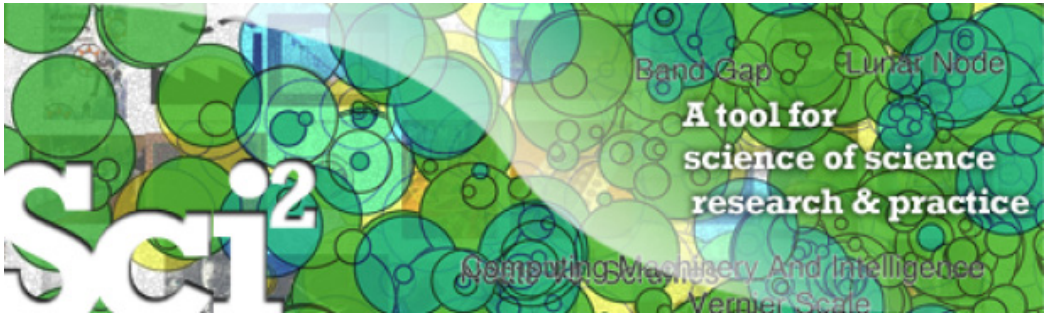
An Introduction to Open Source Tools for Data Analysis and Information Visualization

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



OPEN Refine

Gephi
makes graphs handy

Tutorial Overview

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

Software, Datasets, Plugins, and Documentation

- 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

Tutorial Overview

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

Visualization Framework

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

Visualization Framework

LEVELS

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

TYPES

Statistical Analysis
page 44

Knowledge Cartography
page 135

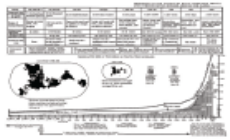
Productivity of Russian life sciences research teams
page 105

Number of scientists versus population and R&D costs versus GNP.
page 103

WHEN: Temporal Analysis
page 48

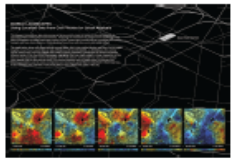
Visualizing decision-making processes
page 95

Key events in the development of the video tape recorder
page 85

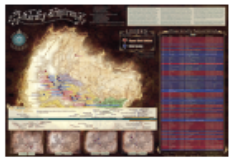


Increased travel and communication speeds
page 83

WHERE: Geospatial Analysis
page 52



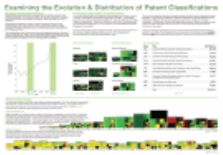
Cell phone usage in Milan, Italy
page 109



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

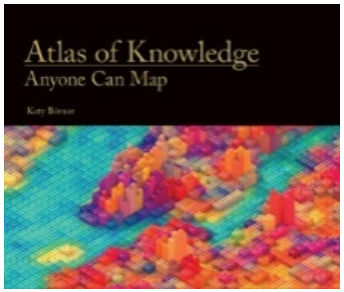
Product space showing co-export patterns of countries
page 93

WITH WHOM: Network Analysis
page 60

World Finance Corporation network
page 87

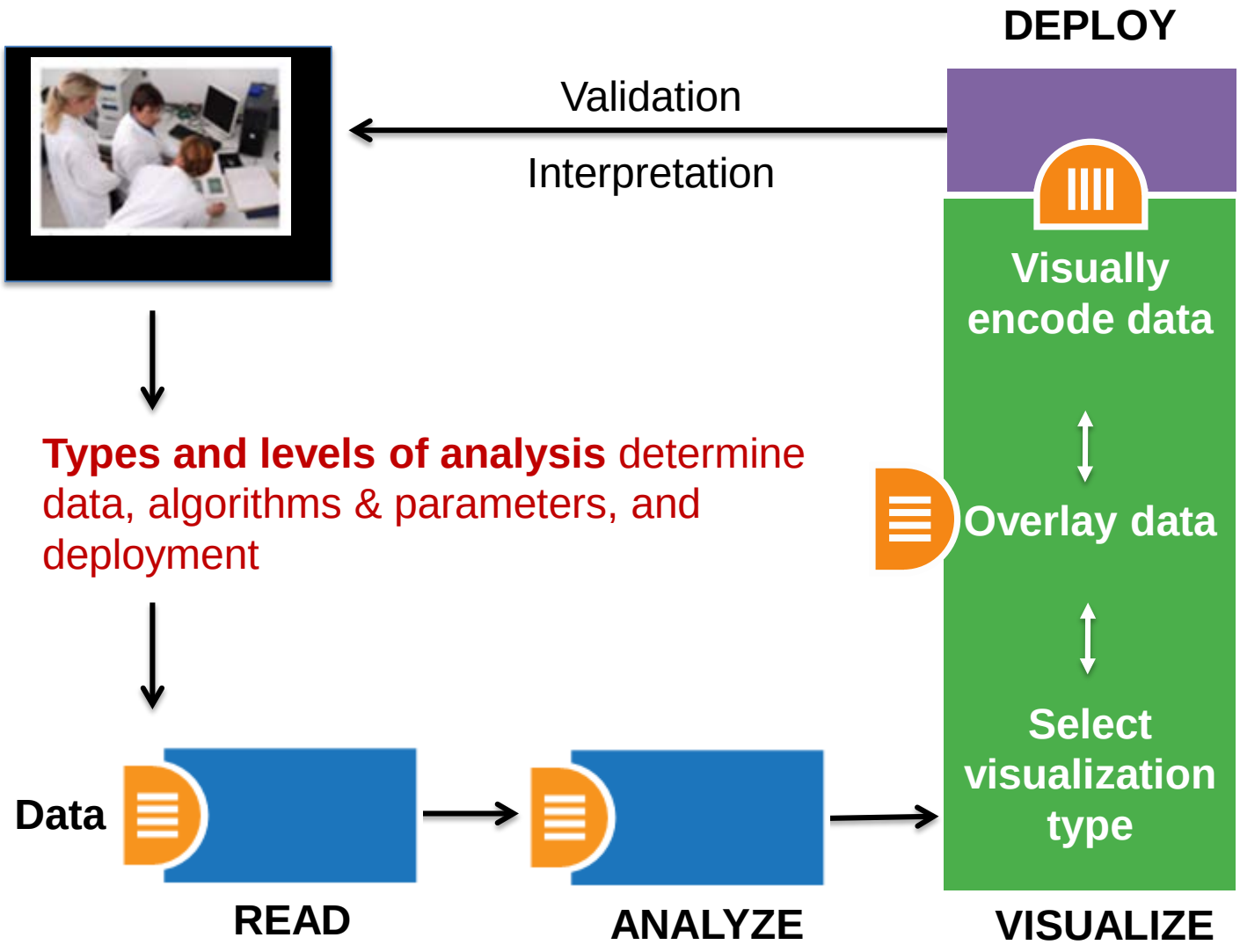
Electronic and new media art networks
page 133

World-wide scholarly collaboration networks
page 157

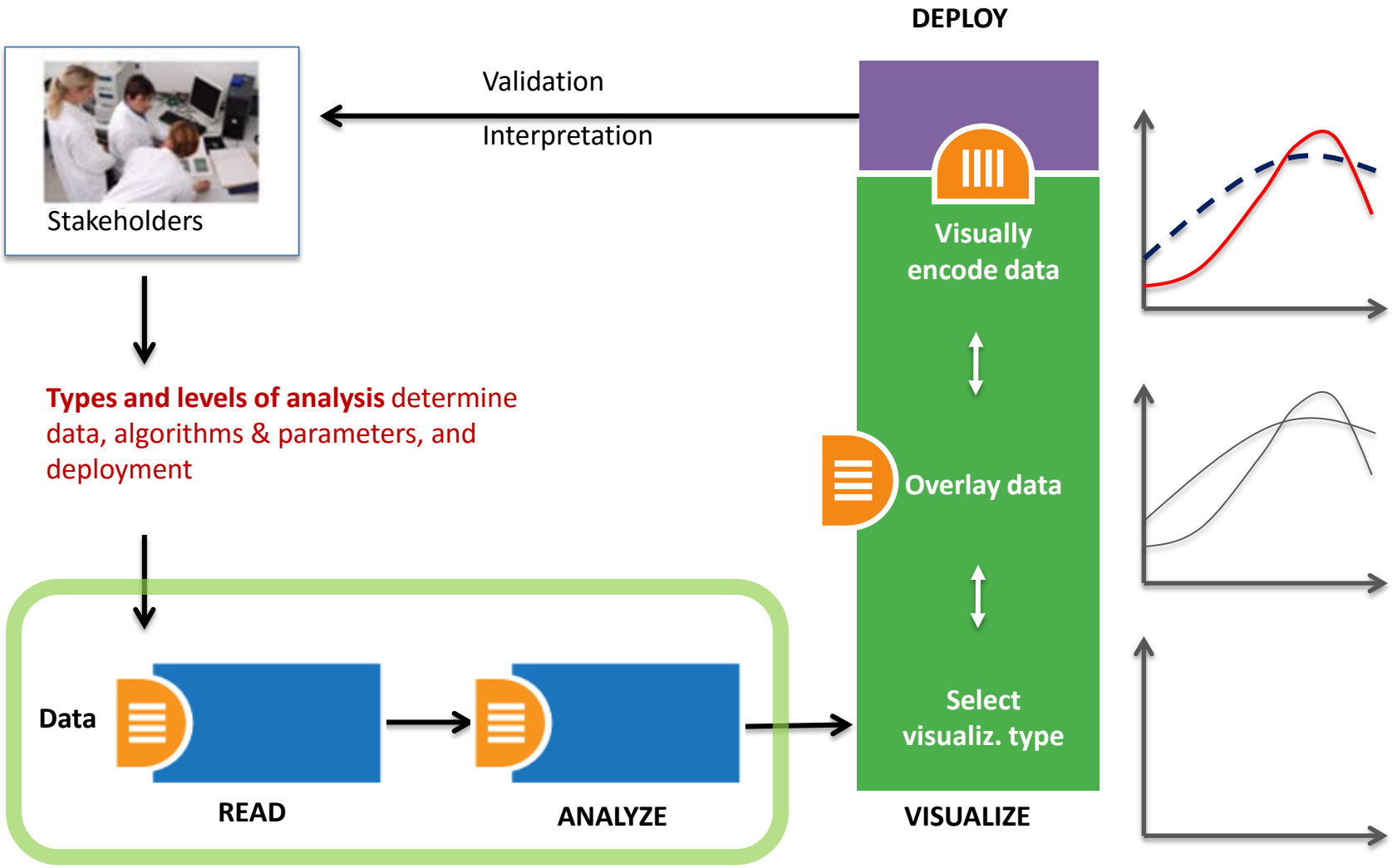


See page 5

Workflow Design and Implementation

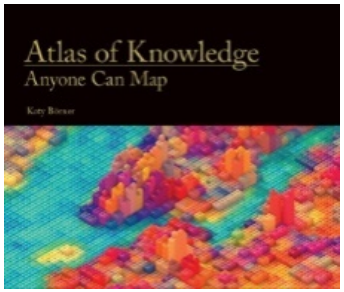


Tying It All Together: Visualization Framework and Workflow Design



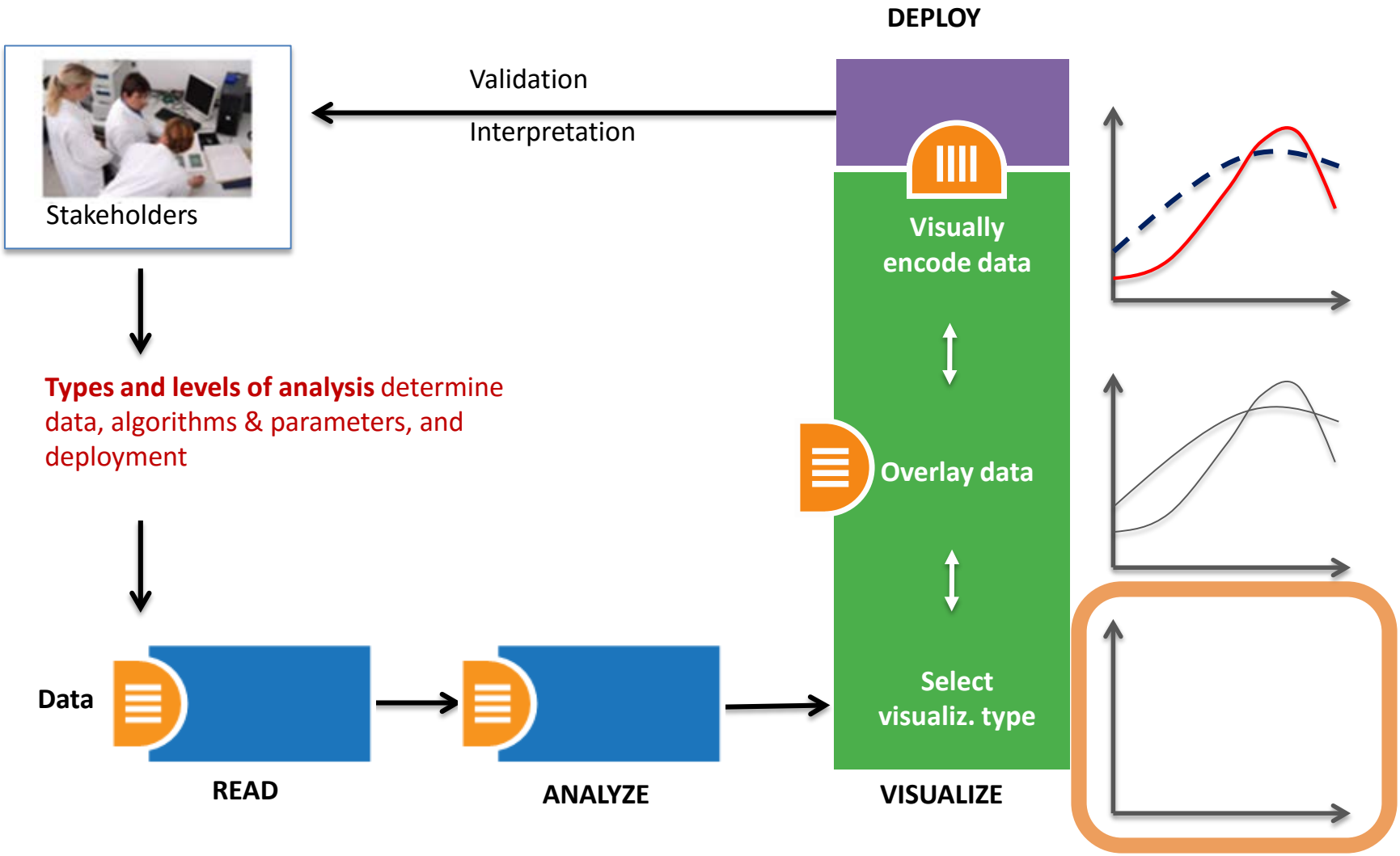
Tying It All Together: Visualization Framework and Workflow Design

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
• categorize/cluster • order/rank/sort • distributions (also outliers, gaps) • comparisons • trends (process and time) • geospatial • compositions (also of text) • correlations/relationships	• nominal • ordinal • interval • ratio	• table • chart • graph • map • network layout	• geometric symbols point line area surface volume • linguistic symbols text numerals punctuation marks • pictorial symbols images icons statistical glyphs	• spatial position • retinal form color optics motion	• overview • zoom • search and locate • filter • details-on-demand • history • extract • link and brush • projection • distortion



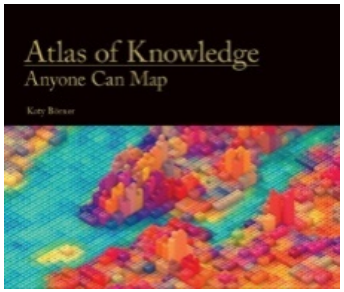
See page 24

Tying It All Together: Visualization Framework and Workflow Design



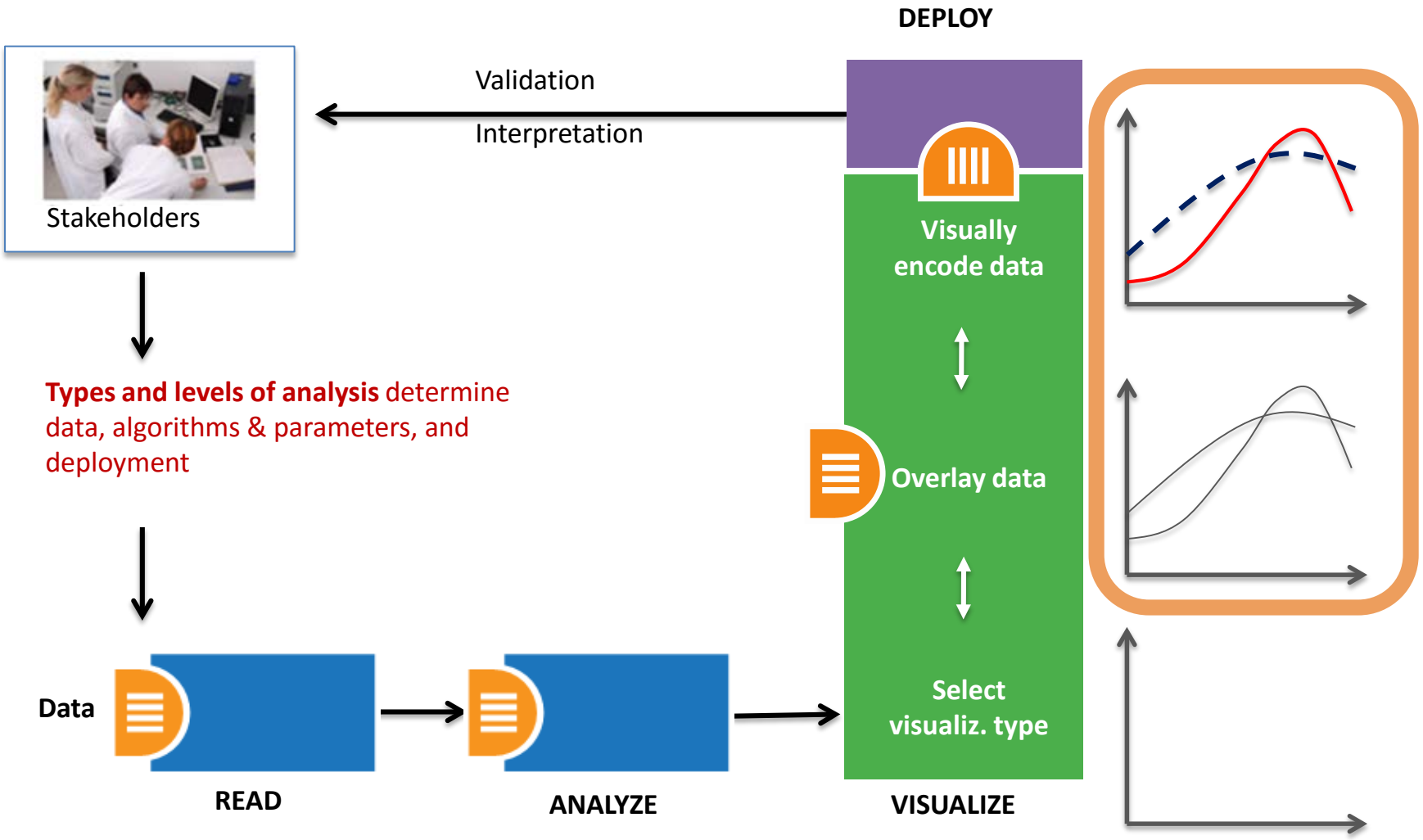
Tying It All Together: Visualization Framework and Workflow Design

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
• categorize/cluster • order/rank/sort • distributions (also outliers, gaps) • comparisons • trends (process and time) • geospatial • compositions (also of text) • correlations/relationships	• nominal • ordinal • interval • ratio	• table • chart • graph • map • network layout	• geometric symbols point line area surface volume • linguistic symbols text numerals punctuation marks • pictorial symbols images icons statistical glyphs	• spatial position • retinal form color optics motion	• overview • zoom • search and locate • filter • details-on-demand • history • extract • link and brush • projection • distortion



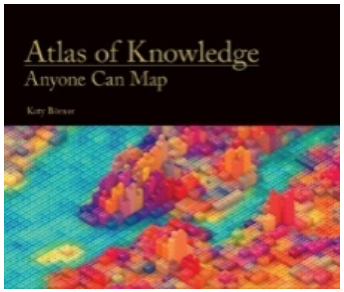
See page 24

Tying It All Together: Visualization Framework and Workflow Design



Tying It All Together: Visualization Framework and Workflow Design

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
• categorize/cluster • order/rank/sort • distributions (also outliers, gaps) • comparisons • trends (process and time) • geospatial • compositions (also of text) • correlations/relationships	• nominal • ordinal • interval • ratio	• table • chart • graph • map • network layout	• geometric symbols point line area surface volume • linguistic symbols text numerals punctuation marks • pictorial symbols images icons statistical glyphs	• spatial position • retinal form color optics motion	• overview • zoom • search and locate • filter • details-on-demand • history • extract • link and brush • projection • distortion



See page 24

Questions?

Tutorial Overview

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

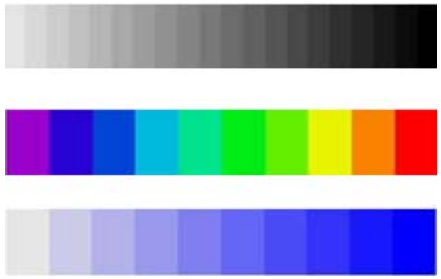
Position: x, y; possibly z

Form:

- Size
- Shape
- Orientation/Rotation

Color:

- Value (Lightness)
- Hue (Tint)
- Saturation (Intensity)



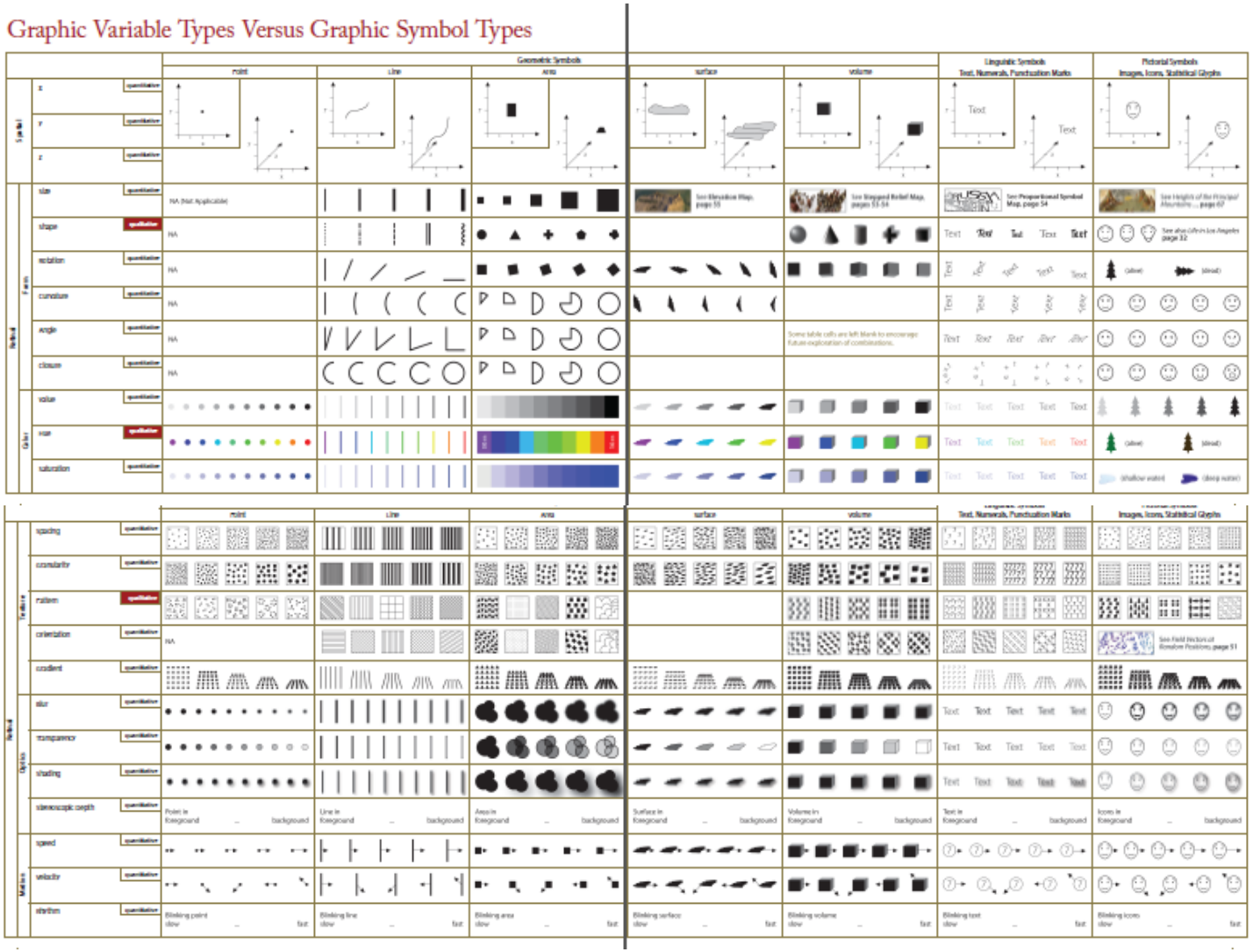
Texture:

- Pattern, Rotation, Coarseness, Size, Density Gradient

Optics:

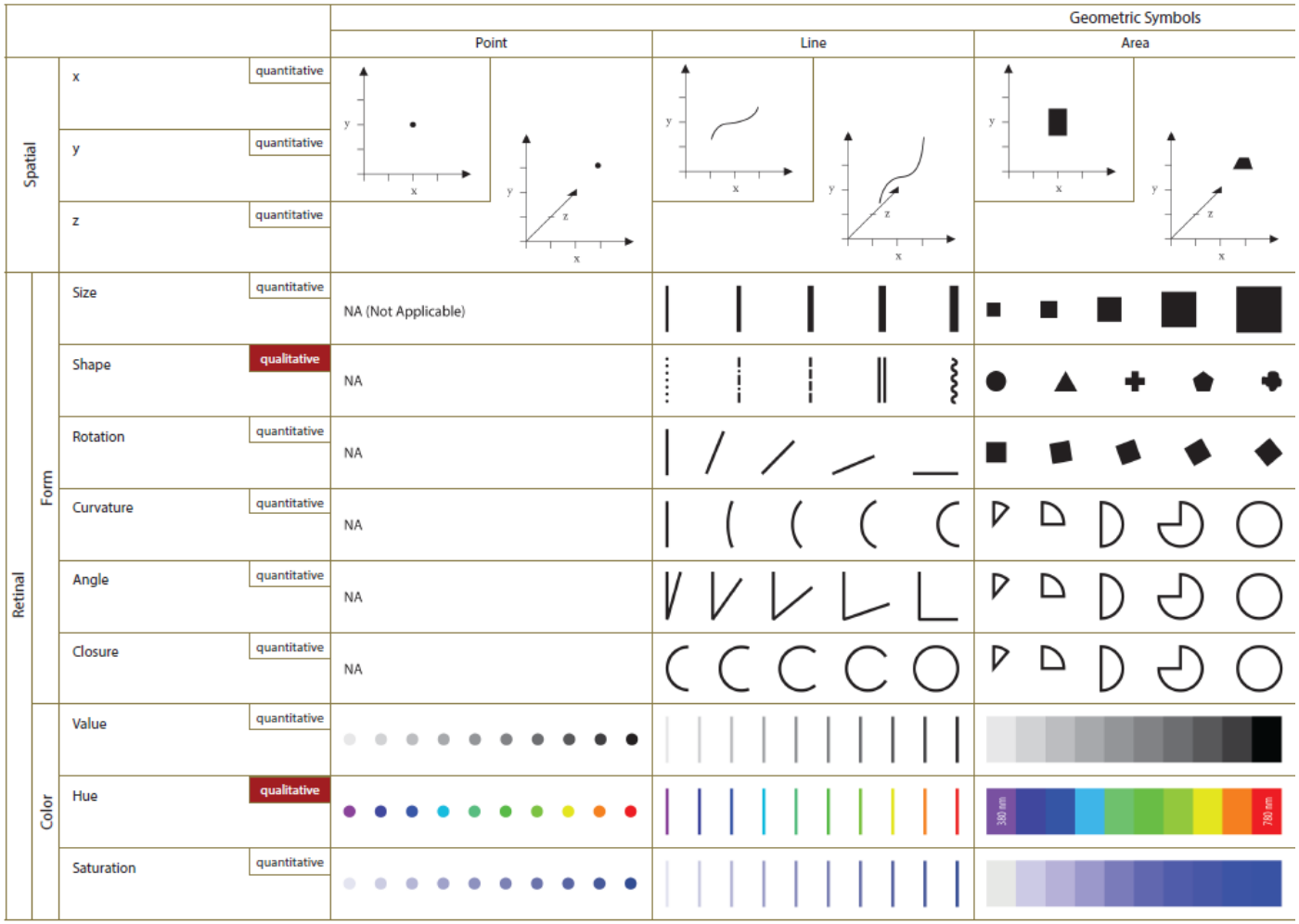
- Crispness, Transparency, Shading

Workflow Design: Graphic Variable Types



Workflow Design: Graphic Variable Types

Graphic Variable Types Versus Graphic Symbol Types



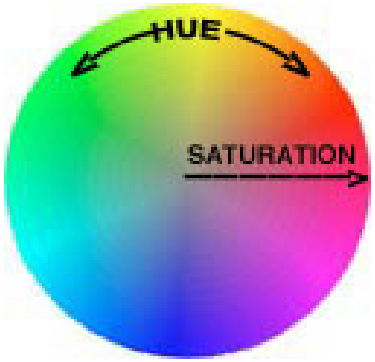
Workflow Design: Color

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**

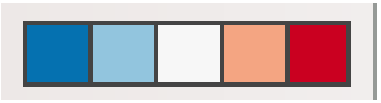
  (Tint) related to the wavelength of the stimulus. Categorical and should never be used to encode magnitude. Need to select sequence carefully, e.g., yellow through orange to red.

