness 15.0
- Rescale weight
- Color custom [92,148,83] ...
- Opacity 100.0
- Curved
- Radius 0.0

Edge Arrows

- Size 3.0

Edge Labels

- Show Labels
- Font Arial 10 Plain ...
- Color original ...
- Shorten label
- Max characters 30
- Outline size 0.0
- Outline color custom [255,255,255] ...
- Outline opacity 80.0

Preview Settings - Color

Edge Color
Configures the color of the edges. Edges can either a color on their own (original) or uses incident nodes color.

- Original
- Mixed
- Source
- Target
- Custom

Choose a Color

Hue: 112

Sat: 44

Bri: 69

Red: 109

Green: 176

Blue: 99

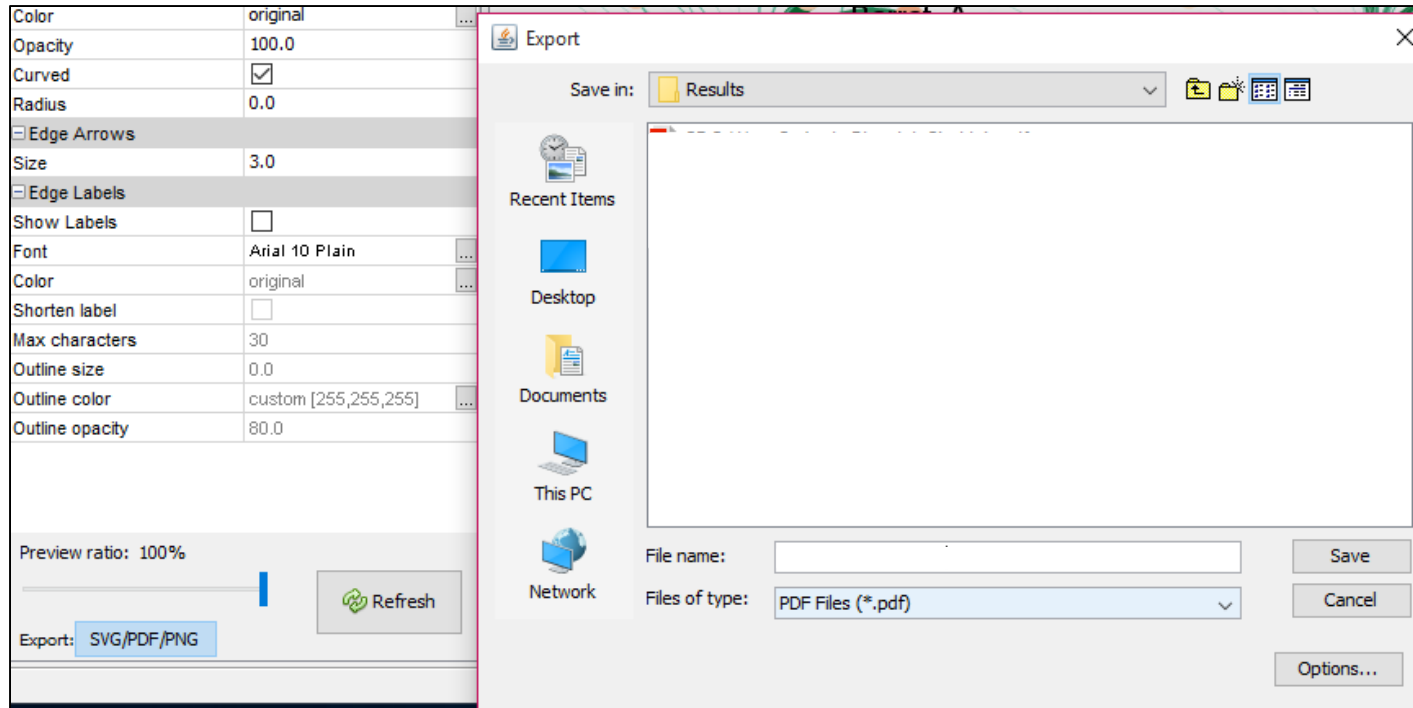
Hex: 6DB063

Preview ratio: 100%

Export: SVG/PDF/PNG

Refresh

Background Reset zoom - +



Questions?