- **Saturation**

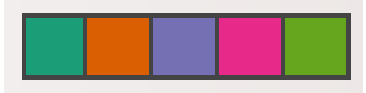
  (Intensity) is related to how much white content is in the stimulus. Monochromatic hues are very highly saturated. Higher saturated (purer) colors appear in the foreground while low saturation (dull) colors fade into background.

Workflow Design: Color Schemes

- **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**

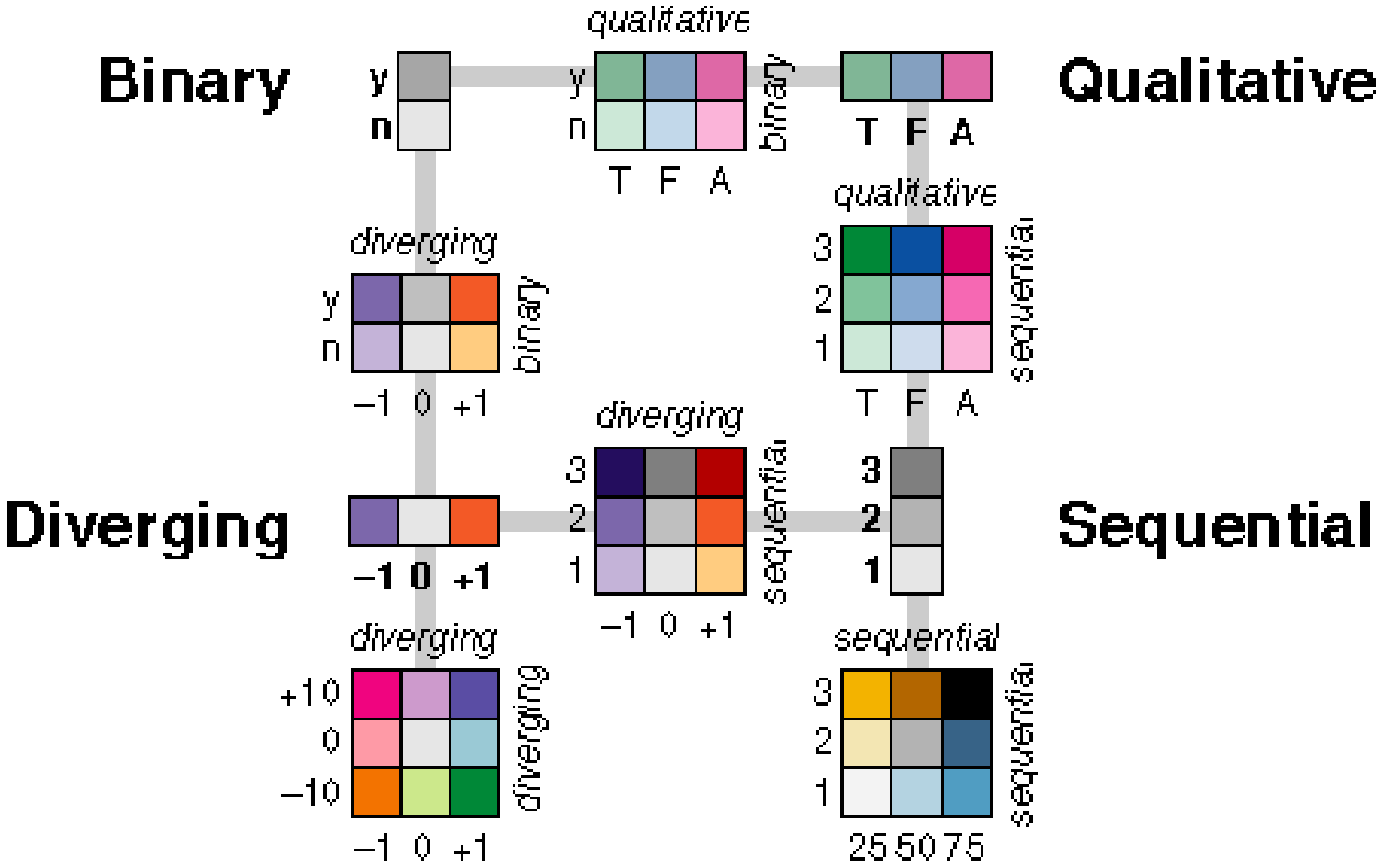


  (Bi-polar) put equal emphasis on mid-range critical values and extremes at both ends of the data range. The critical class or break in the middle of the legend is emphasized with light colors and low and high extremes are emphasized with dark colors that have contrasting hues
- **Qualitative schemes**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

Workflow Design: Color Schemes



See [Color Use Guidelines for Mapping and Visualization](#) by Cynthia Brewer

Questions?

Tutorial Overview

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

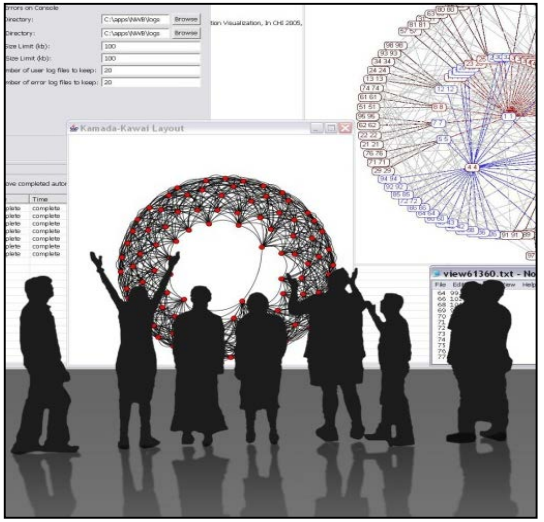
Macroscopes

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

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

Microscopes

Telescopes



Macroscopes

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

Tutorial Overview

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

Sci2 Tool – Supported Data Formats

Input:

Network Formats

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

Scientometric Formats

- ISI (*.isi)
- Bibtex (*.bib)
- Endnote Export Format (*.enw)
- Scopus csv (*.scopus)
- NSF csv (*.nsf)

Other Formats

- Pajek Matrix (*.mat)
- TreeML (*.xml)
- Edgelist (*.edge)
- CSV (*.csv)

Output:

Network File Formats

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

Image Formats

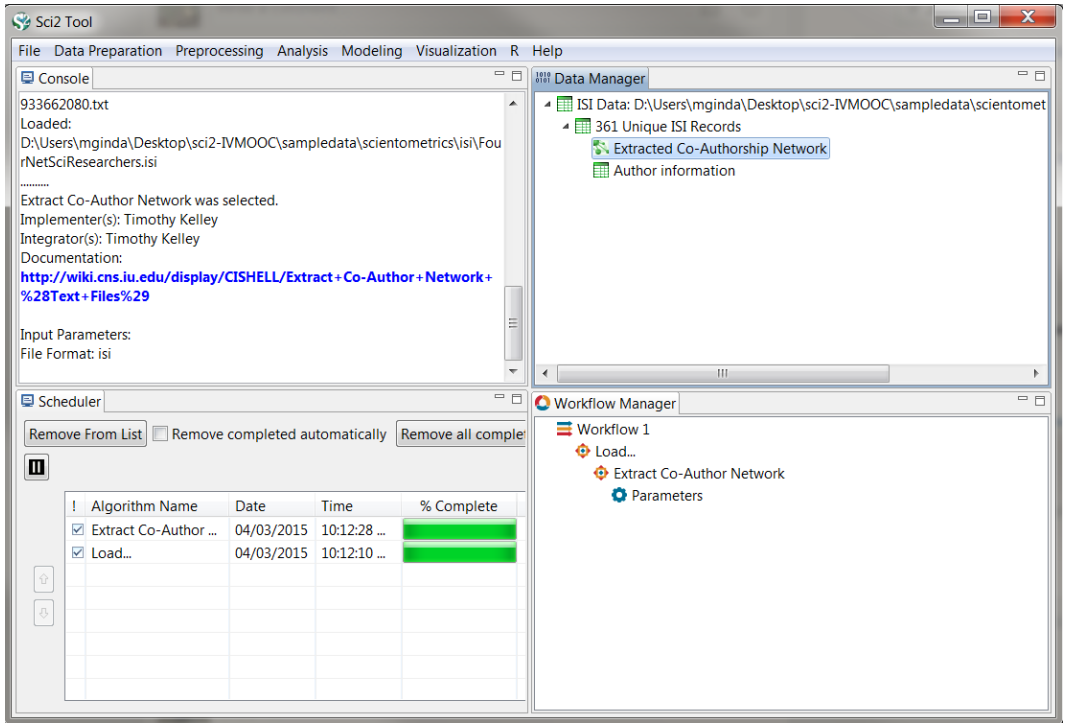
- JPEG (*.jpg)
- PDF (*.pdf)
- PostScript (*.ps)

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

Sci2 Tool – Interface Components

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

Troubleshooting

Problem: Sci2 will not run on my computer…

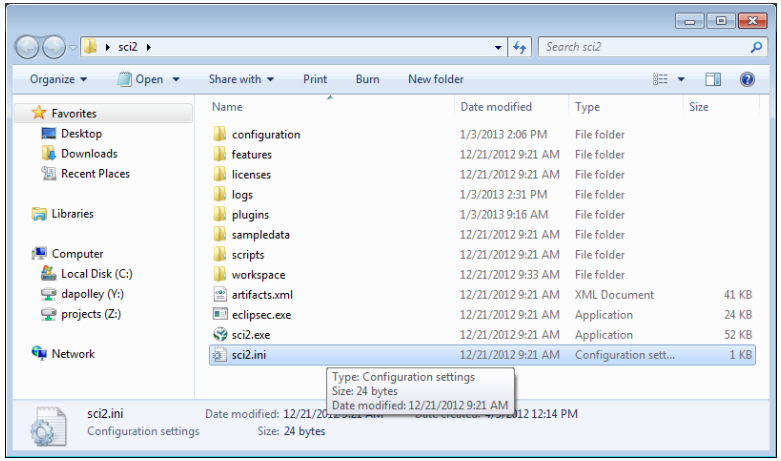
Solution: Sci2 runs on Java 1.6 (32 bit) or newer. If you are having trouble installing Sci2, you may want to install the latest 32 bit version of Java. You can run 32 bit and 64 bit versions of Java simultaneously.

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
```



Troubleshooting

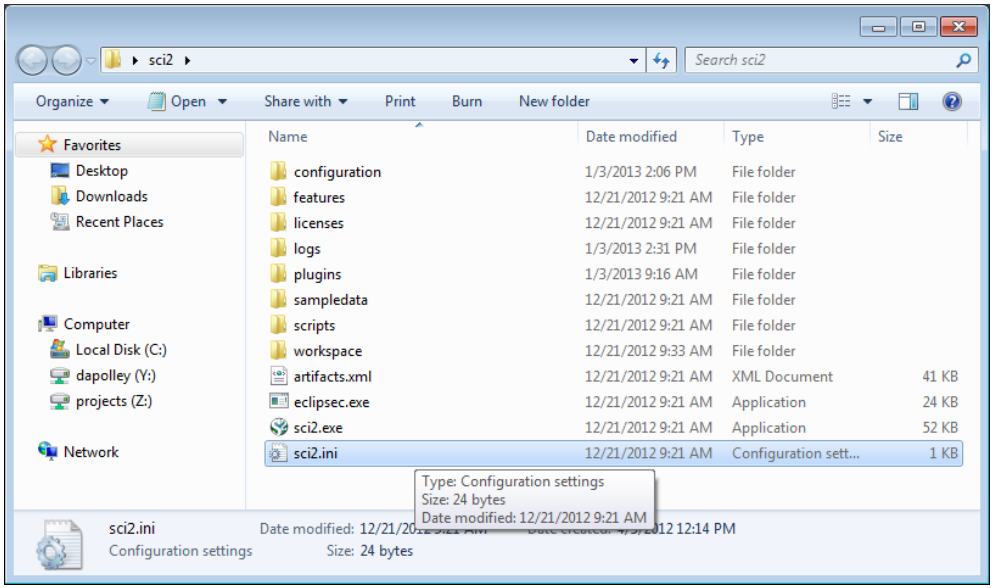
Problem: Sci2 is taking a long time to complete a process…

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.

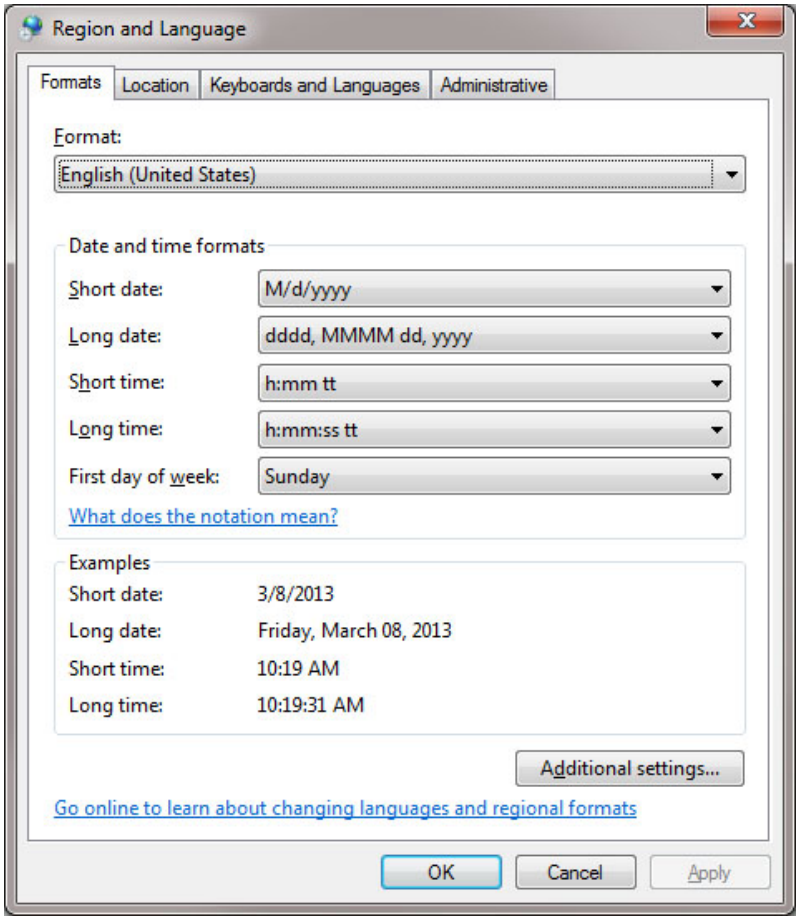


Troubleshooting

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

Tutorial Overview

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

Open Refine

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.

Tutorial Overview

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

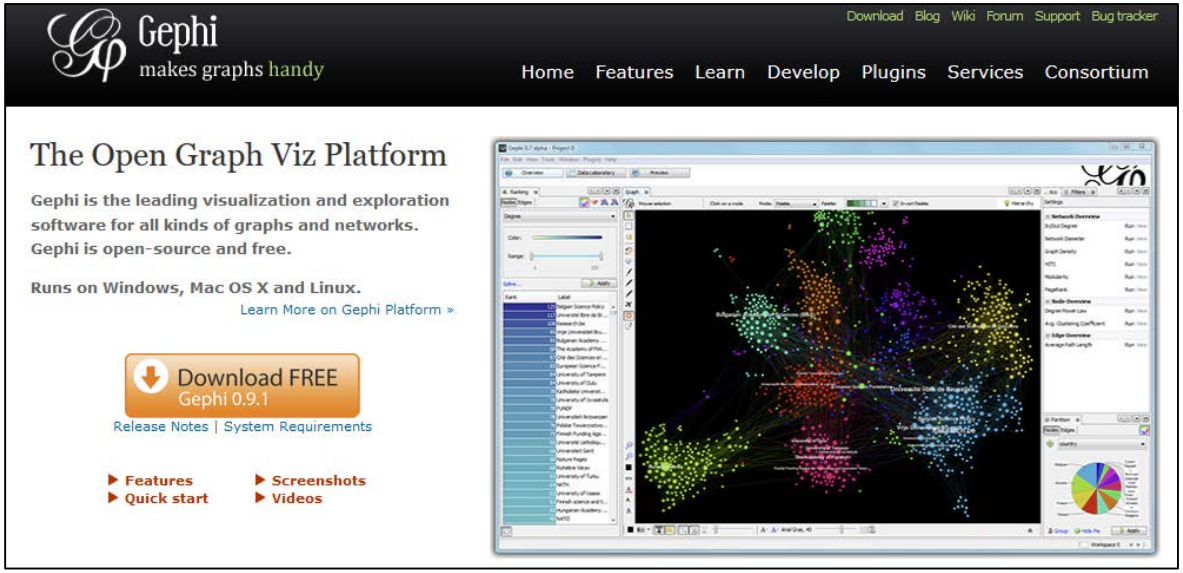
9:45 Geospatial Analysis
- Overview of geospatial analysis and mapping
- Geospatial Analysis: Geocoding with OpenRefine
- Geospatial Analysis: Proportional Symbol Map using CDC

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



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 resouces.

Questions?

Tutorial Overview

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

Introduction to Geospatial Analysis

- **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

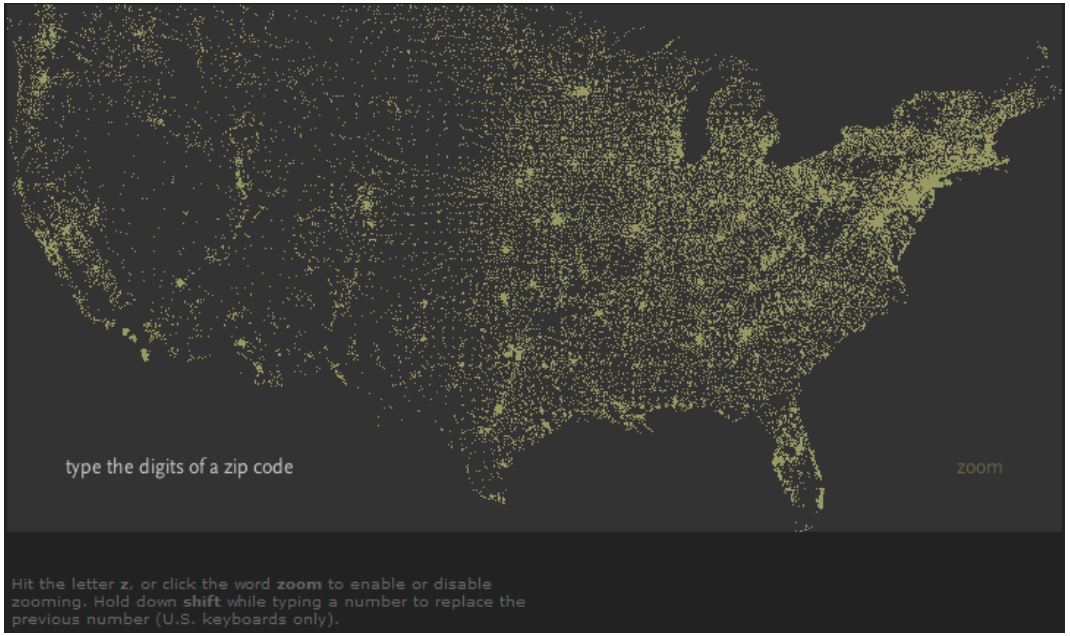
- ***Geocode***: Location of a record (e.g., address, census tract, postal code, geographic coordinates).
- ***Geographic coordinates***: Locations on the surface of the Earth expressed in degrees of latitude and longitude.
- ***Geodesic*****:** The shortest distance between two points on the surface 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

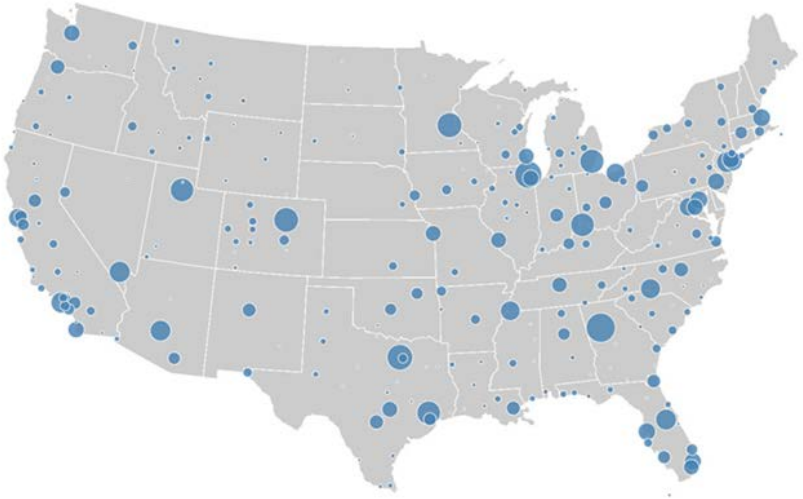


Map Types

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.

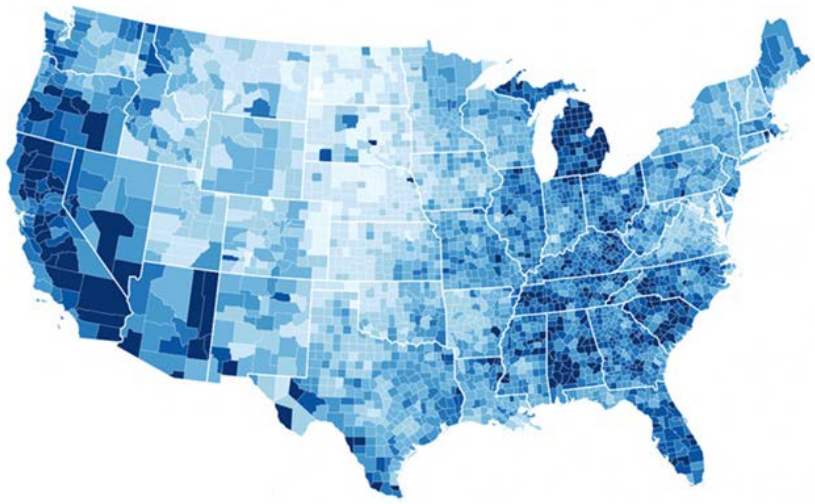
Do NOT use for densities, ratios, or scales, which should be rendered as choropleth map.



Choropleth map

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

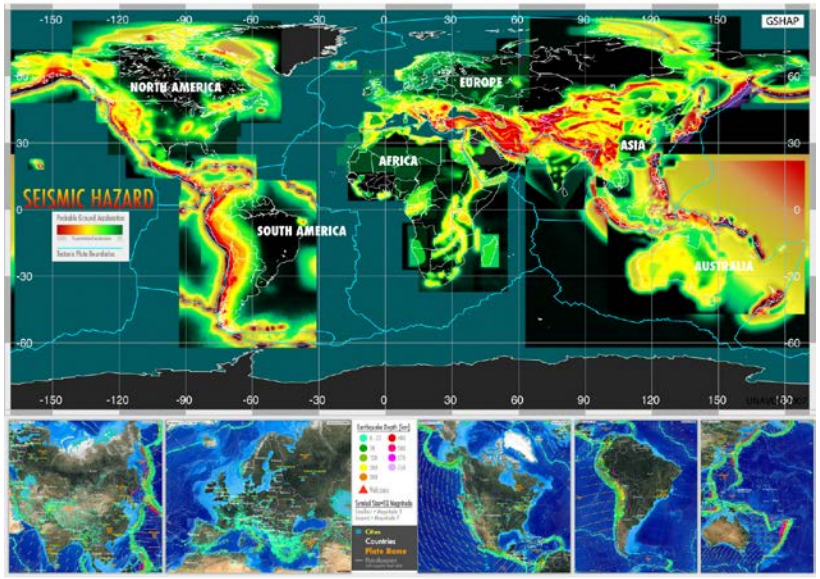
Each artificial collection unit is called a ***chronogram*** and has a distinctive color or shading.



Map Types

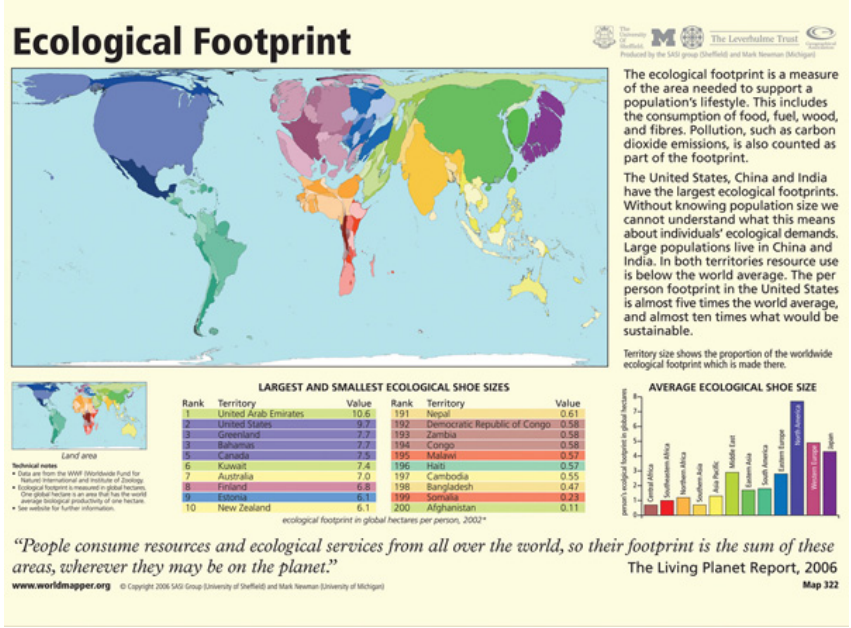
Heat (isopleth) maps

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

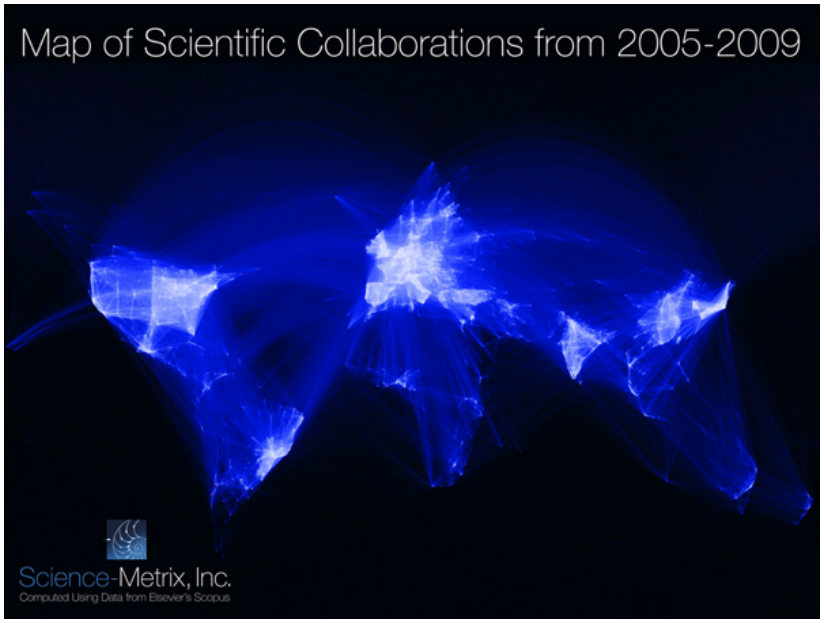
are not drawn to scale. Instead, they distort geographical areas in proportion to data values. Familiarity with regions is necessary. Mostly used for world, continental, and country maps.



Map Types

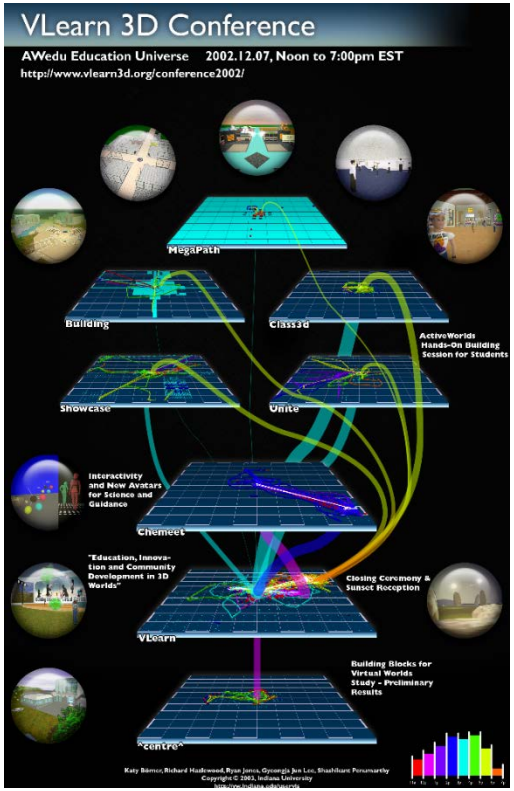
Flow maps

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?

Tutorial Overview

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

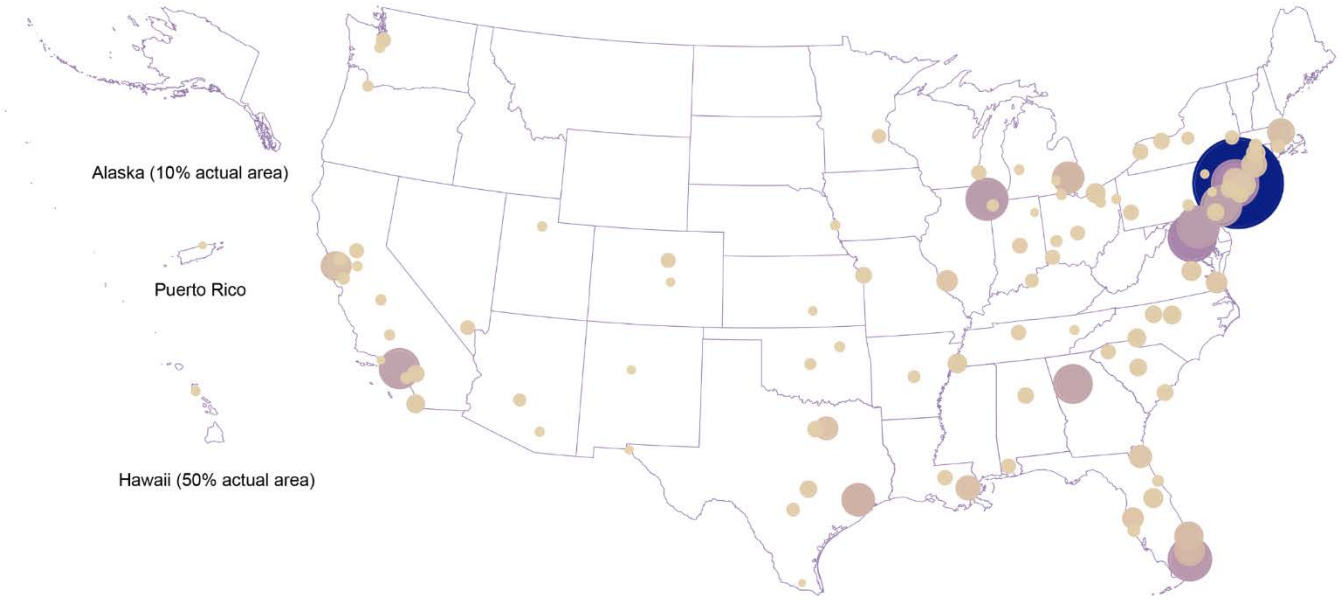
4:00 Wrap-up
4:30 Adjourn

Proportional Symbol Map: CDC HIV/AIDS Diagnosis Data

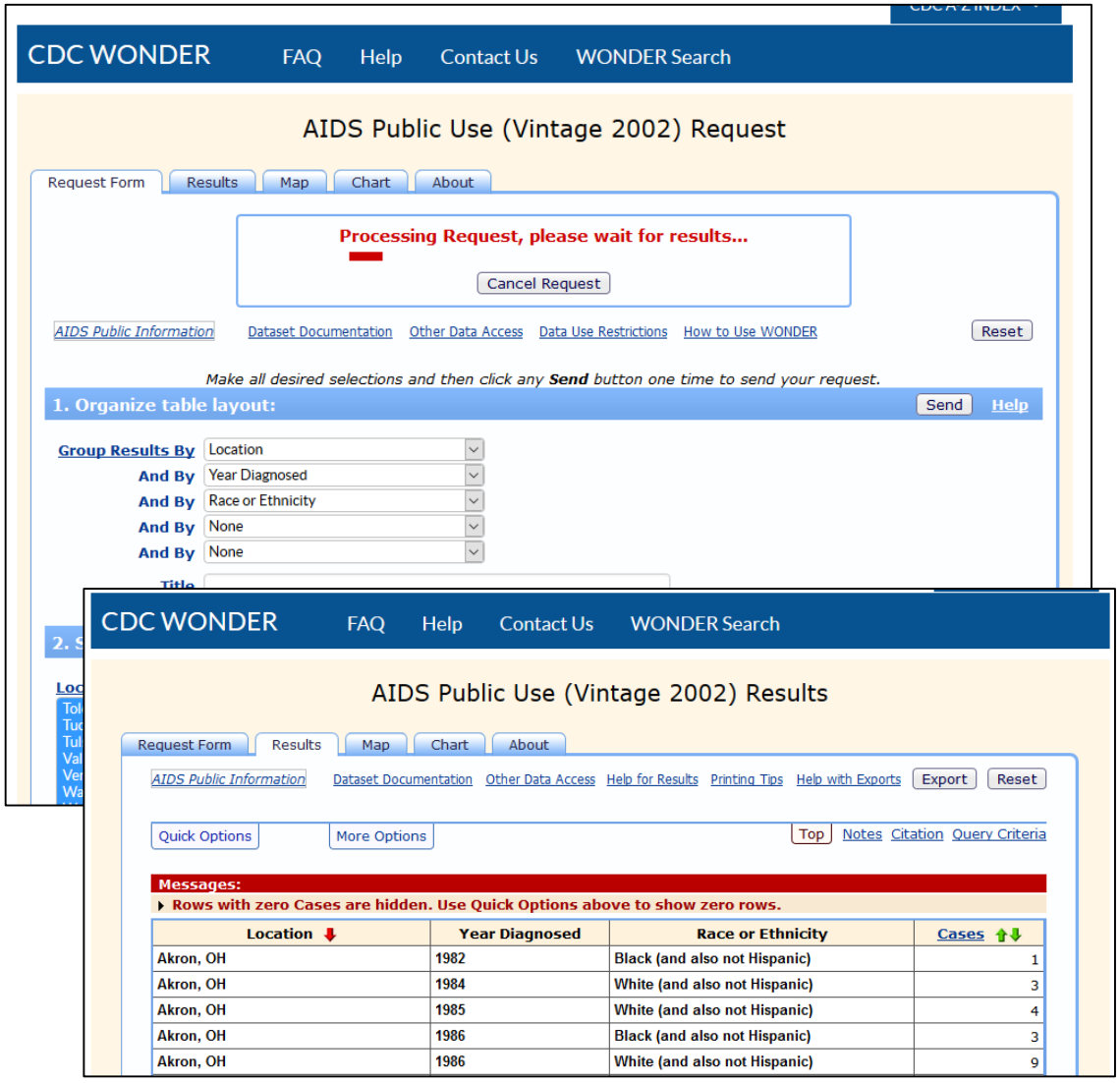
Proportional symbol map

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**.**



Proportional Symbol Map: CDC HIV/AIDS Diagnosis Data

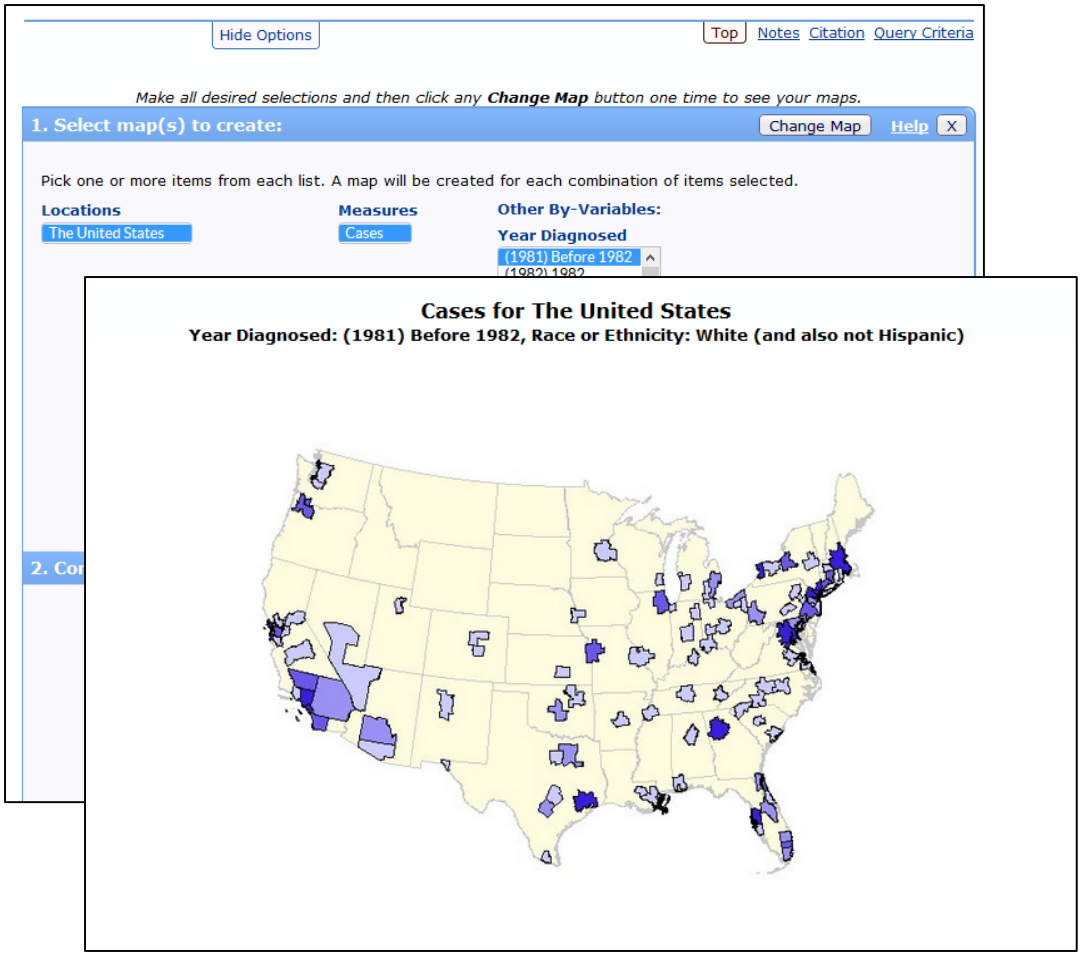


The AIDS Public Information Dataset U.S. Surveillance

"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."

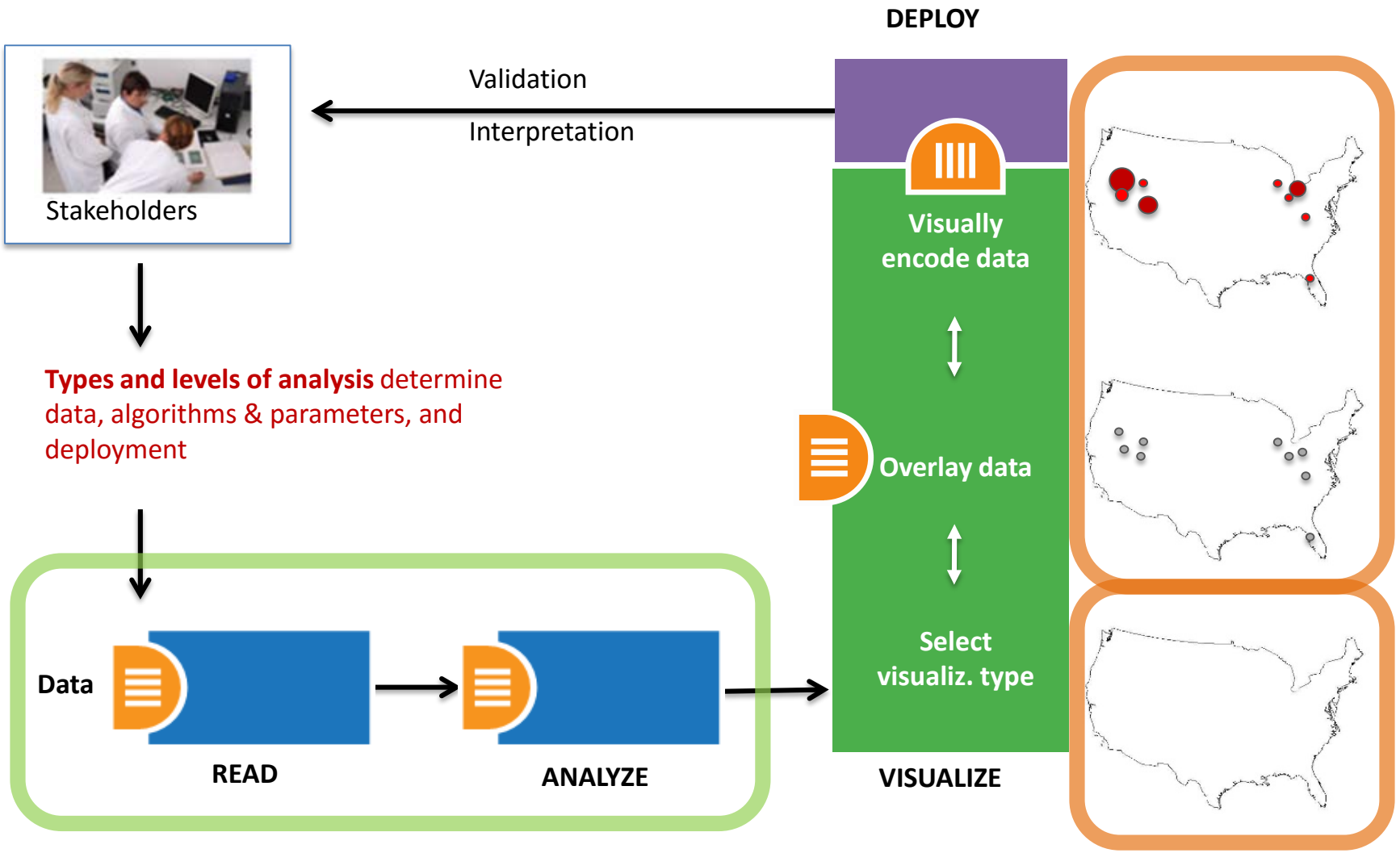
This workflow uses the **1981 – 2002 Archive case reports**: By date, place, demographics, case definition, risk factors and vital status.

Proportional Symbol Map: CDC HIV/AIDS Diagnosis Data

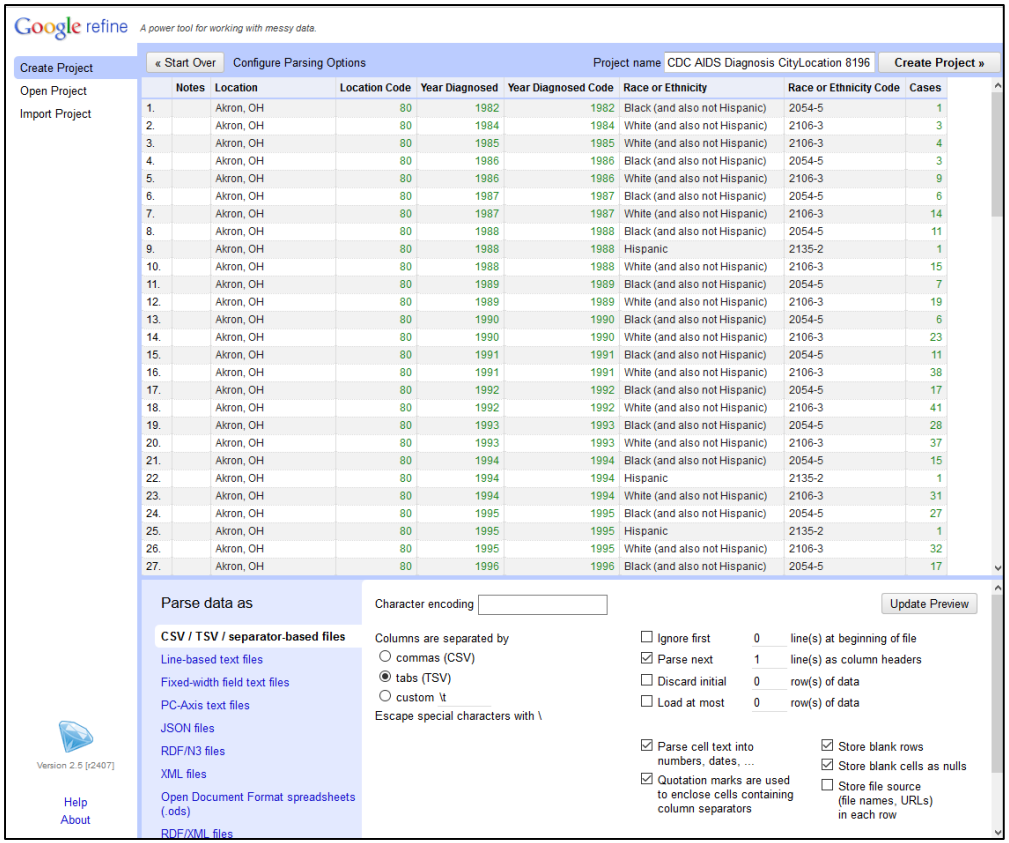


- Mapping on this data base is available, that provides many methods to subset queried data across variables provided.
- Limited to JPEG raster image.
- Works only with data sets available on CDC Wonder.

Workflow Design and Implementation



Open Refine: Review of CDC HIV/AIDS Diagnosis data



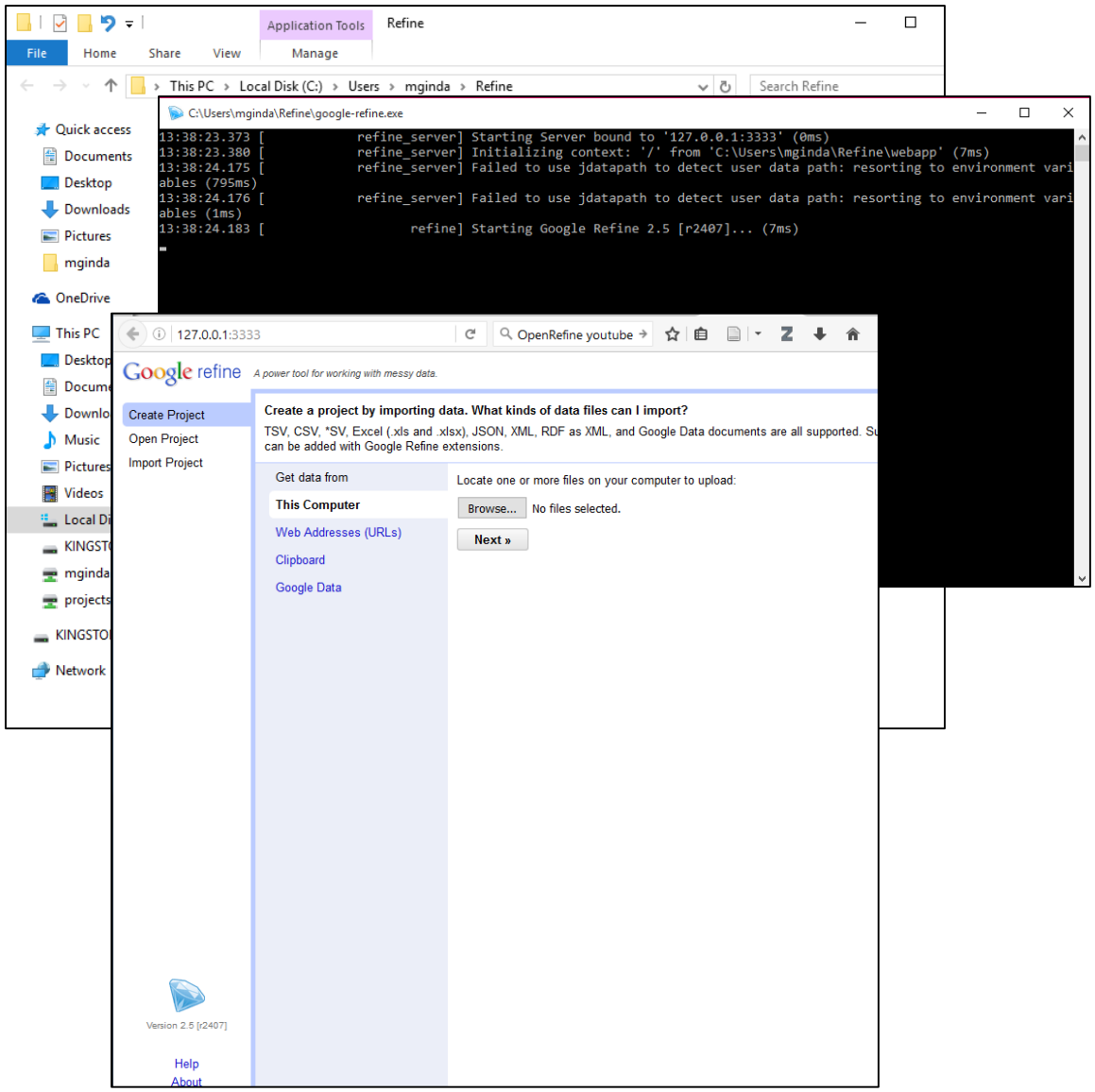
	Notes	Location	Location Code	Year Diagnosed	Year Diagnosed Code	Race or Ethnicity	Race or Ethnicity Code	Cases
1.		Akron, OH	80	1982	1982	Black (and also not Hispanic)	2054-5	1
2.		Akron, OH	80	1984	1984	White (and also not Hispanic)	2106-3	3
3.		Akron, OH	80	1985	1985	White (and also not Hispanic)	2106-3	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

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 Googles 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 the Google geocoder.

Open Refine: Creating an CDC HIV/AIDS Diagnosis data project



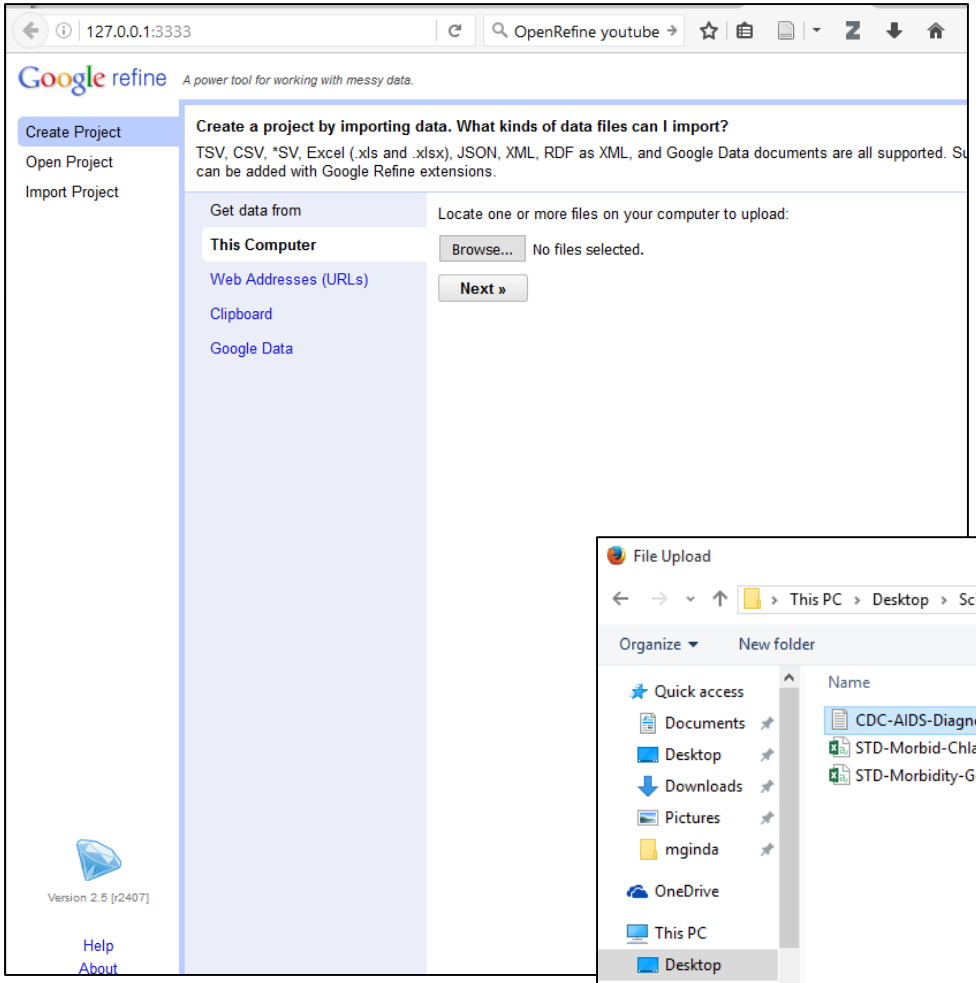
First we need to start OpenRefine. Navigate to the Desktop folder Refine.

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

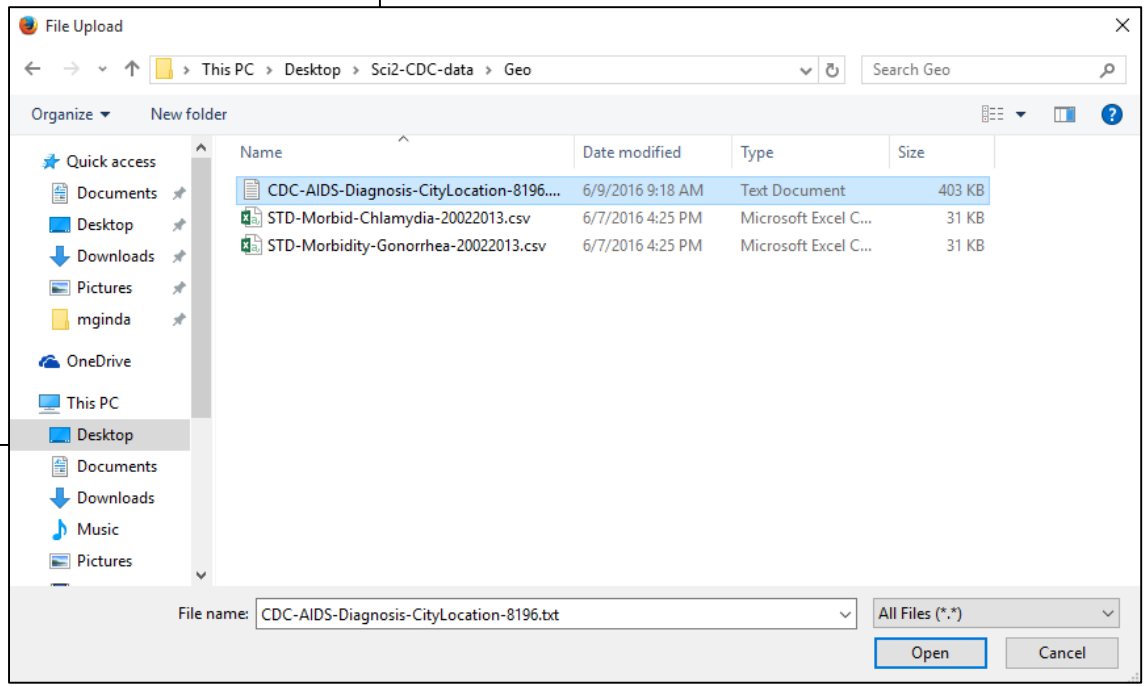
A console window will appear, this may be ignored for now, but can indicate error messages and process log.

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

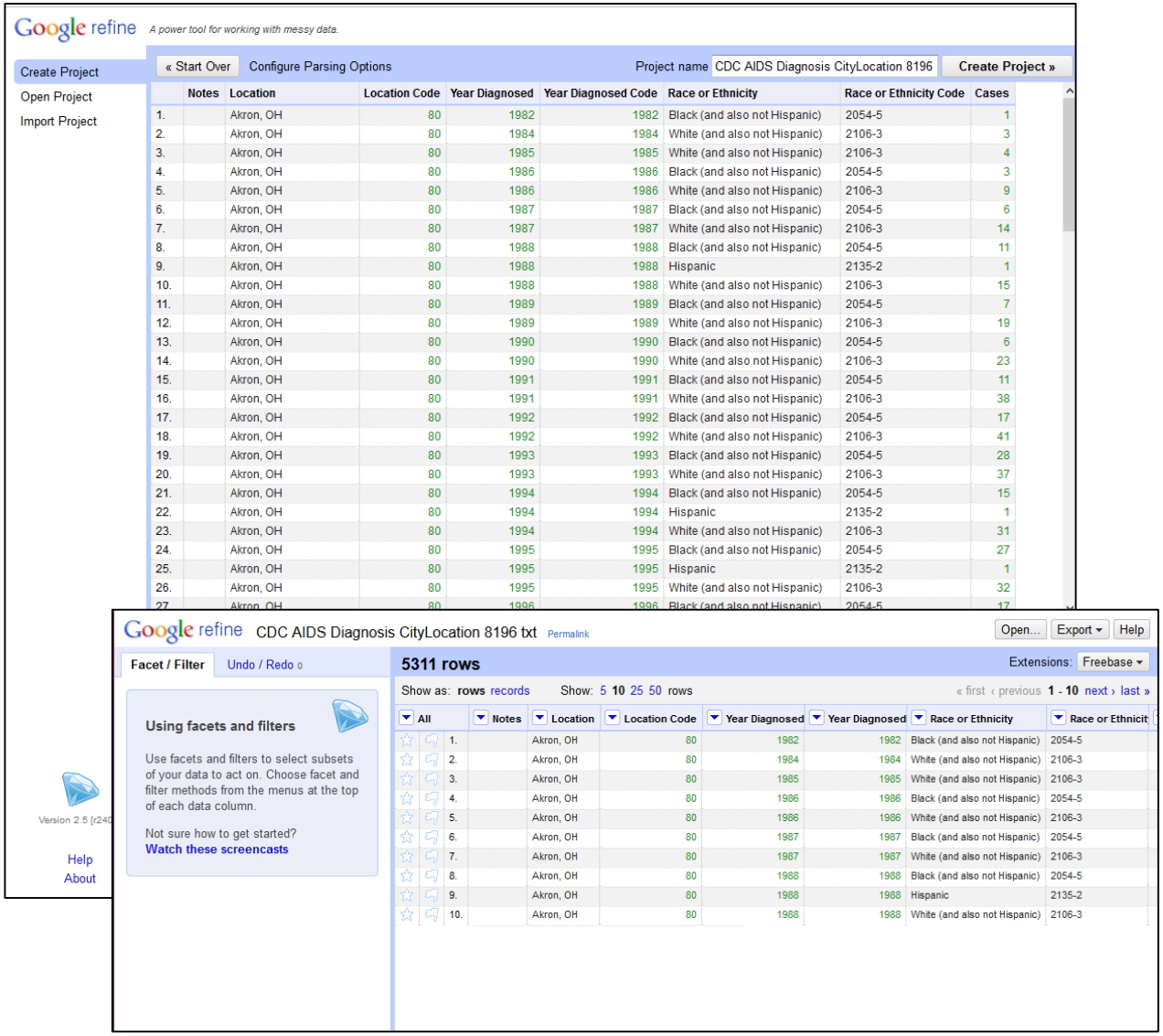
Open Refine: Creating an CDC HIV/AIDS Diagnosis data project



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.



Open Refine: Creating an CDC HIV/AIDS Diagnosis data project



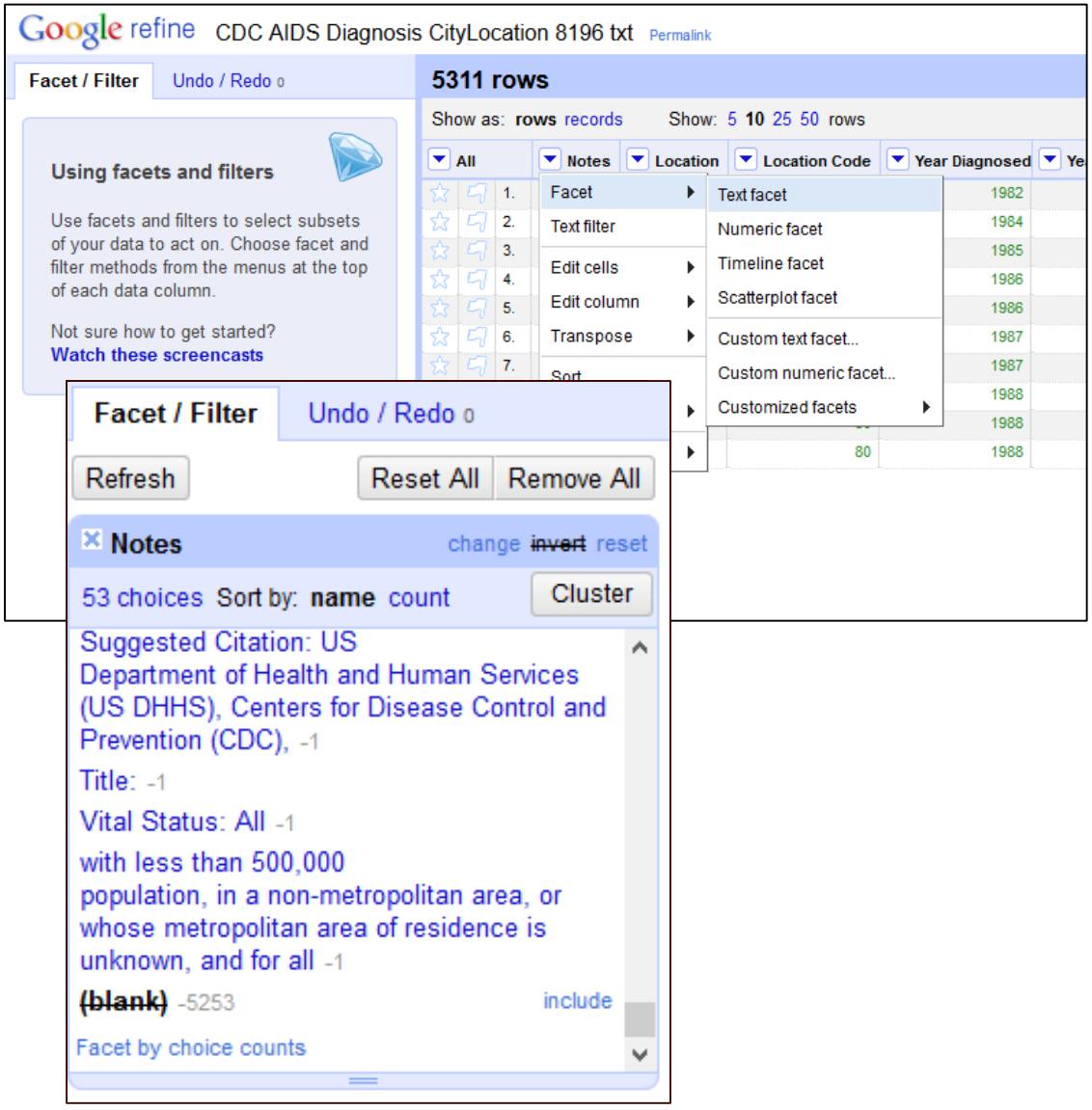
	Notes	Location	Location Code	Year Diagnosed	Year Diagnosed Code	Race or Ethnicity	Race or Ethnicity Code	Cases
1.		Akron, OH	80	1982	1982	Black (and also not Hispanic)	2054-5	1
2.		Akron, OH	80	1984	1984	White (and also not Hispanic)	2106-3	3
3.		Akron, OH	80	1985	1985	White (and also not Hispanic)	2106-3	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

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.

Open Refine: Review of CDC HIV/AIDS Diagnosis data

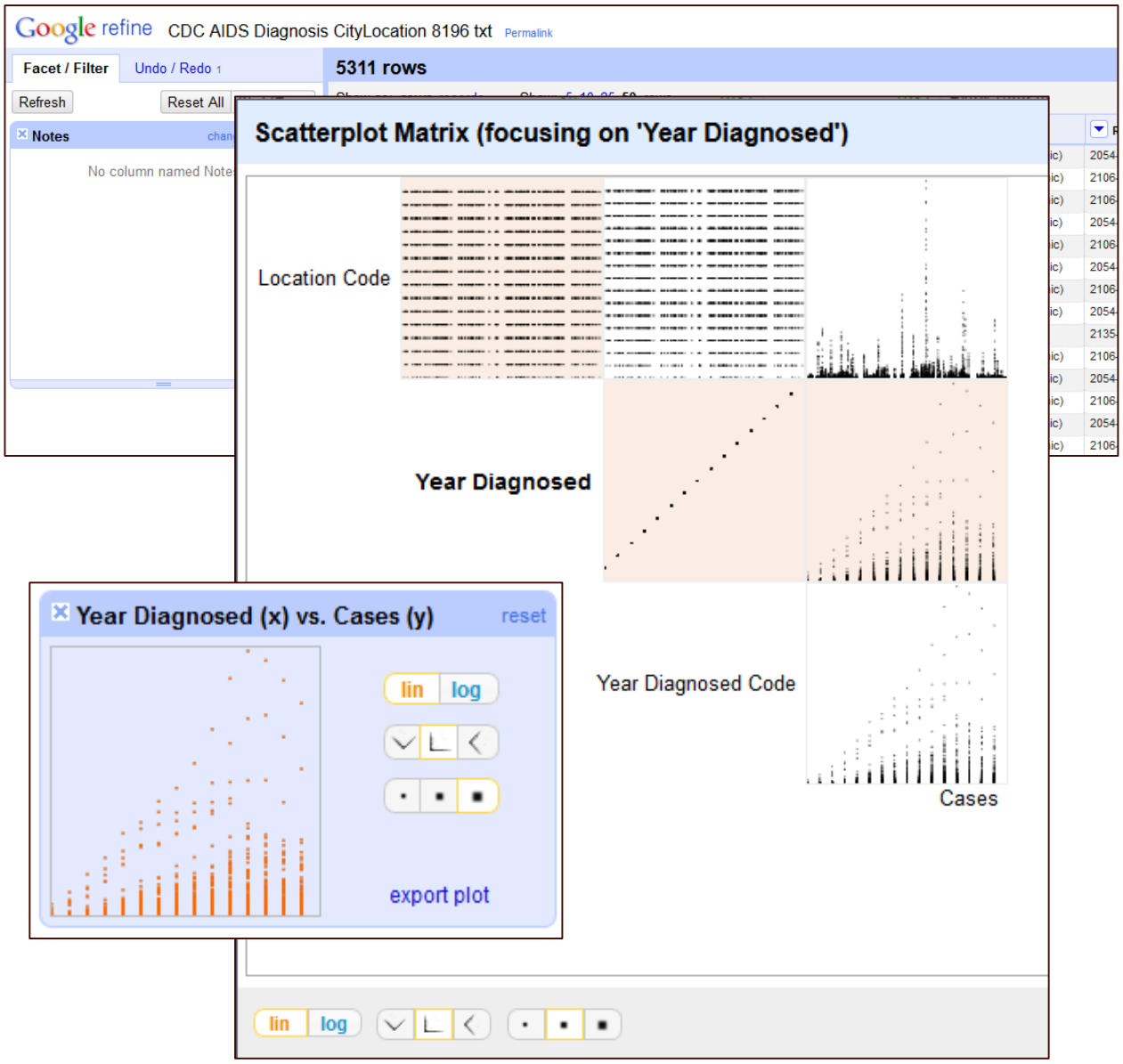


The first task is to review the fields in the data set. The “notes” column is a good place to start, as it looks empty.

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.

Open Refine: Review of CDC HIV/AIDS Diagnosis data

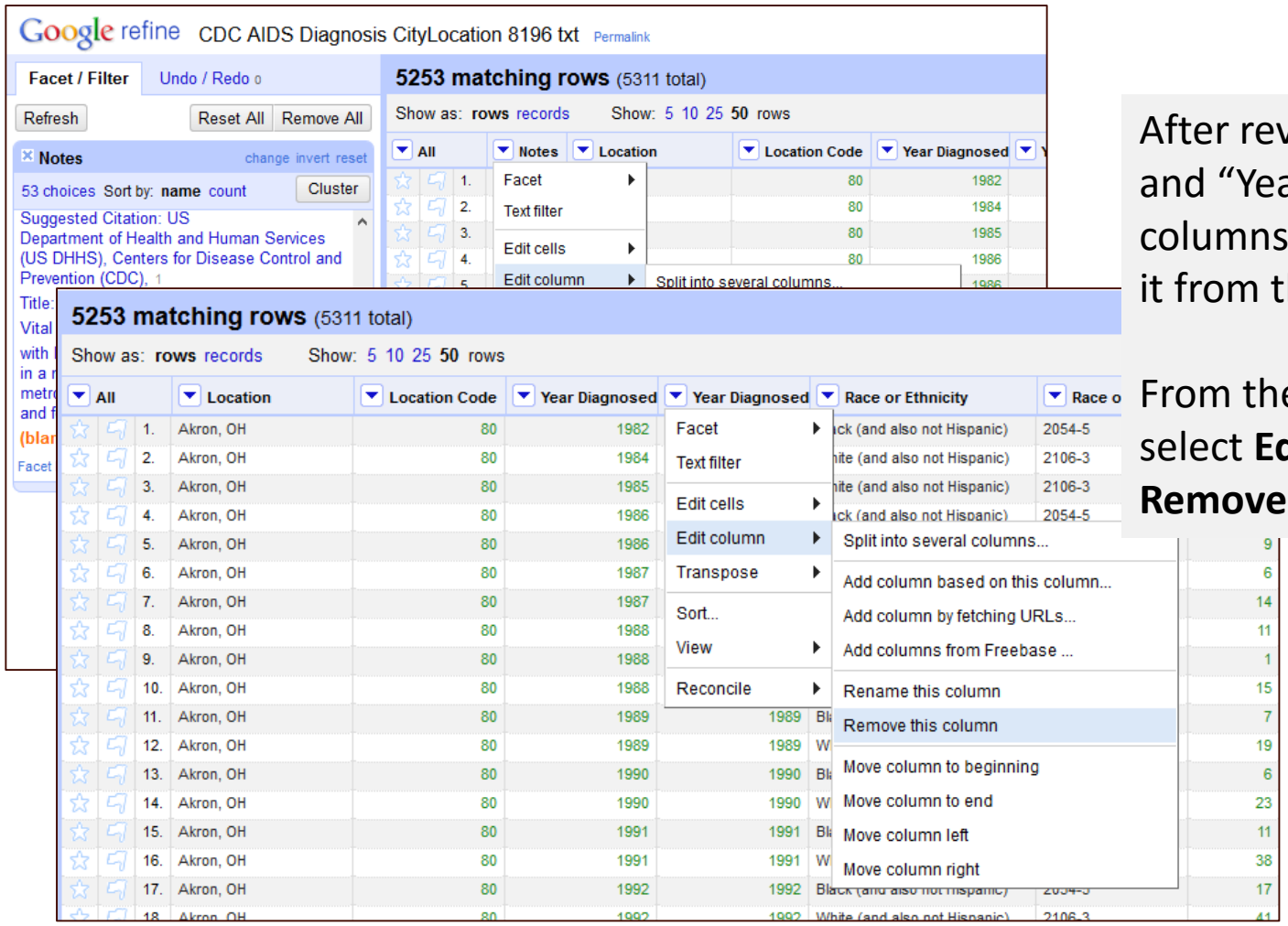


Next, lets review the **"Year Diagnosed"** columns. It appears that the both copies of the column are duplicated, but to find out we can use the "Scatter facet" to look at the relationships between numeric variables.

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.

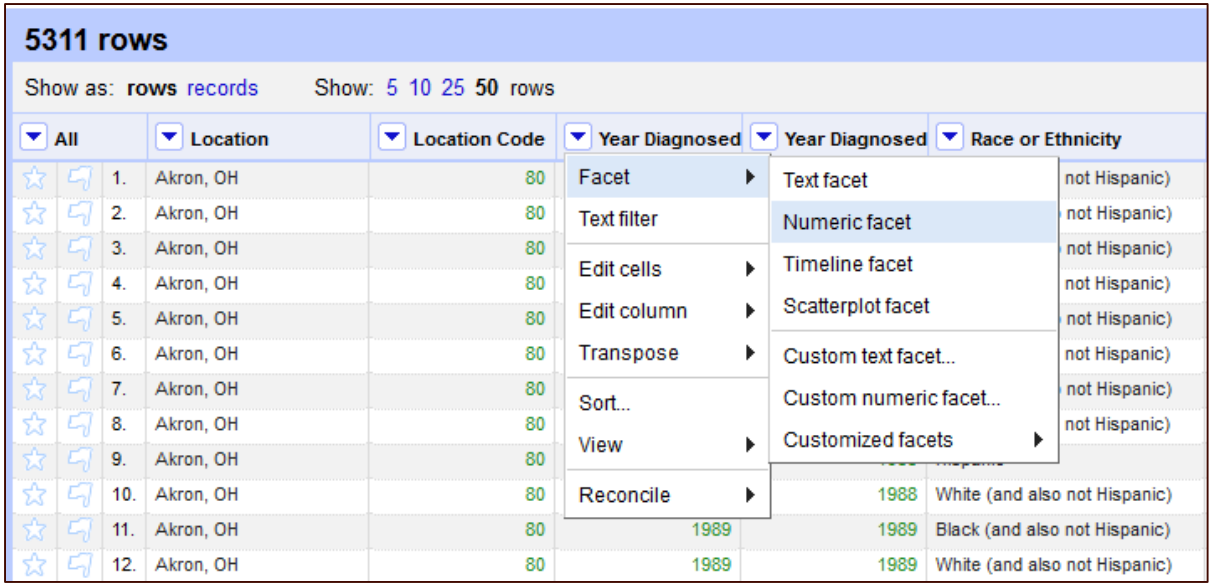
Open Refine: Cleaning the CDC HIV/AIDS Diagnosis data

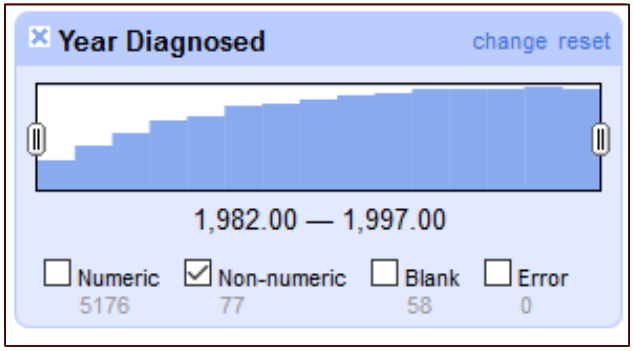


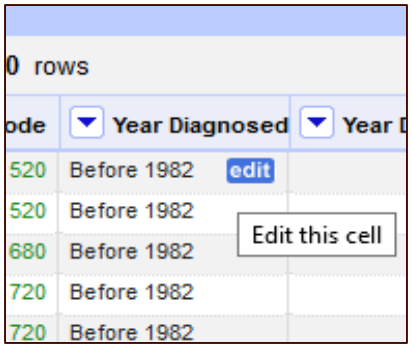
After reviewing the "Notes" and "Years diagnosed" columns, we need to remove it from the data.

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

Open Refine: Cleaning the CDC HIV/AIDS Diagnosis data







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.

Uncheck all but “non-numeric”.

Open Refine: Cleaning the CDC HIV/AIDS Diagnosis data

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

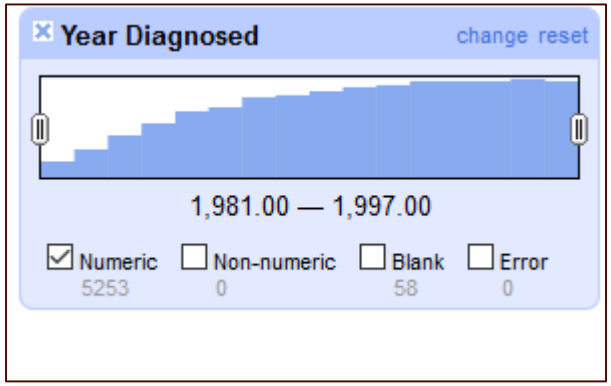
Data type: number

1981

Apply (Enter) | Apply to All Identical Cells (Ctrl-Enter) | Cancel (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.

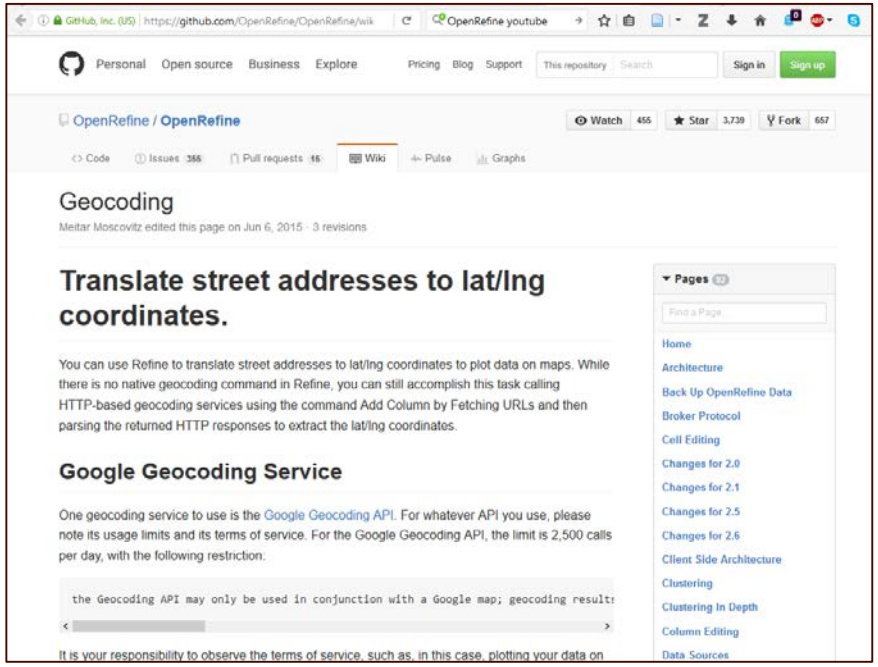


To see values, in the “Year Diagnosed” numeric filter, move the check from “Non-numeric” to “Numeric.” Note: “Blank” rows are leftover from the notes field. Keep this filter open.

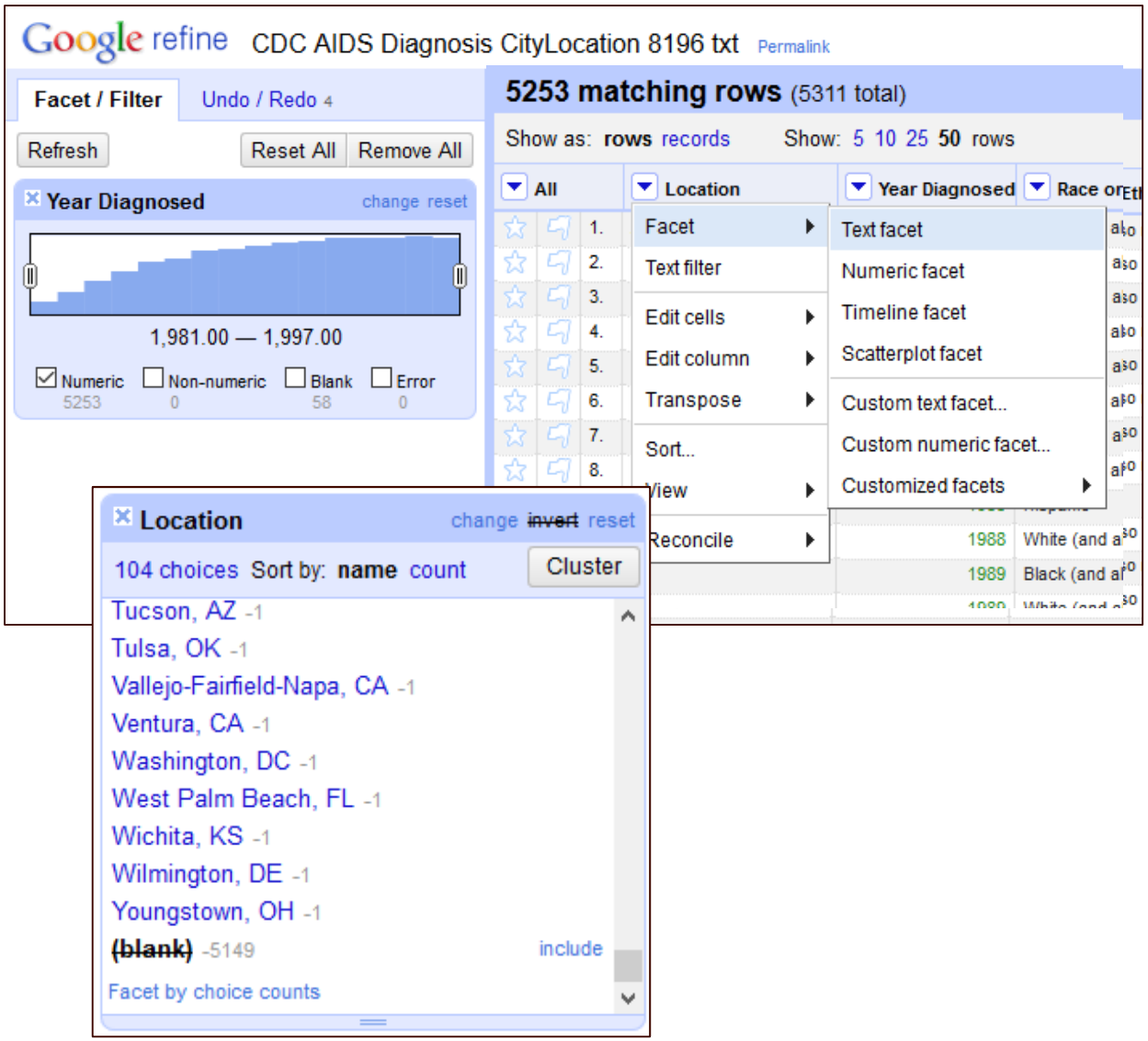
Open Refine: Geocoding CDC HIV/AIDS Diagnosis data

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



Open Refine: Geocoding HIV/AIDS Diagnosis data

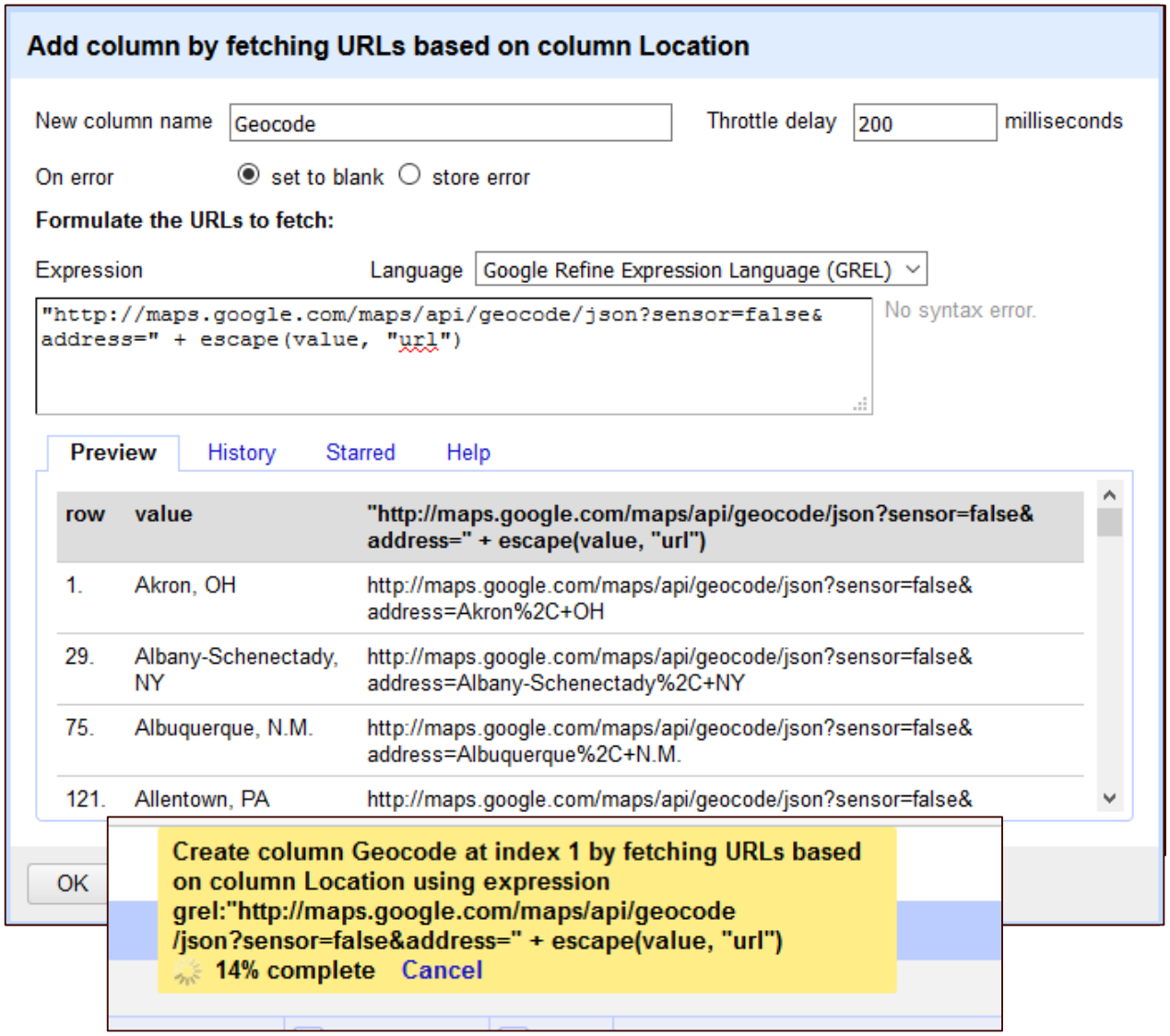


Users are limited to 2500 records per day. To avoid the data fetch limits, we must reduce the location records to the minimal amount of geolocations.

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.

Open Refine: Geocoding the HIV/AIDS Diagnosis data



Fetching

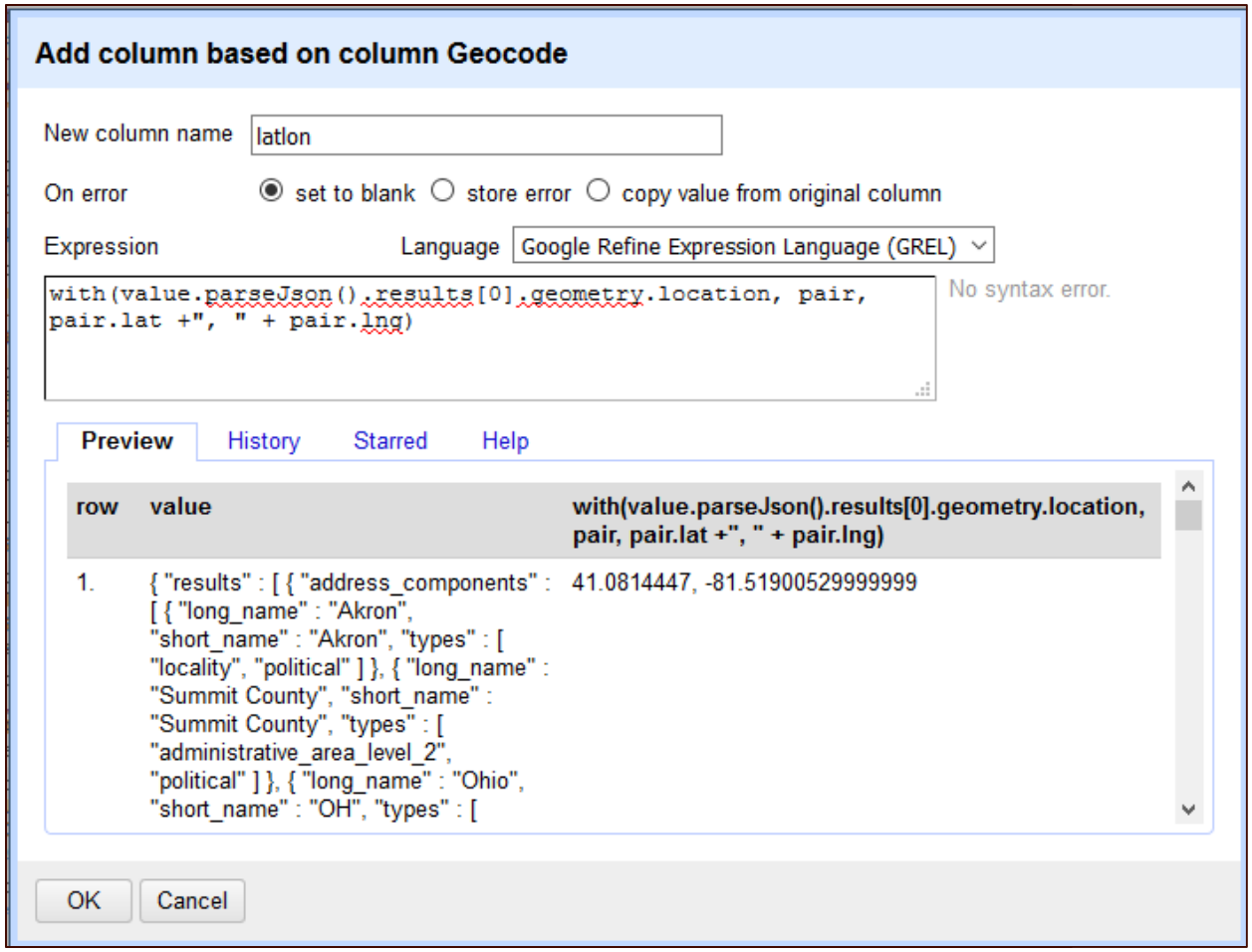
Set up Locations column to fetch geocode data by selecting in the drop down menu **Edit column** > **Add column by fetching url**, which opens a new window.

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")
```

Open Refine: Geocoding the CDC HIV/AIDS Diagnosis data



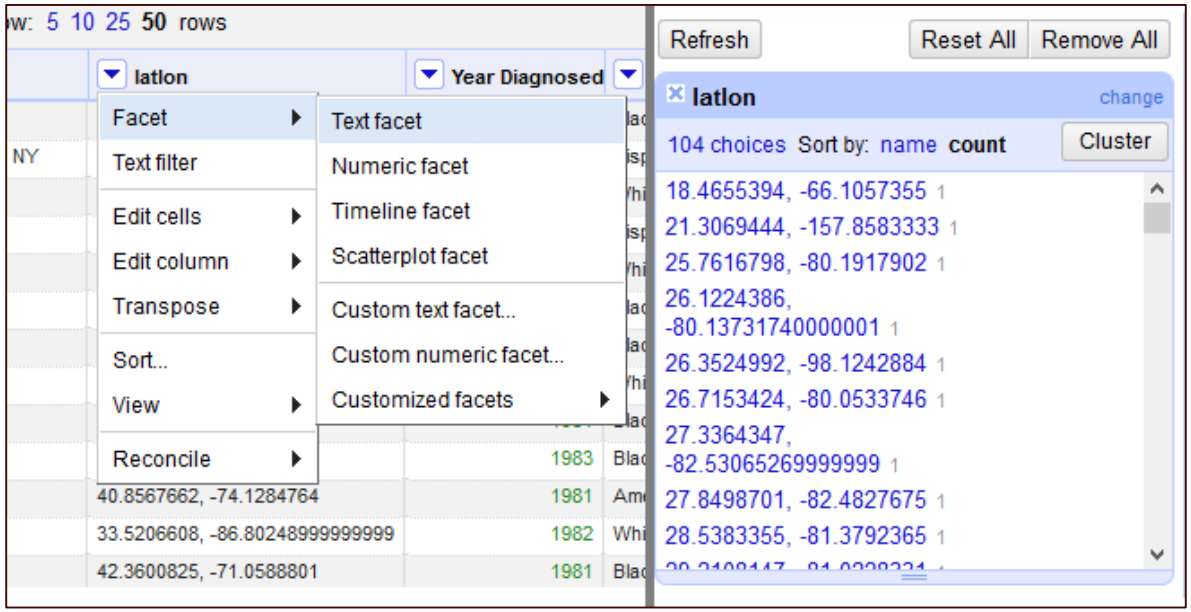
Parsing

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: Geocoding the CDC HIV/AIDS Diagnosis data

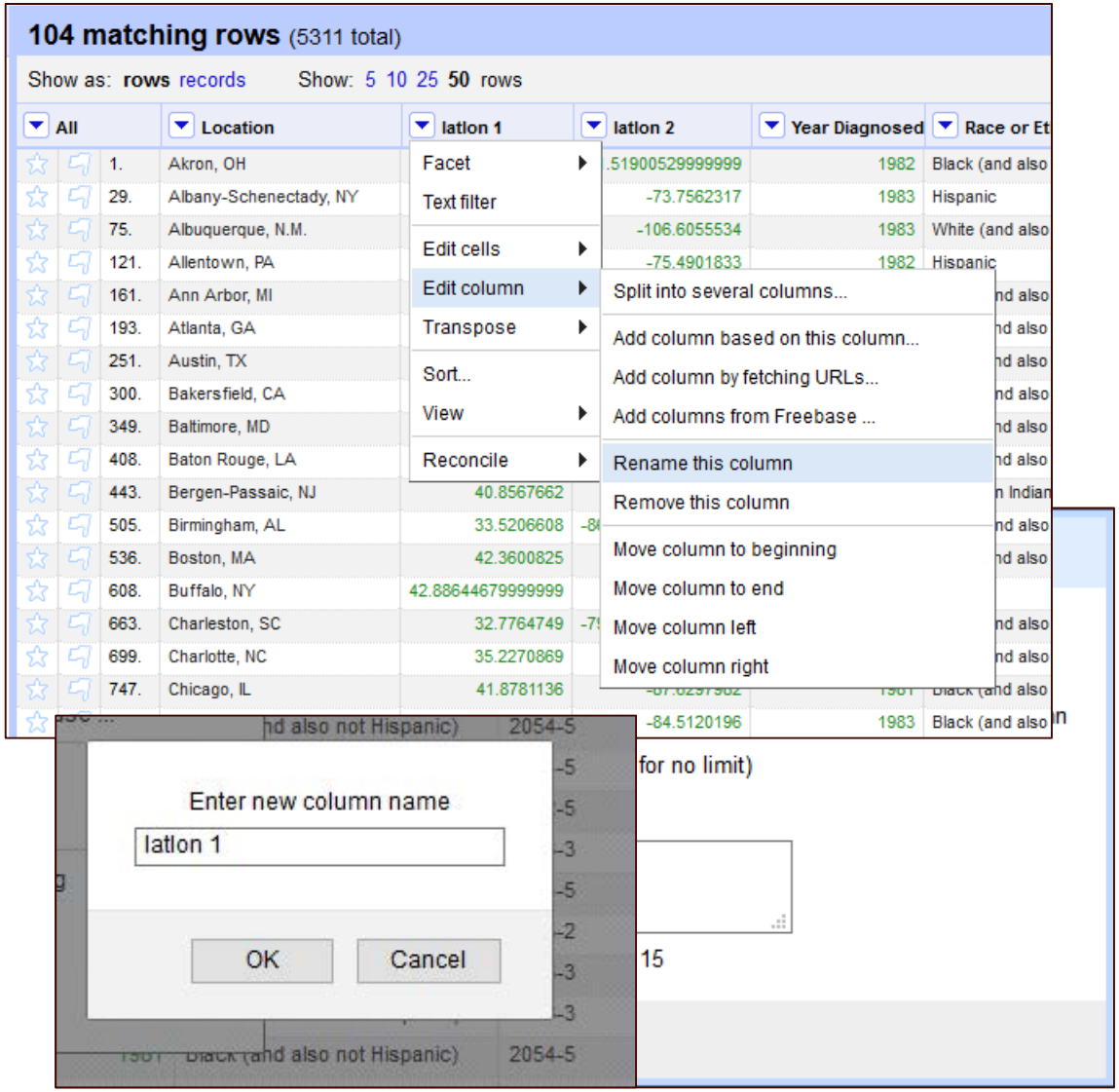


After parsing, we need to remove the **Geocode** column from the drop down menu select **Edit column > Remove column**

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.

Open Refine: Geocoding the CDC HIV/AIDS Diagnosis data

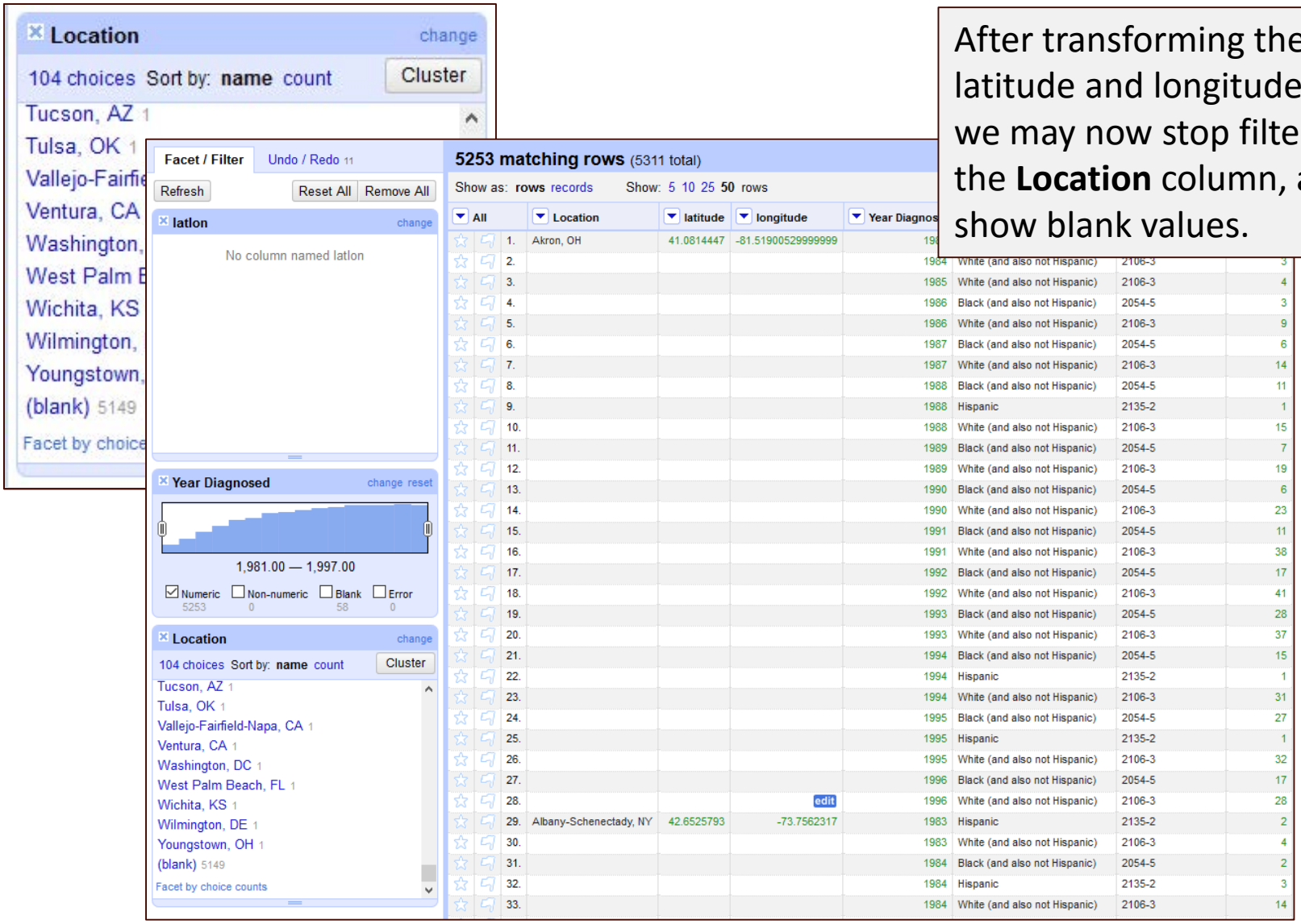


After determining if there are any duplicate locations, we need to split the **latlon** column into two, and then rename them

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".**

Open Refine: Filling the CDC HIV/AIDS Diagnosis data

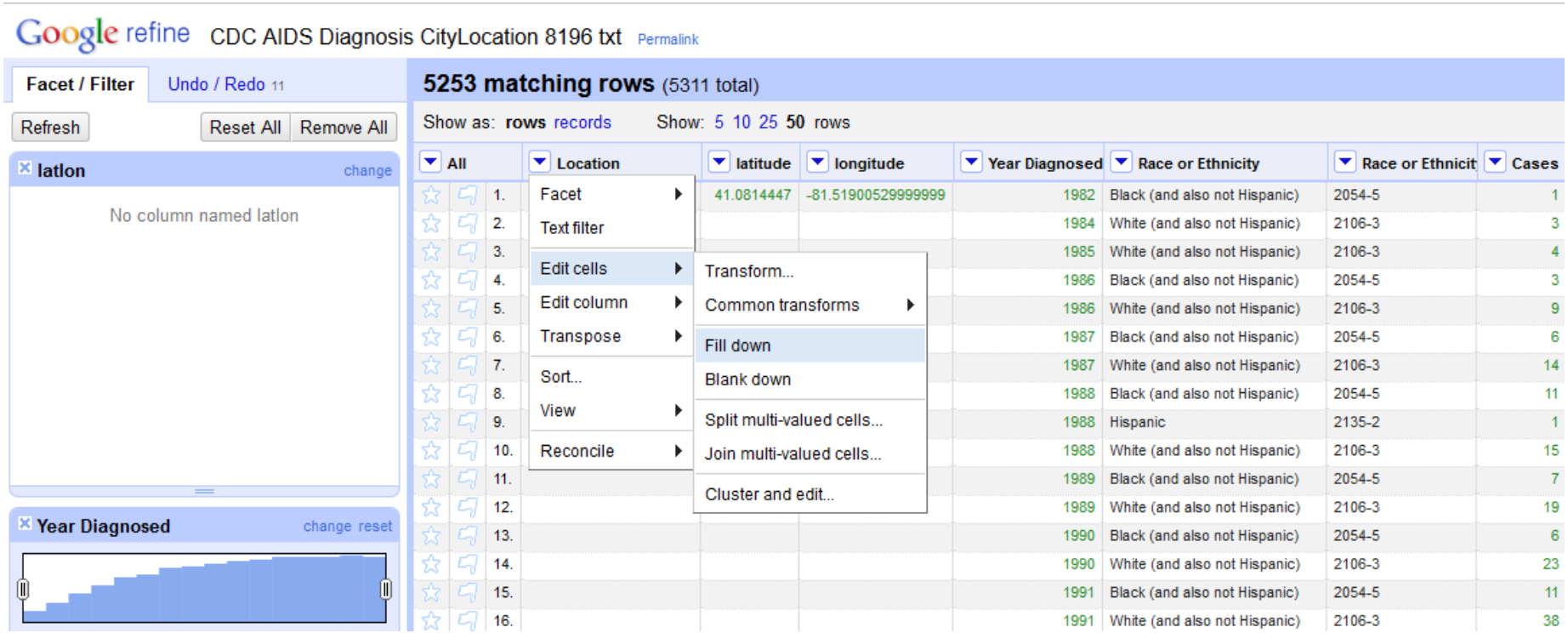


After transforming the latitude and longitude data, we may now stop filtering the **Location** column, and show blank values.

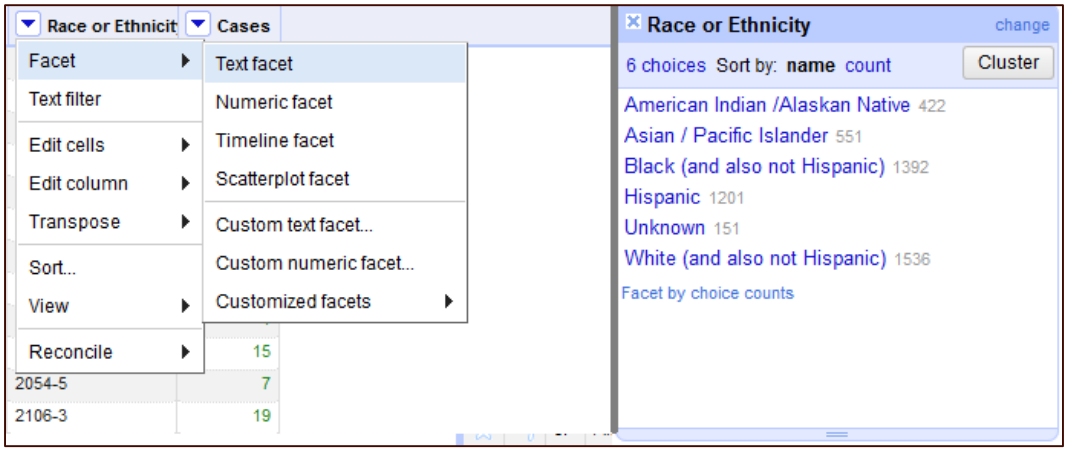
Open Refine: Filling the CDC HIV/AIDS Diagnosis data

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.



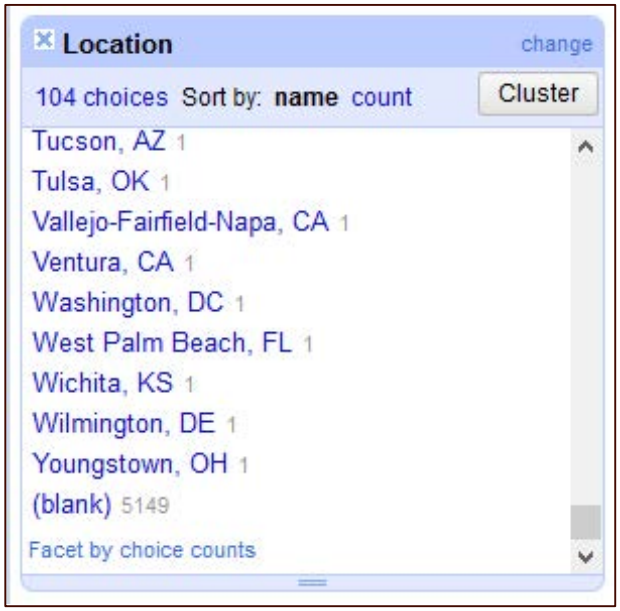
Open Refine: Subsetting the CDC HIV/AIDS Diagnosis data

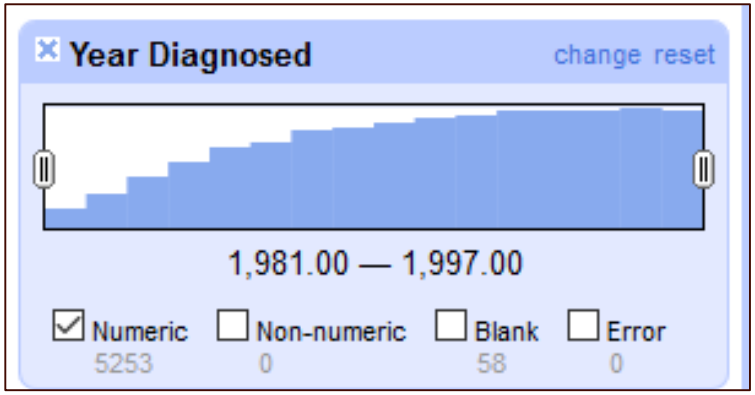


We also may remove the **Race or Ethnicity code** column.

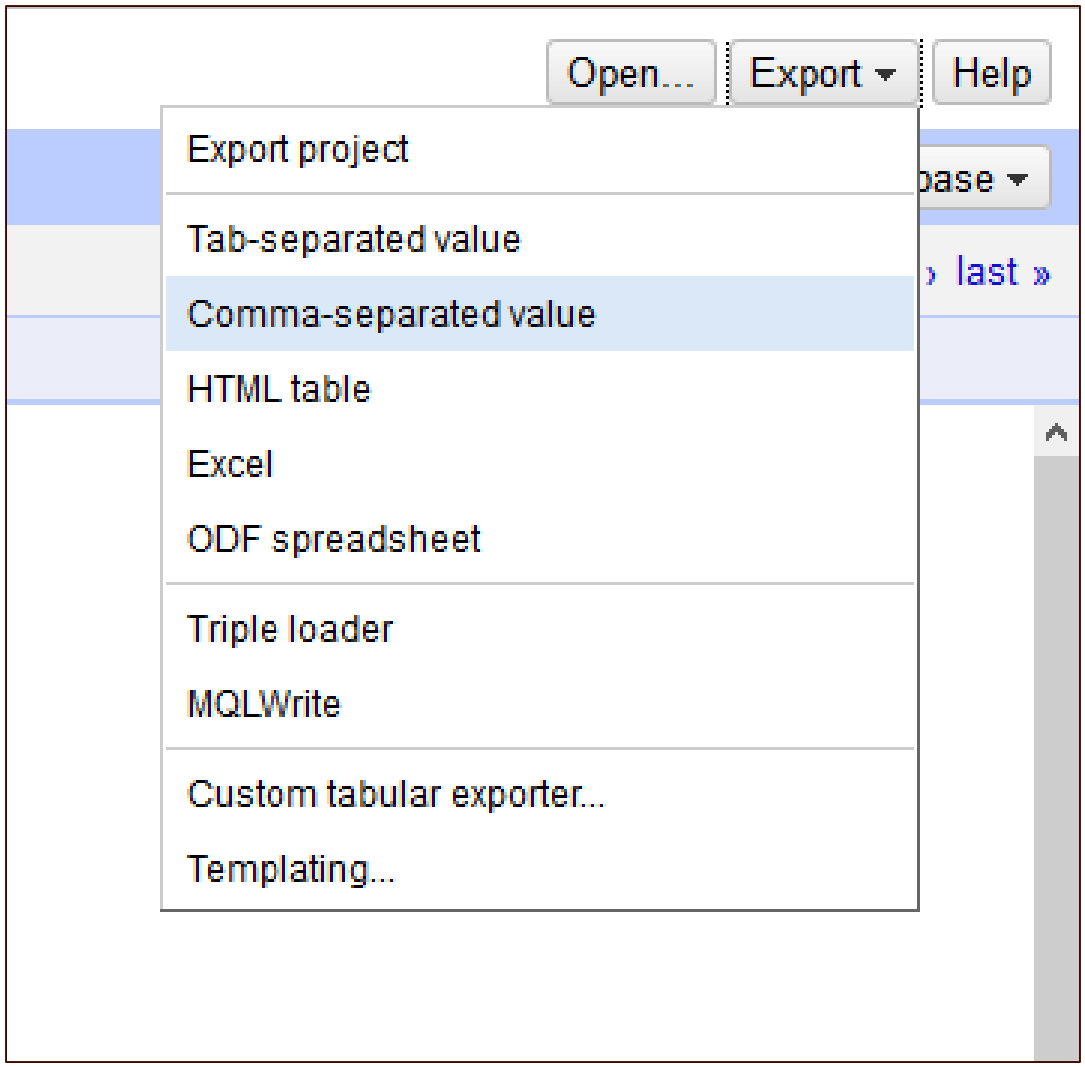
Finally, we want to filter the data set to leave only the values that we are interested. To help, we need to add one more text filter to the **Race or Ethnicity** field.

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





Open Refine: Exporting the CDC HIV/AIDS Diagnosis data



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**.

Tutorial Overview

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

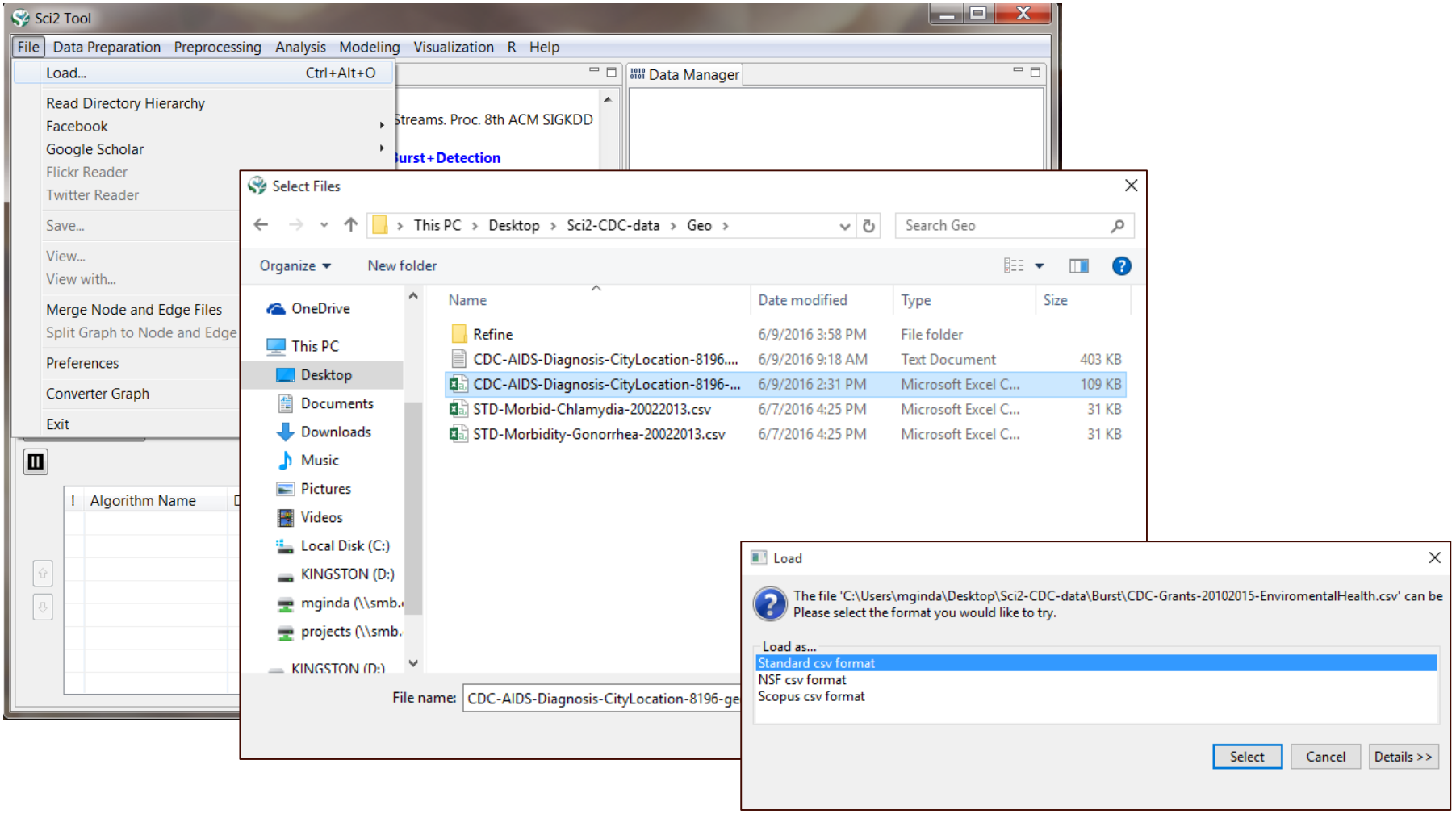
9:45 Geospatial Analysis
- Overview of geospatial analysis and mapping
- Geospatial Analysis: Geocoding with OpenRefine
- Geospatial Analysis: Proportional Symbol Map using CDC

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

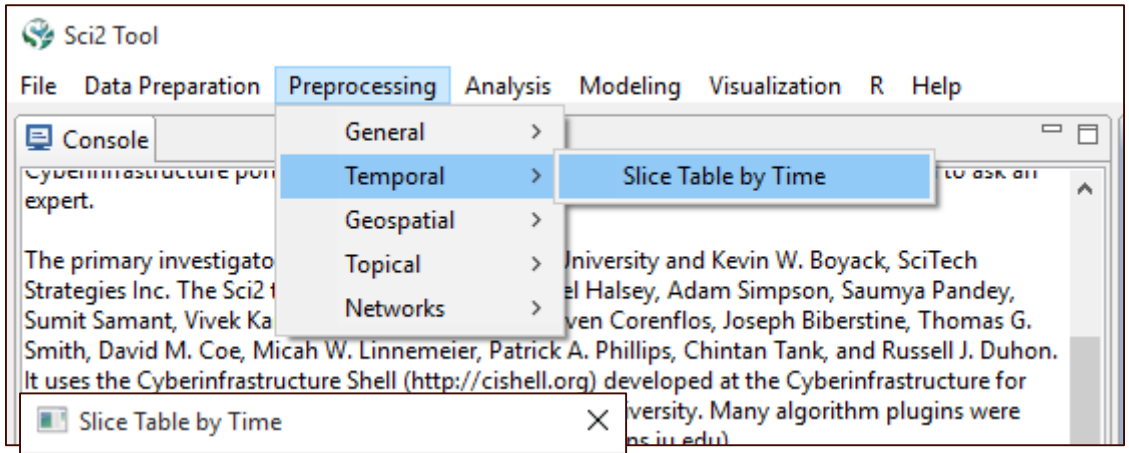
4:00 Wrap-up
4:30 Adjourn

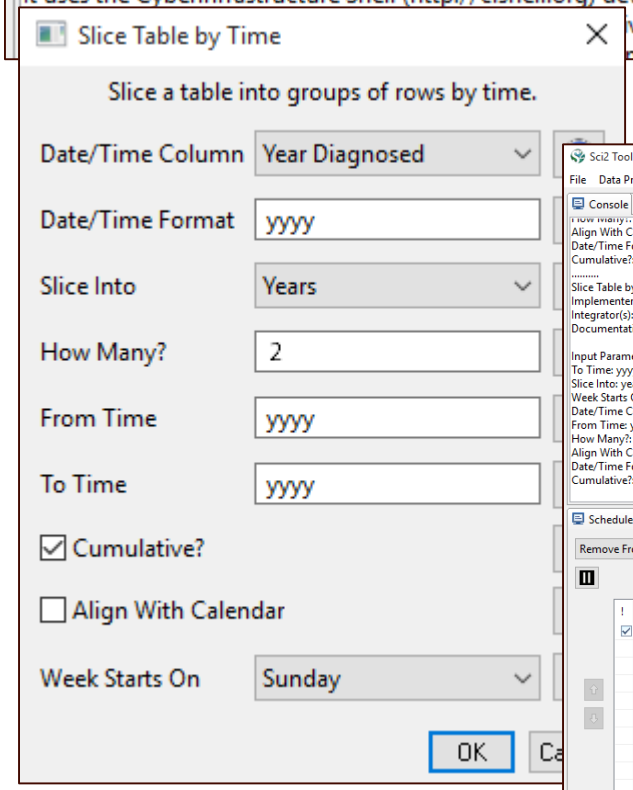
Sci2: Loading the CDC HIV/AIDS Diagnosis data

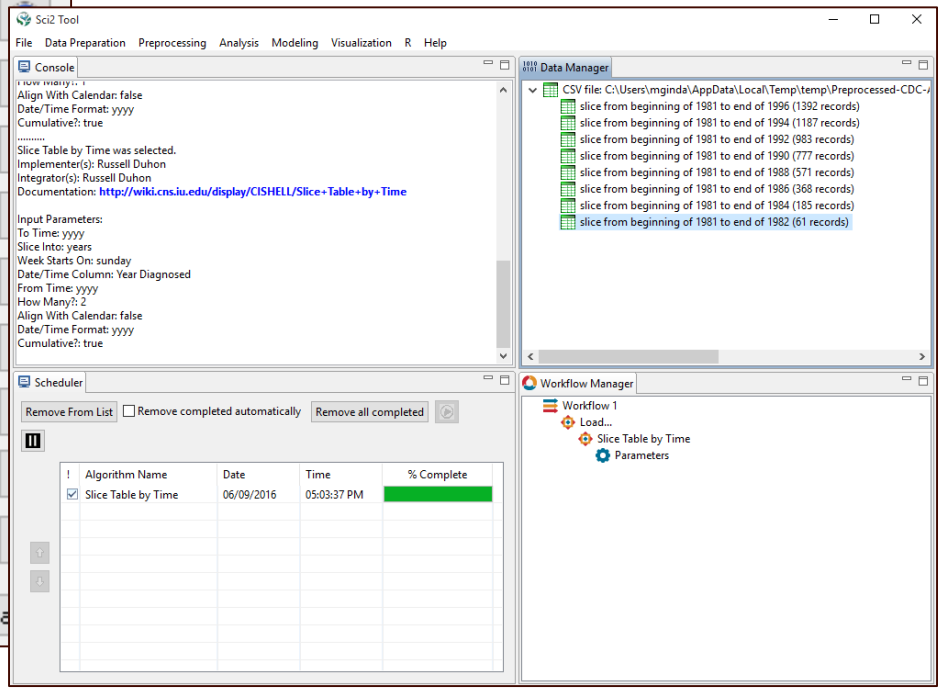


Load ***CDC-AIDS-Diagnosis-CityLocation-8196-geocoded-filtered-race-black.csv***
Located in CDC data directory: Sci2-CDC-data-> burst

Sci2: Time Slicing and Aggregating CDC HIV/AIDS Diagnosis data



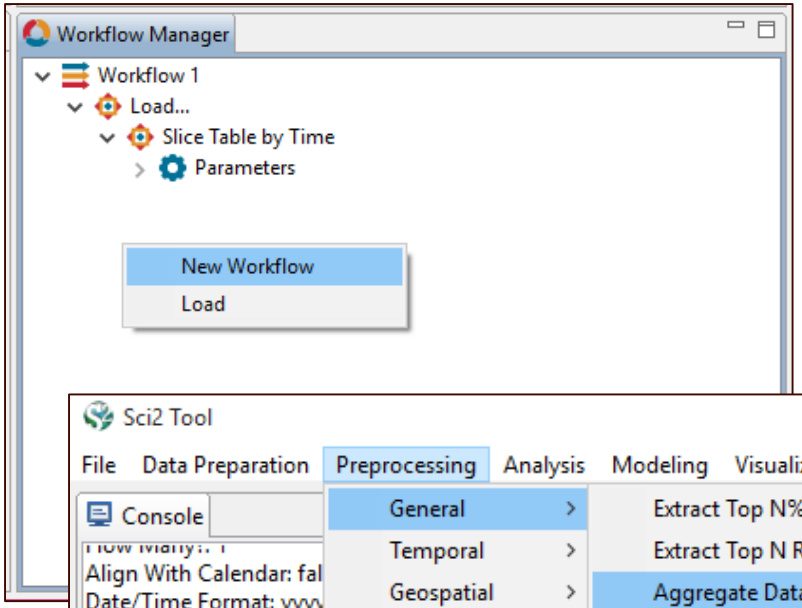




Next, navigate **Preprocessing > Temporal > Slice Table by Time** and then enter the following parameters in the pop-up box.

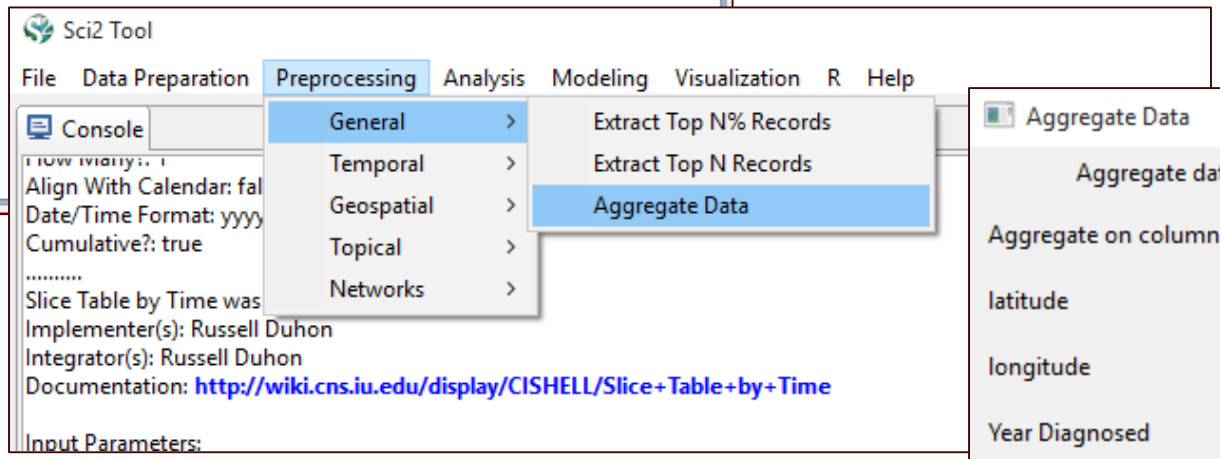
We will create a cumulative series of maps for two year periods.

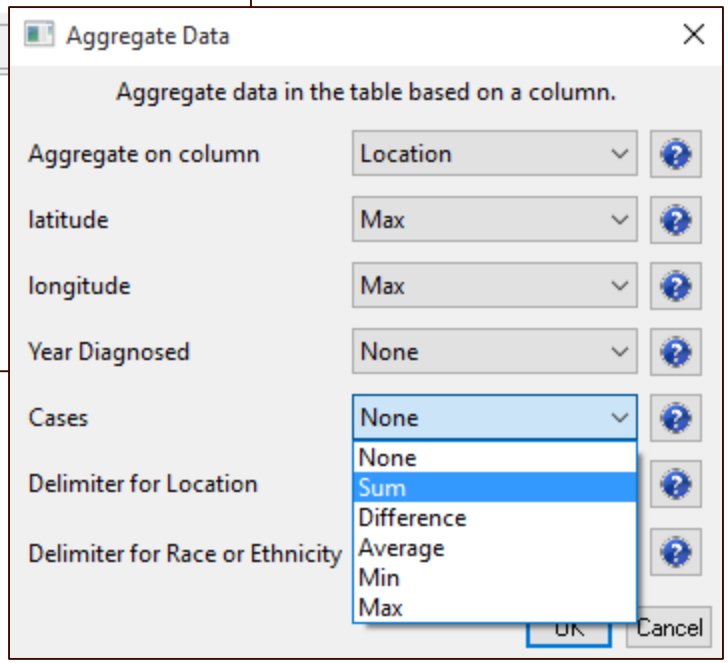
Sci2: Time Slicing and Aggregating CDC HIV/AIDS Diagnosis data



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”:

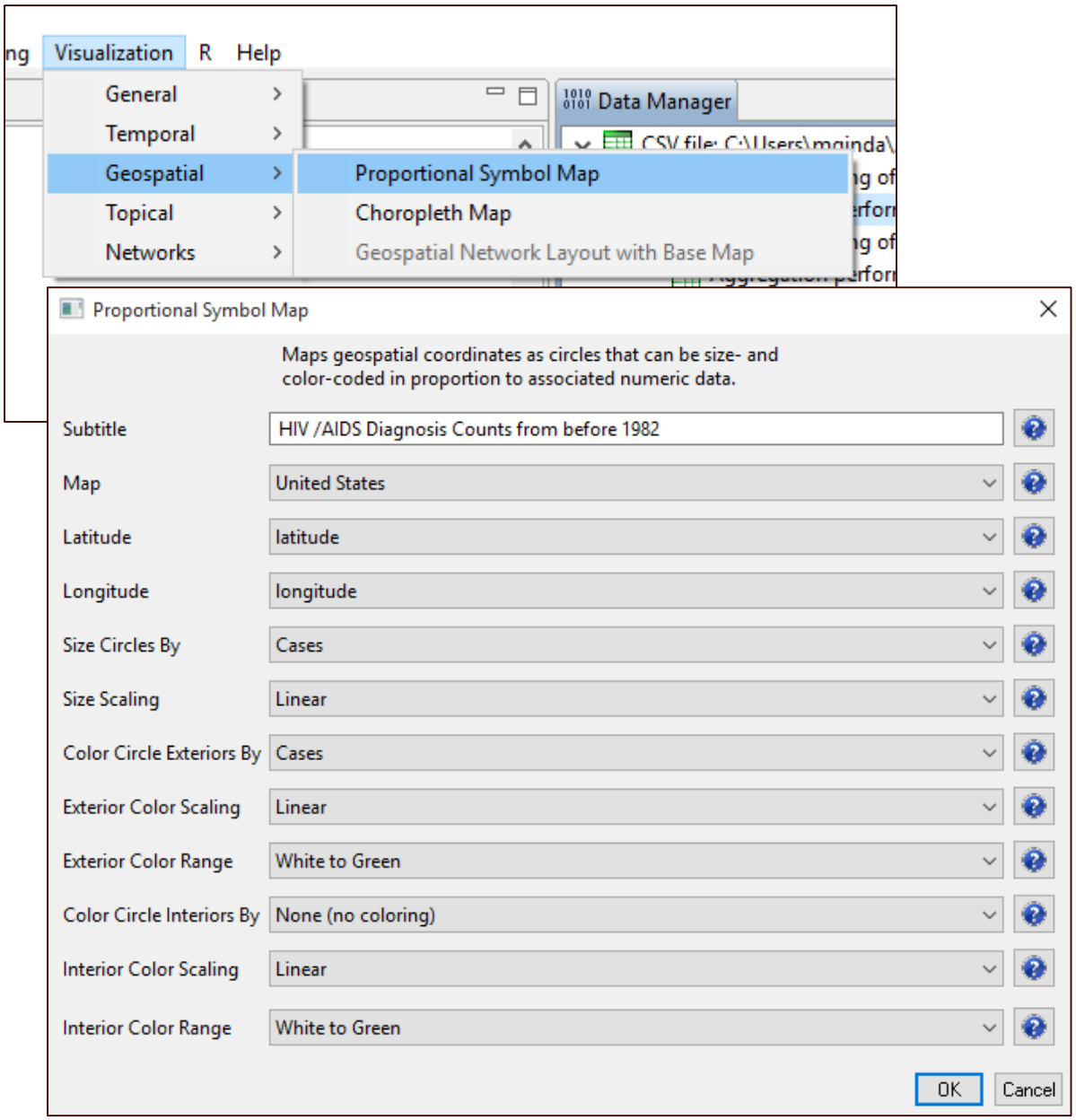
Sci2: Time Slicing and Aggregating CDC HIV/AIDS Diagnosis data

	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.

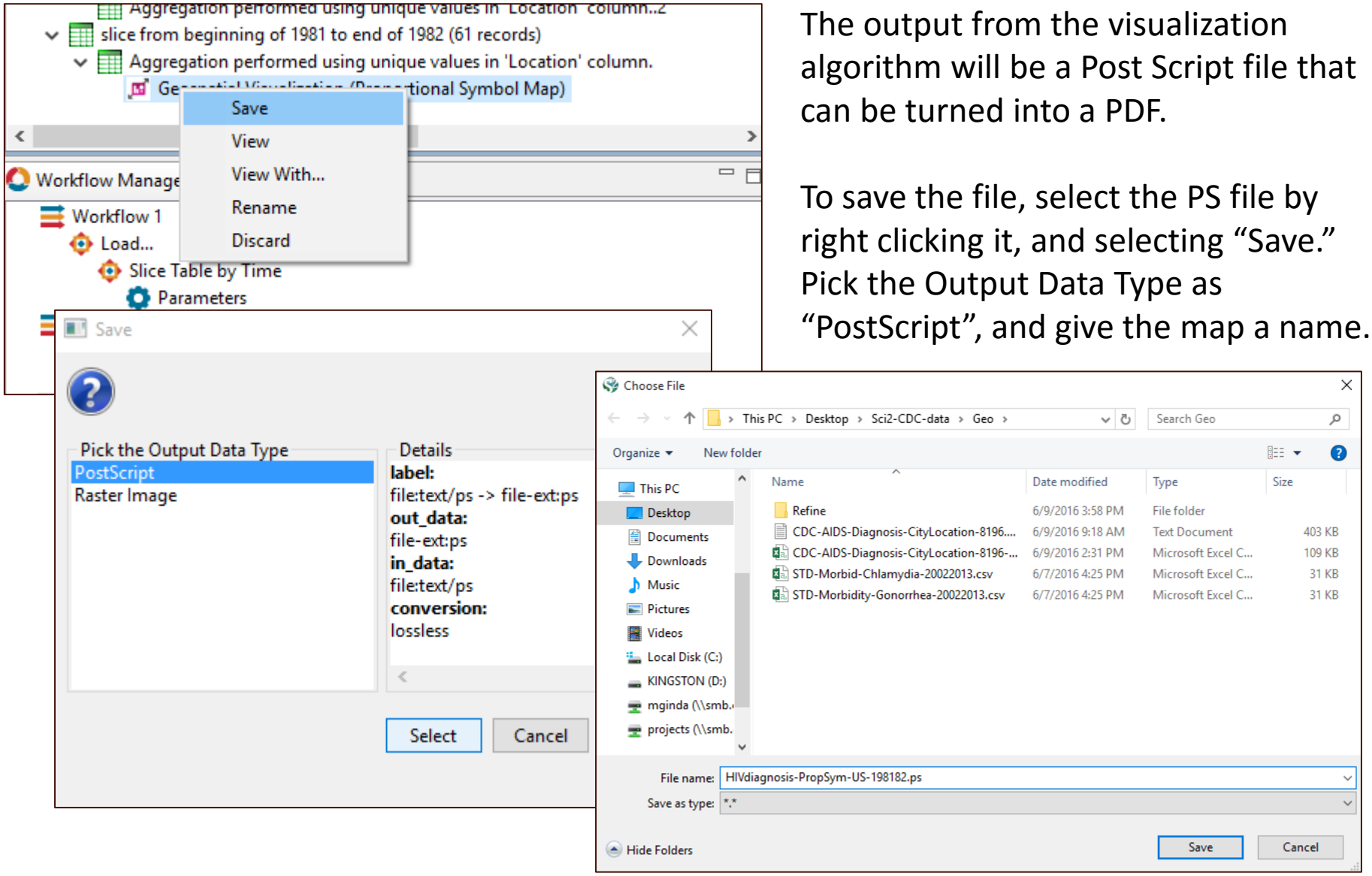
Sci2: Visualizing the CDC HIV/AIDS Diagnosis data



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”.

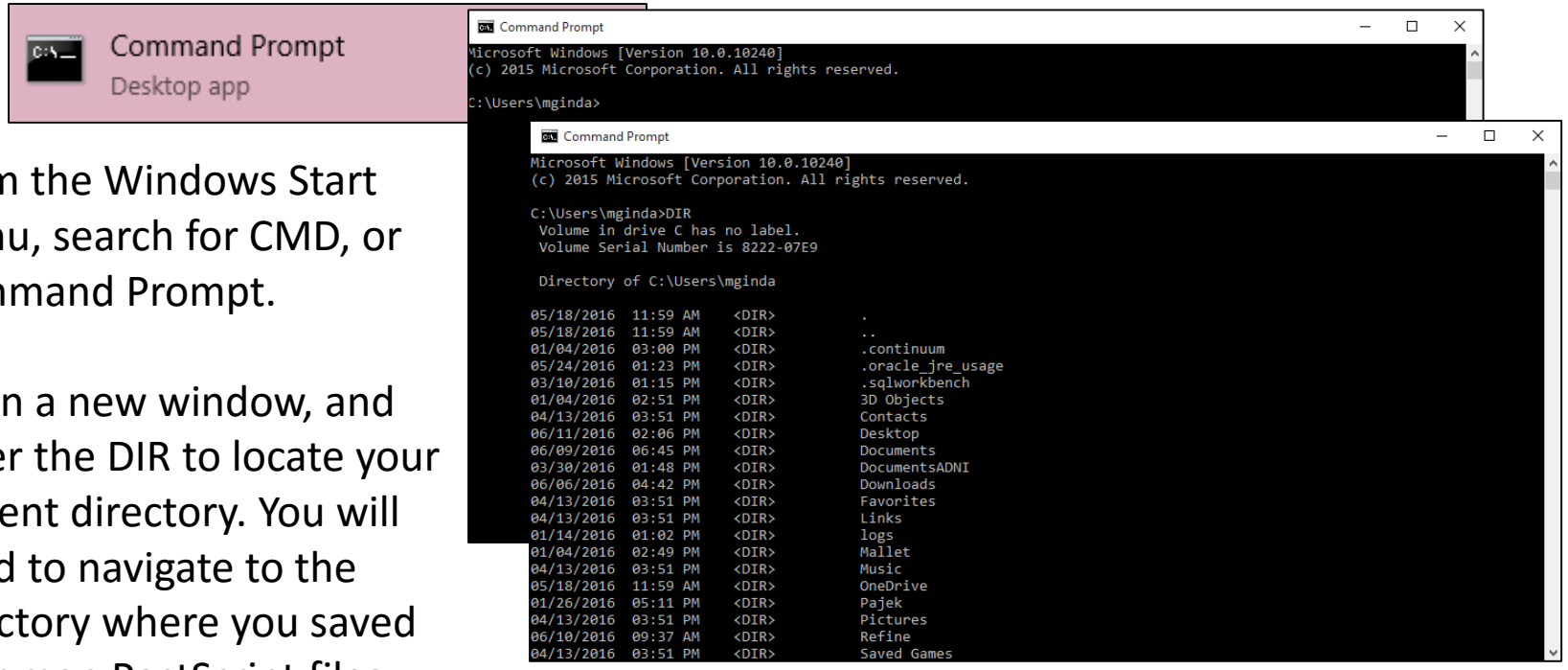
Sci2: Visualizing the CDC HIV/AIDS Diagnosis data



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.

Sci2: Converting the Output PostScript Maps to PDF



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.

Sci2: Converting the Output PostScript Maps to PDF

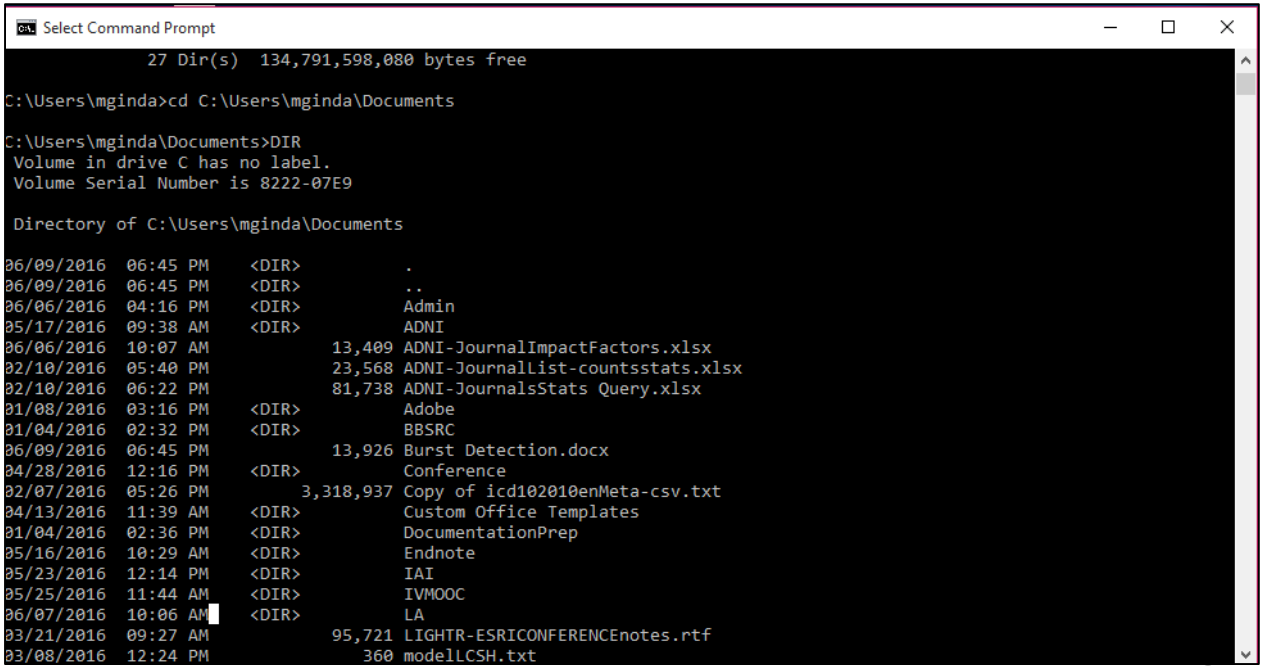
A common place to save a file may be the Desktop or the Documents folder.

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

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

or

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

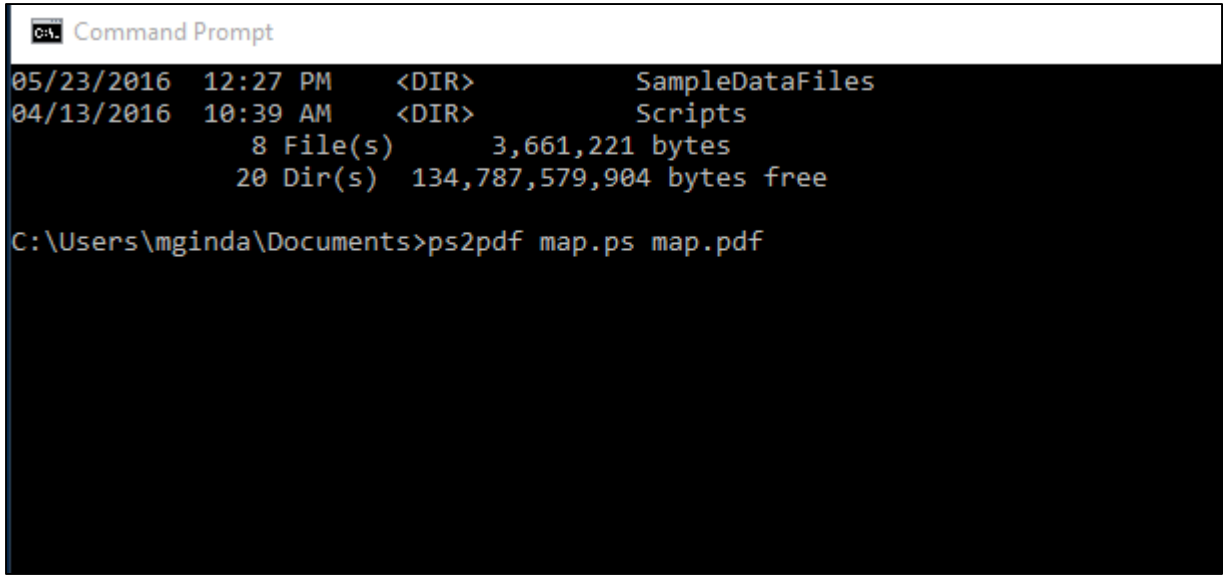


Sci2: Converting the Output PostScript Maps to PDF

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
```

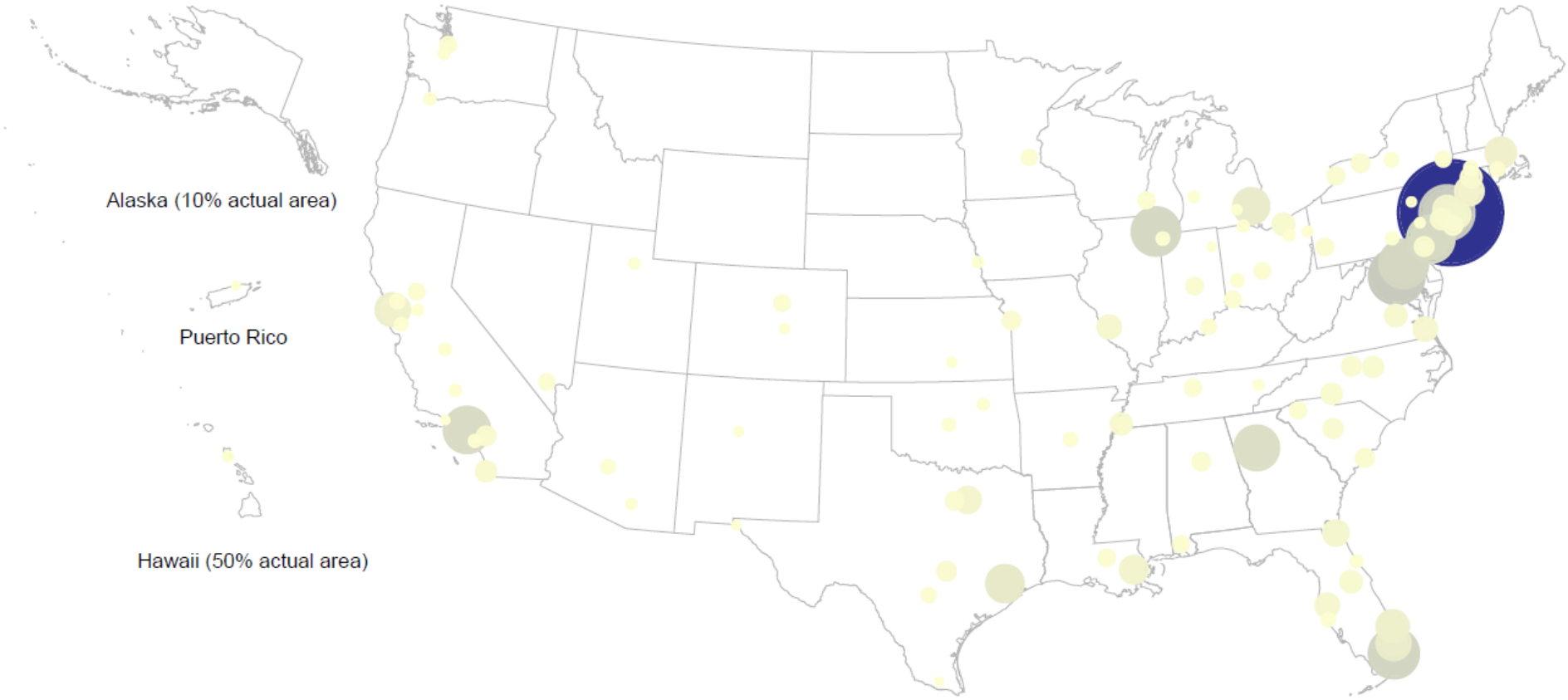
In you leave off the output file name, the resulting PDF file will have the same name as the original file with a .PDF extension into the same directory you navigated to.



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

3 15,978 31,953

Exterior Color (Linear)

Cases

3 15,978 31,953

Area (Linear)

Cases

31,953
15,978
3

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?

Tutorial Overview

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

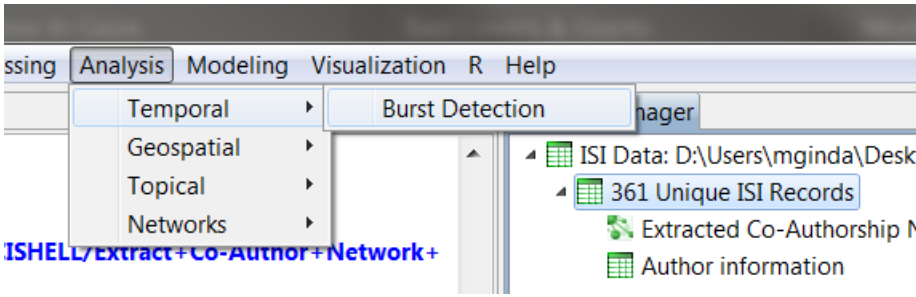
4:00 Wrap-up
4:30 Adjourn

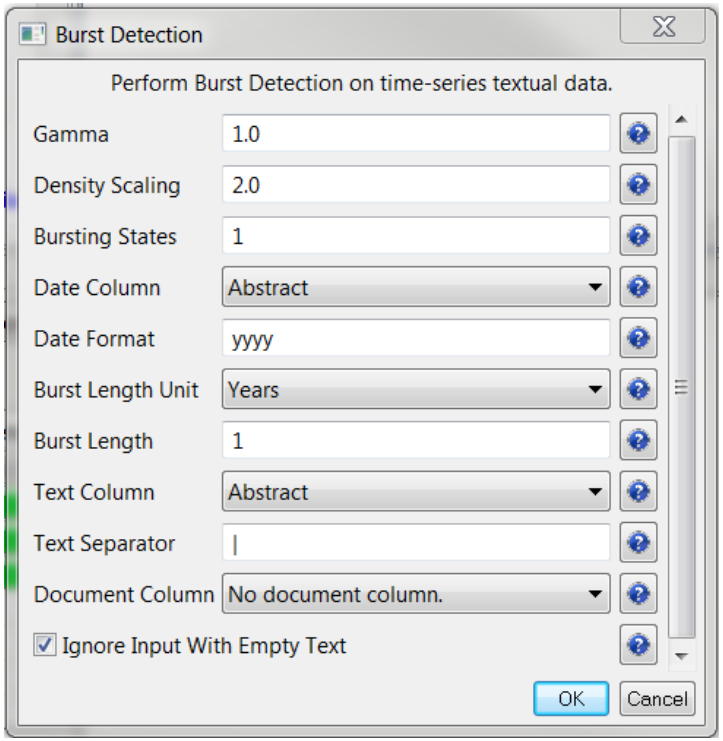
Introduction to Temporal Analysis

- 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.)

Temporal Analysis: Burst Detection

- 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.

Temporal Analysis: Burst Detection

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?

Tutorial Overview

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

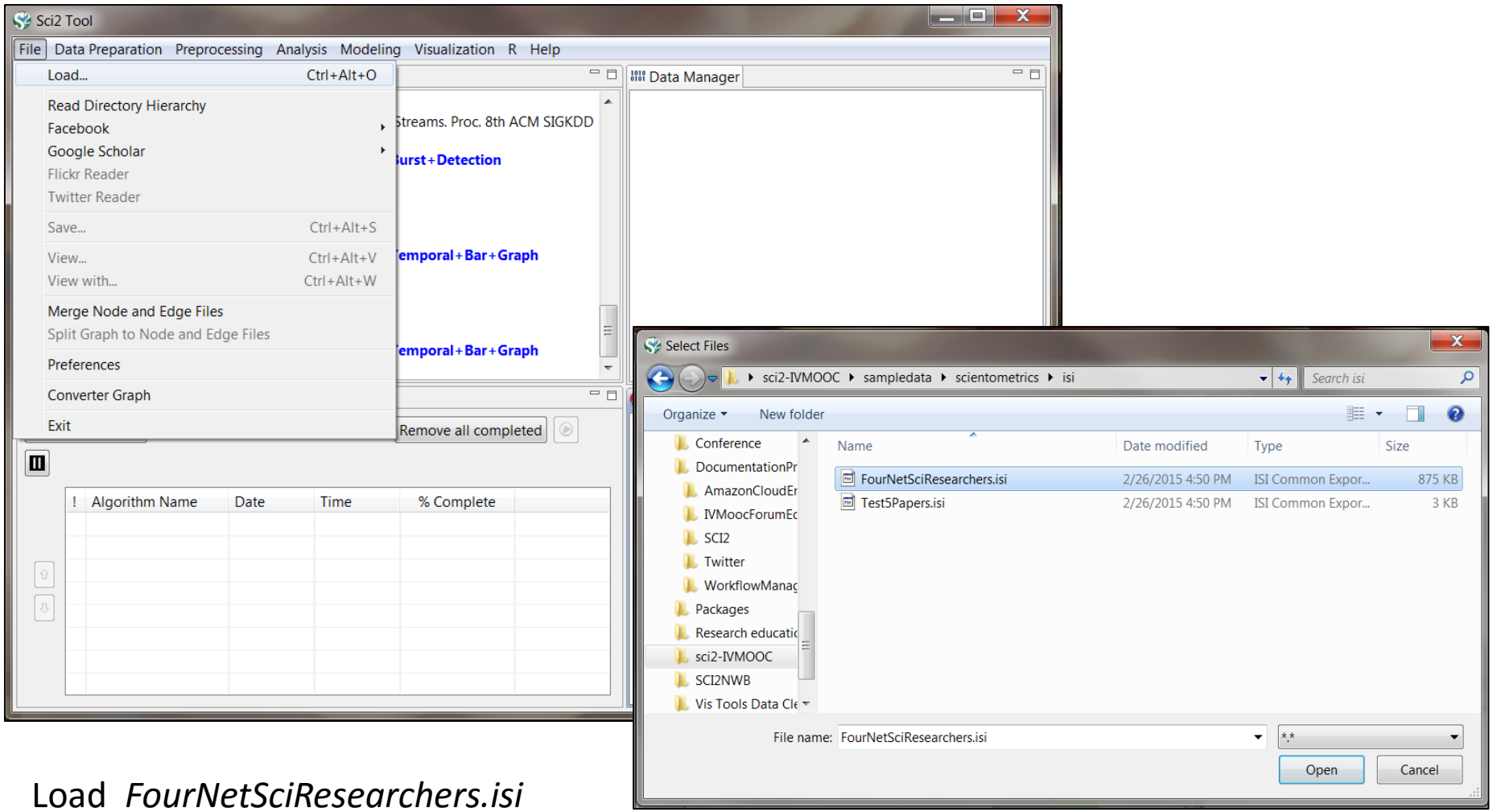
9:45 Geospatial Analysis
- Overview of geospatial analysis and mapping
- Geospatial Analysis: Geocoding with OpenRefine
- Geospatial Analysis: Proportional Symbol Map using CDC

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

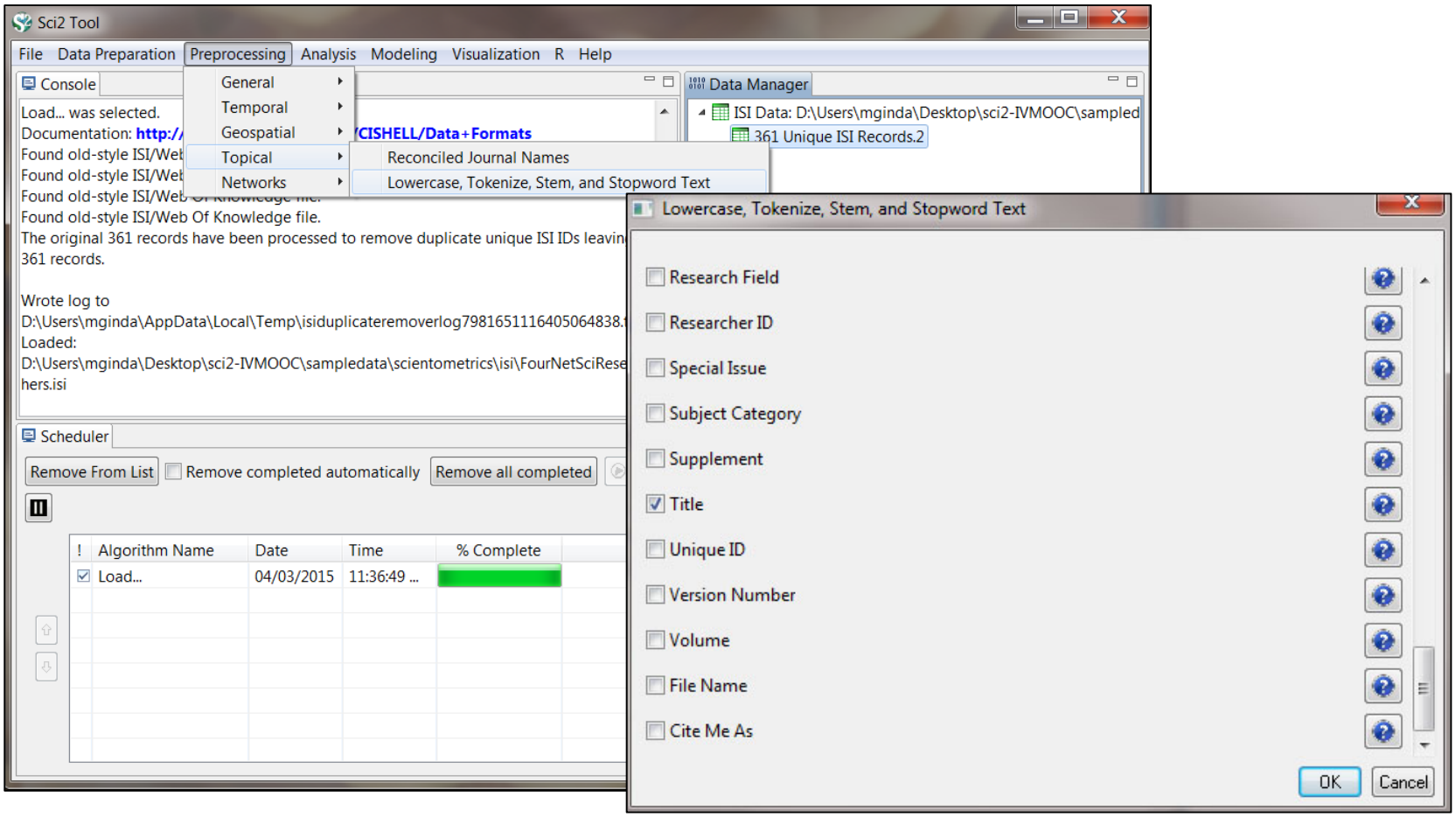
Temporal Analysis: Burst Detection



Load *FourNetSciResearchers.isi*

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

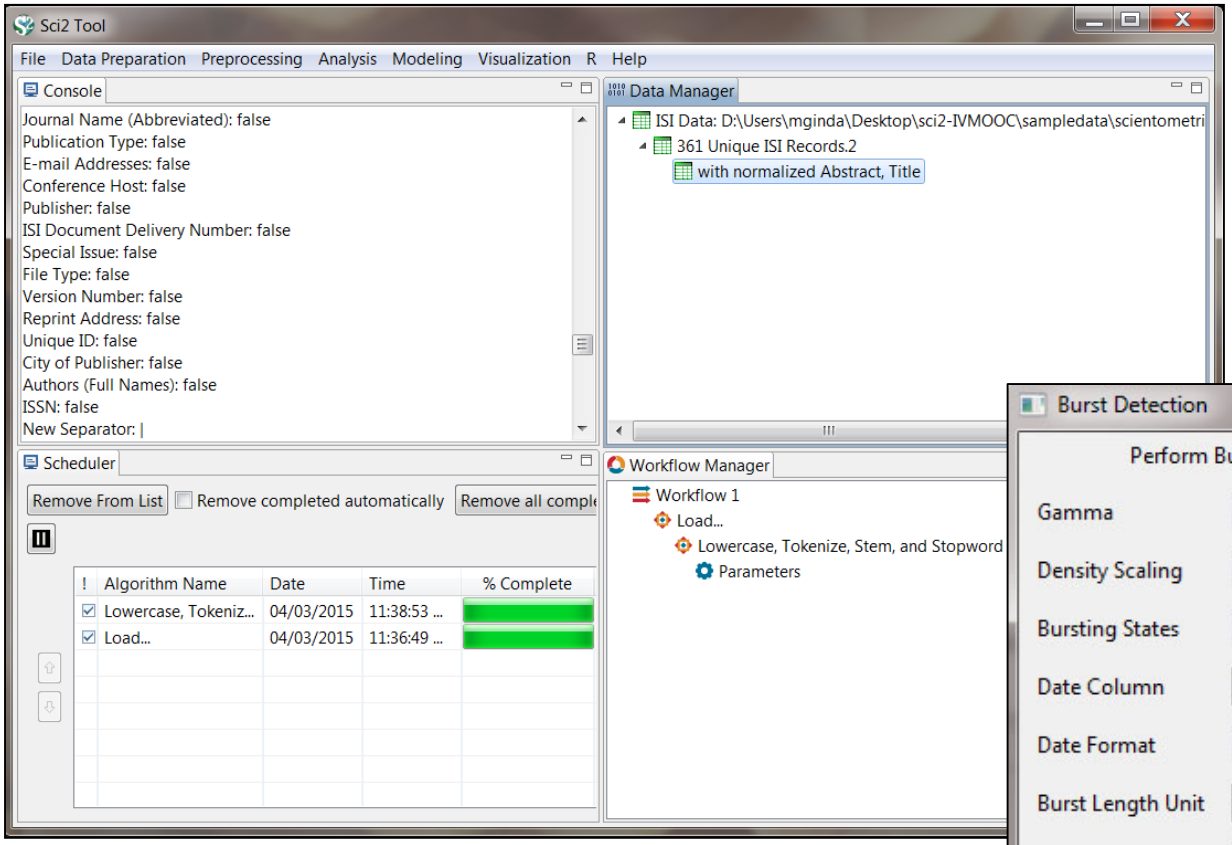
Temporal Analysis: Burst Detection

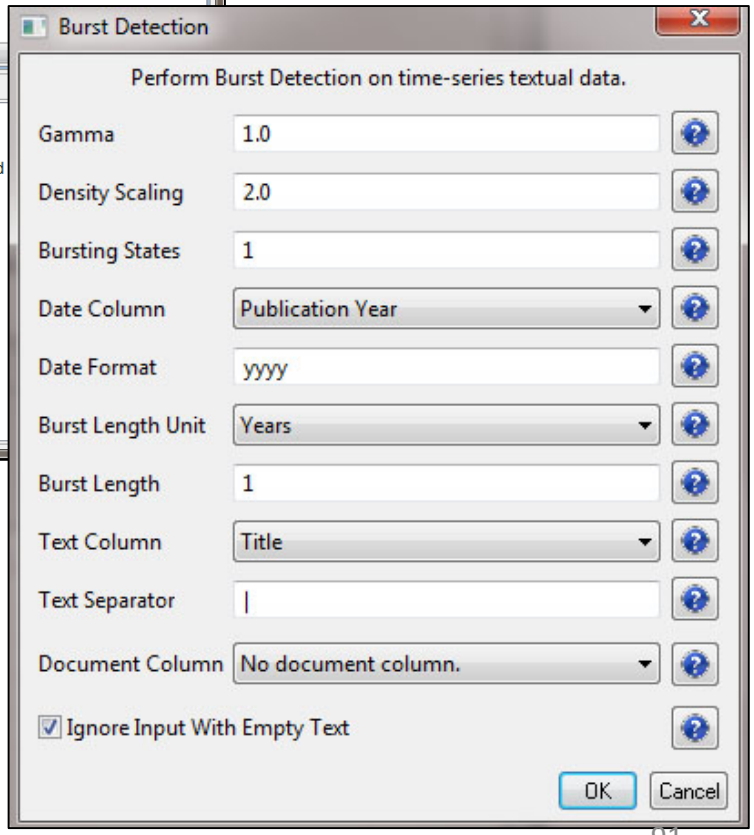


Select *Preprocessing > Topical > Lowercase, Tokenize, Stem, and Stopword Text*

Then select **Title** from the input parameters

Temporal Analysis: Burst Detection



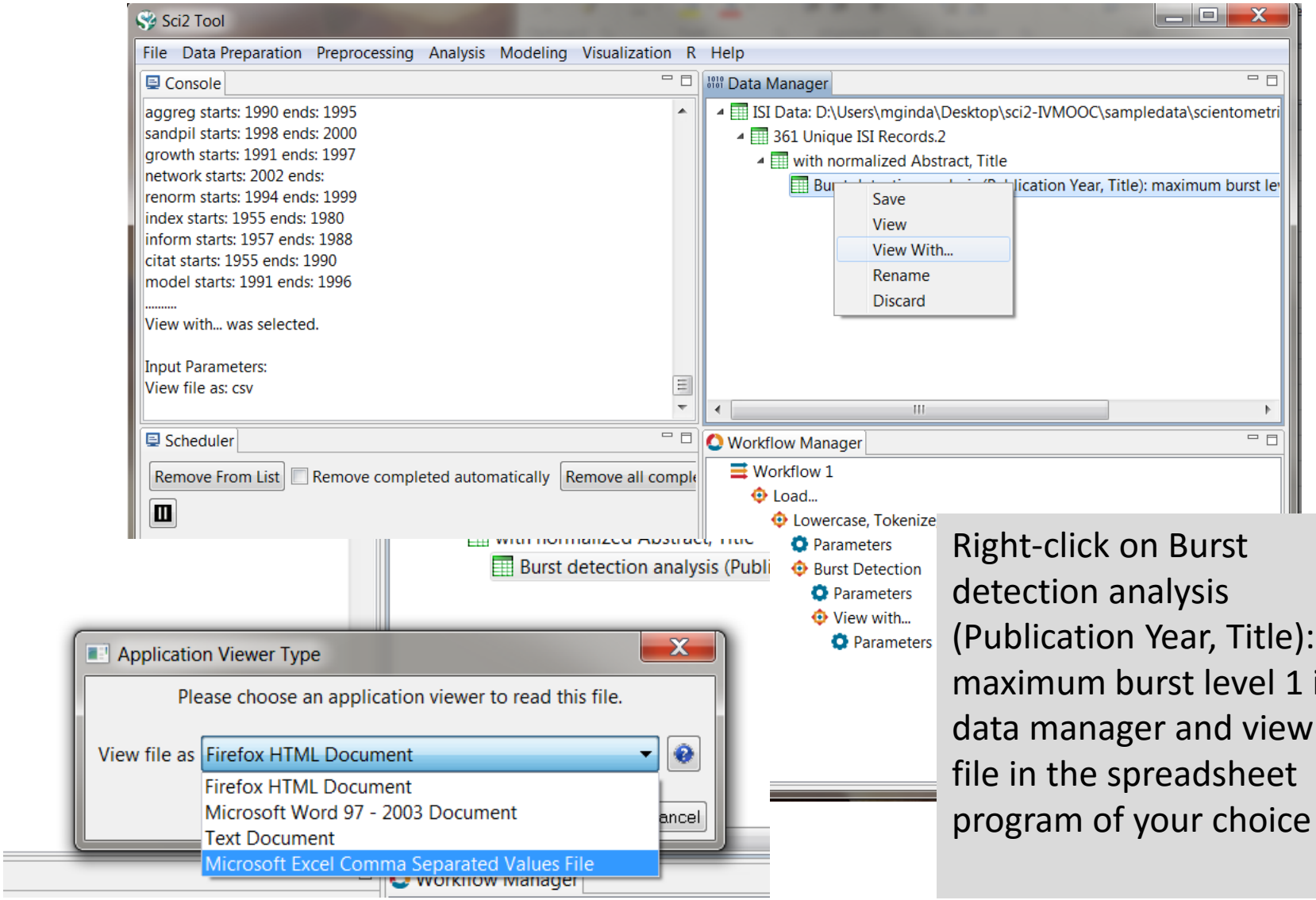


Highlight the table 'with normalized Title'

Select *Analysis > Temporal > Burst Detection*

Then set the parameters to what is shown to the right

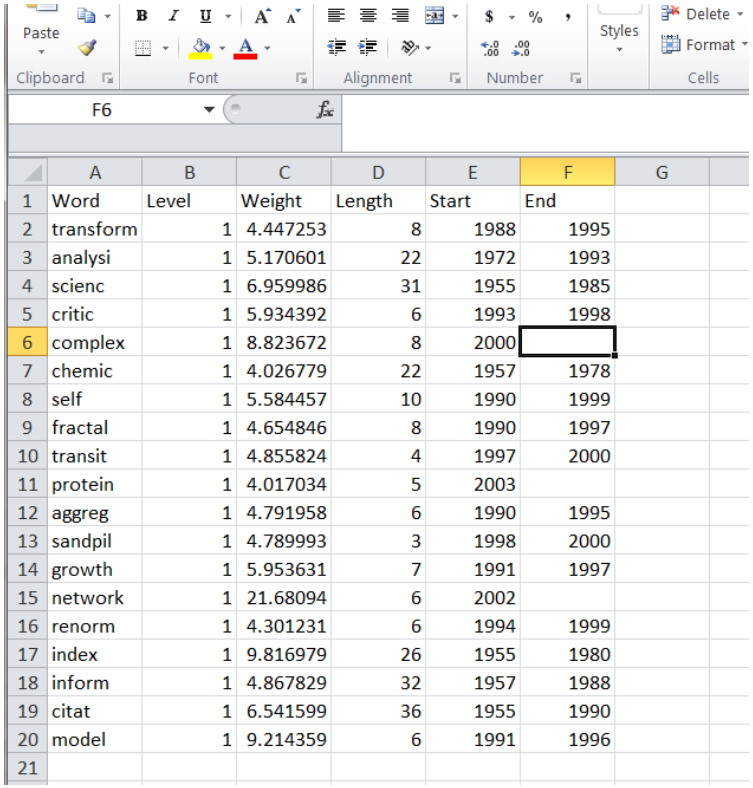
Temporal Analysis: Burst Detection



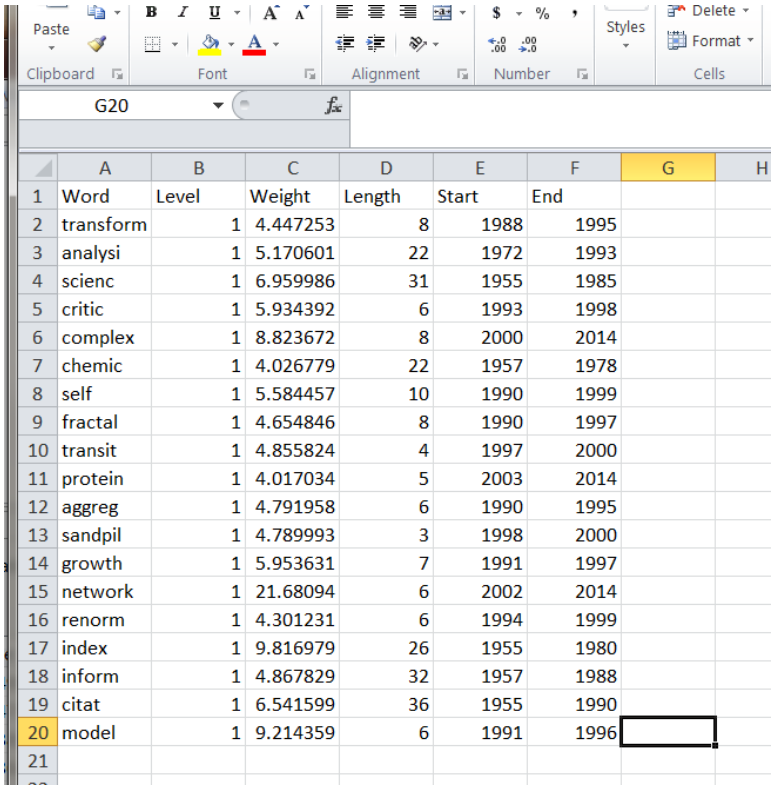
Right-click on Burst detection analysis (Publication Year, Title): maximum burst level 1 in the data manager and view the file in the spreadsheet program of your choice

Temporal Analysis: Burst Detection

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							

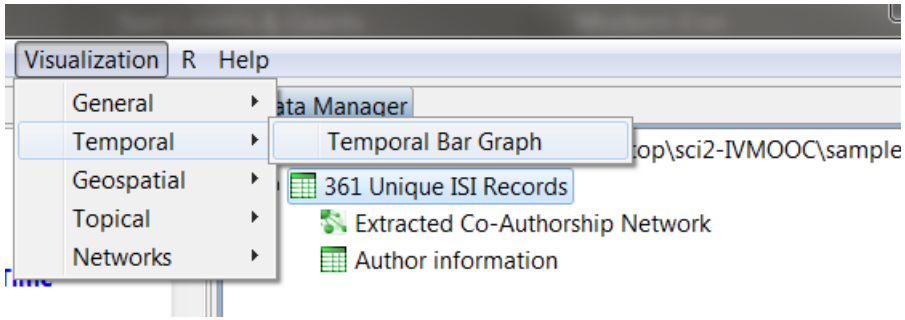


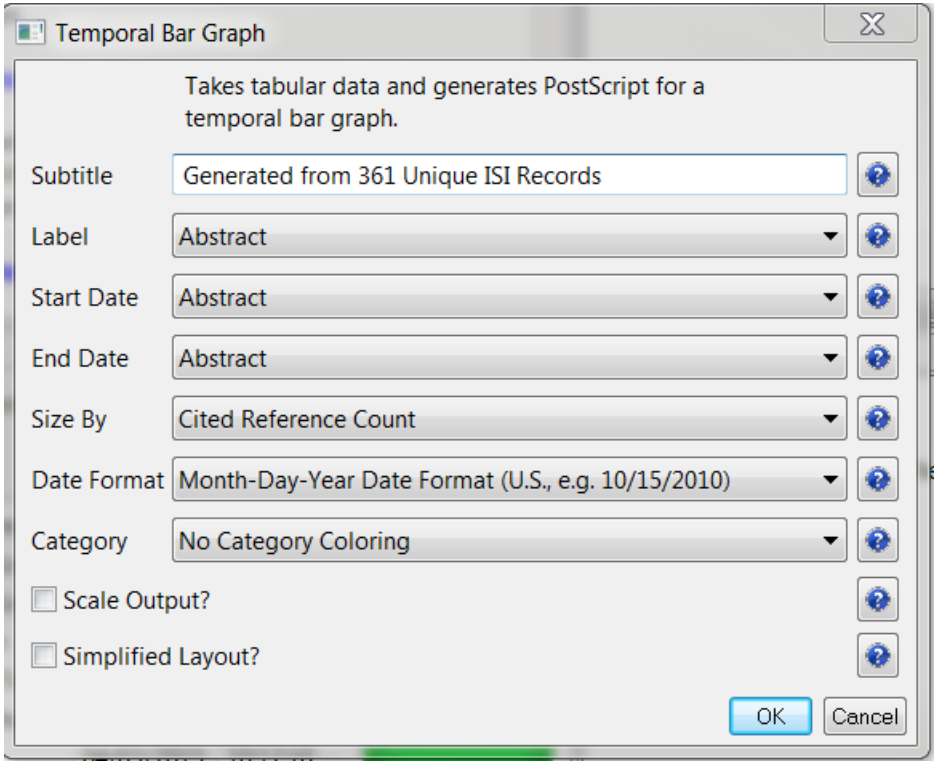
	A	B	C	D	E	F	G	H
1	Word	Level	Weight	Length	Start	End		
2	transform	1	4.447253	8	1988	1995		
3	analysi	1	5.170601	22	1972	1993		
4	scienc	1	6.959986	31	1955	1985		
5	critic	1	5.934392	6	1993	1998		
6	complex	1	8.823672	8	2000	2014		
7	chemic	1	4.026779	22	1957	1978		
8	self	1	5.584457	10	1990	1999		
9	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

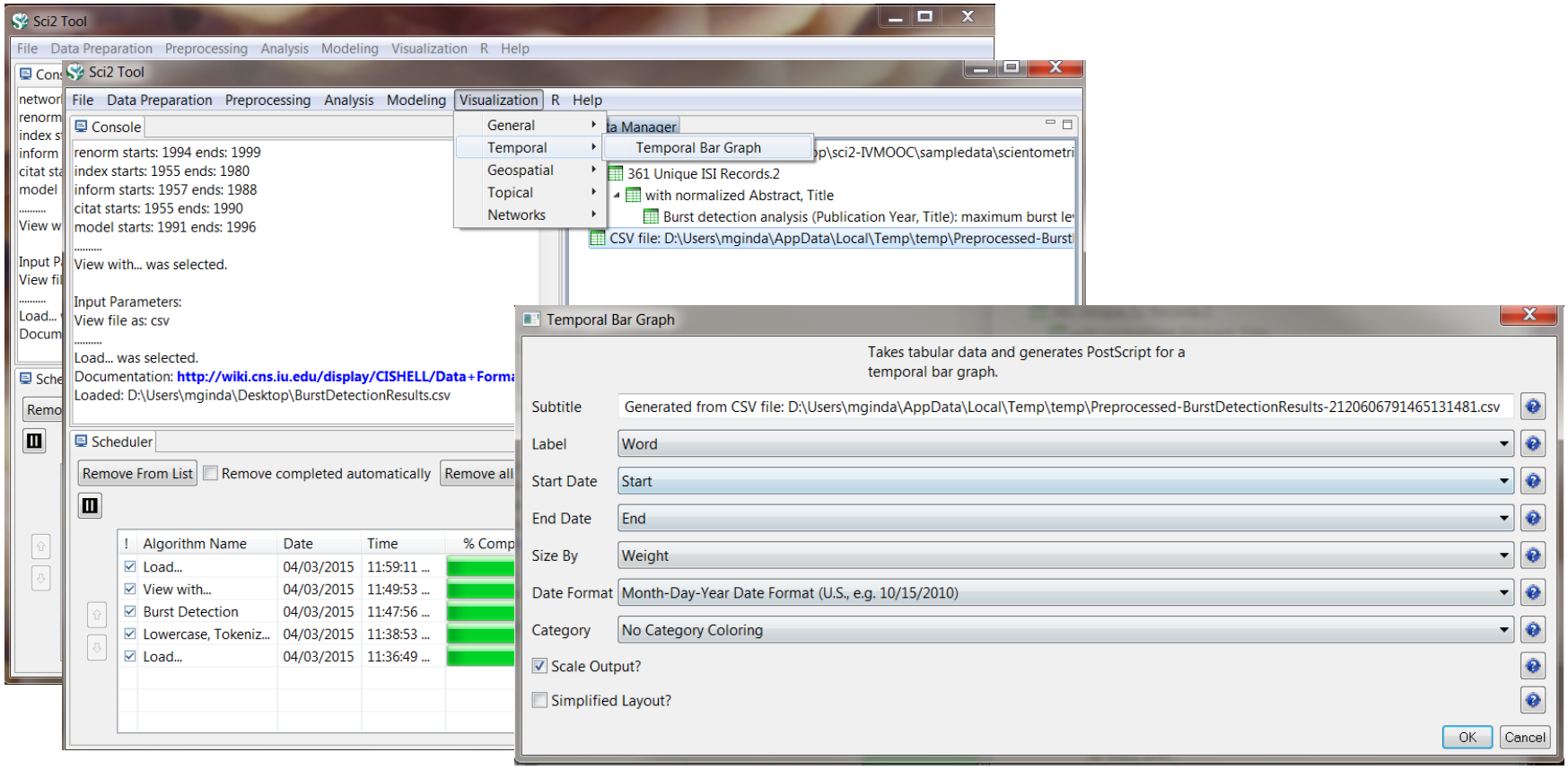
Temporal Analysis: Burst Detection

- 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.





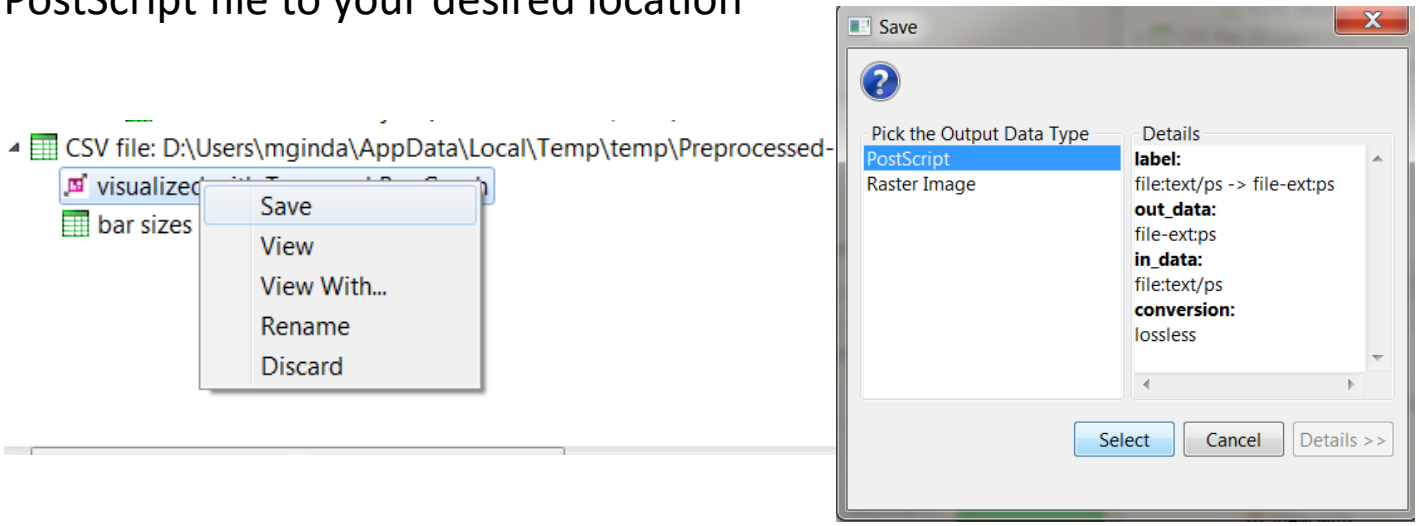
Temporal Analysis: Burst Detection



- 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

Temporal Analysis: Burst Detection

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



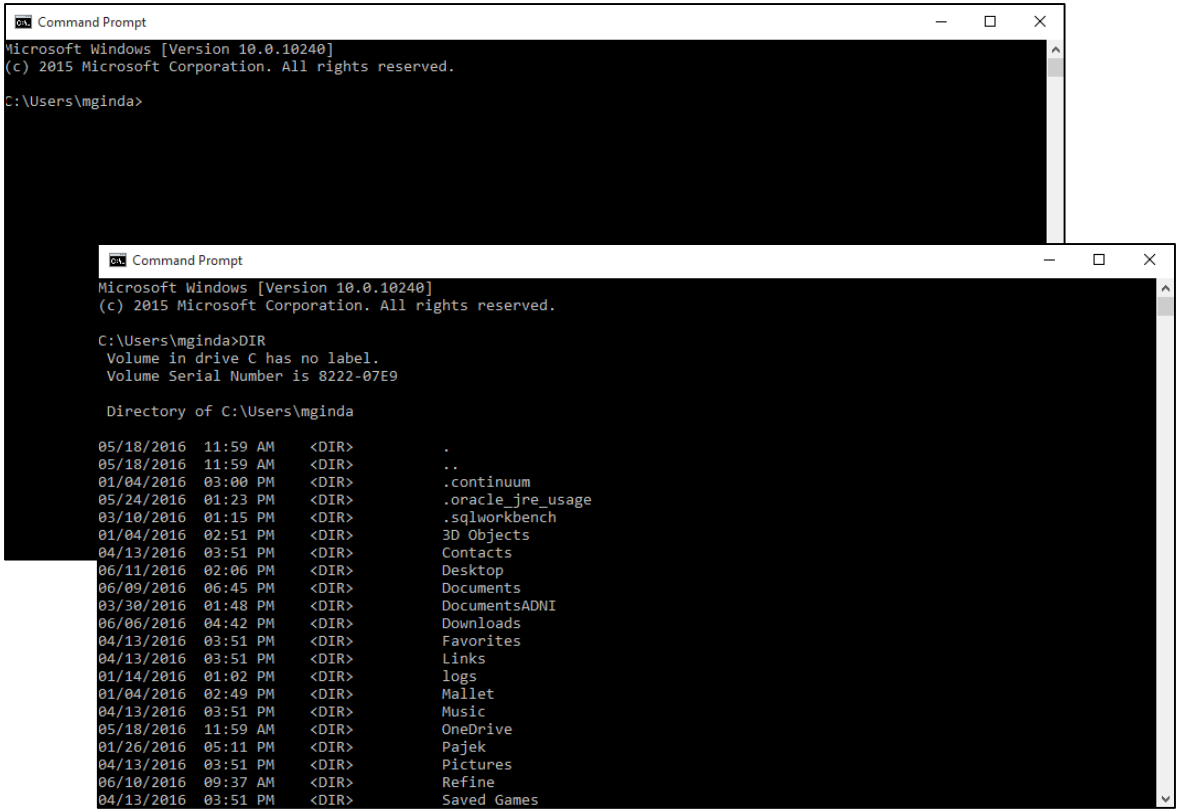
To convert the PS file output to a PDF, we will use GhostScripts command line tool, PS2PDF. Open Windows' Command Prompt. Next

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

Sci2: Converting the Output PostScript Bursts to PDF

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.



Sci2: Converting the Output PostScript Bursts to PDF

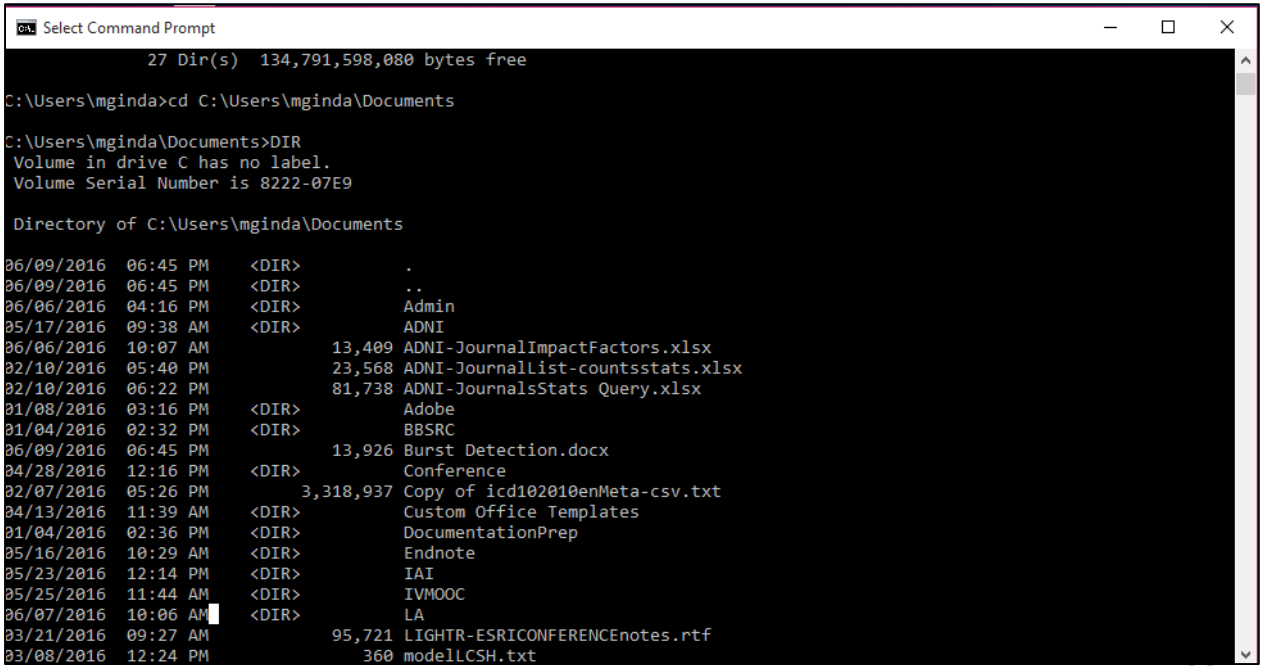
A common place to save a file may be the Desktop or the Documents folder.

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

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

or

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

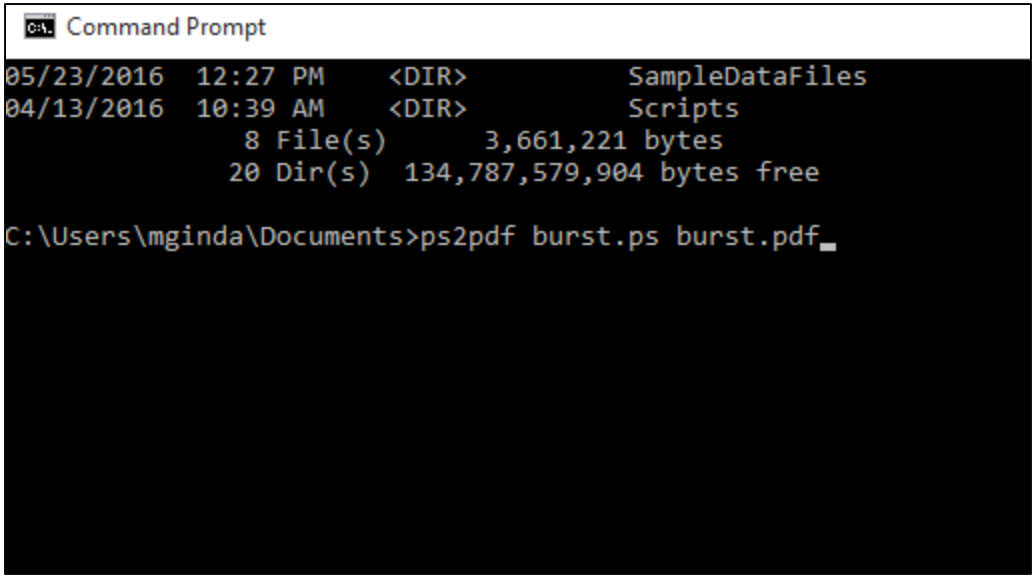


Sci2: Converting the Output PostScript Bursts to PDF

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
```

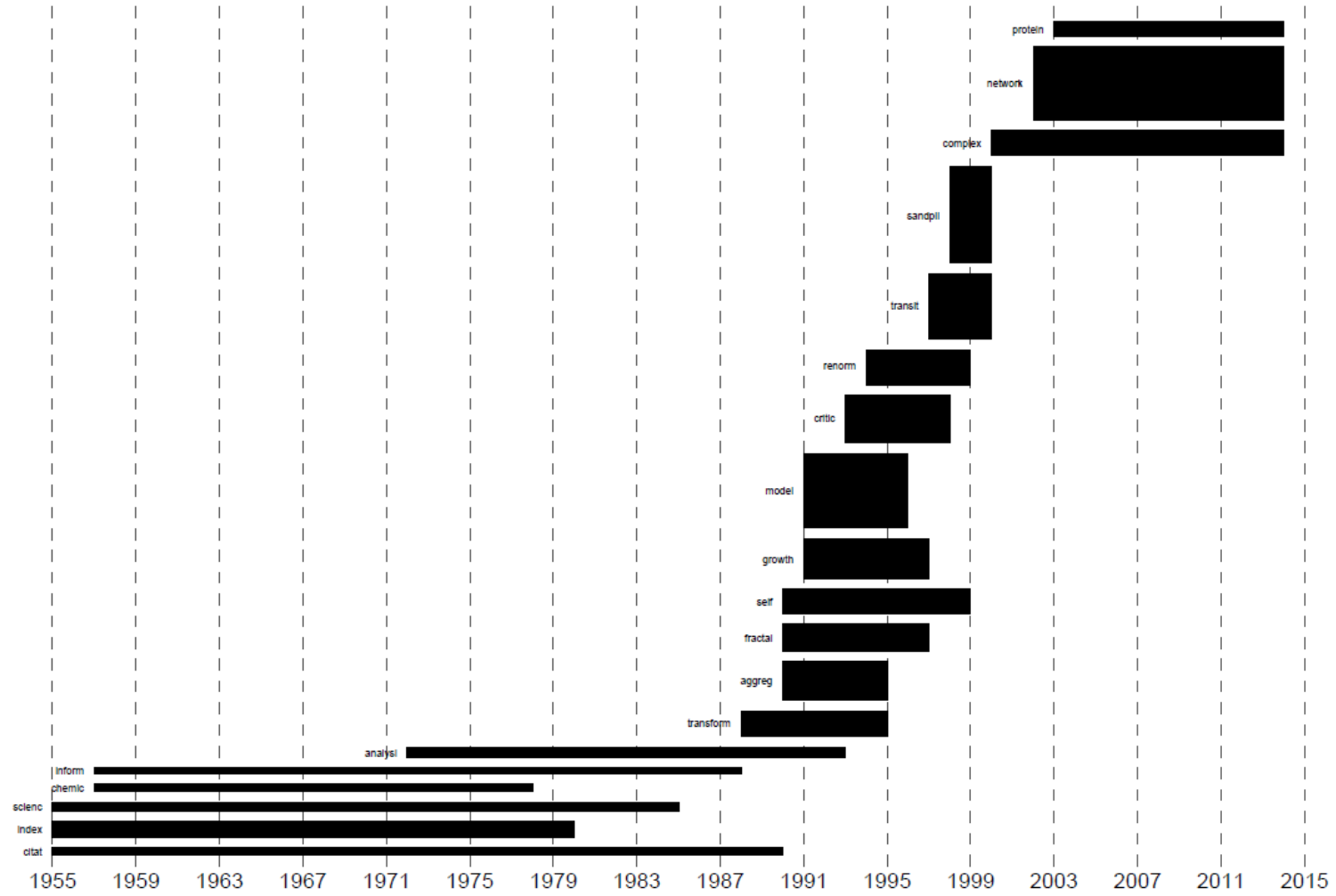
In you leave off the output file name, the resulting PDF file will have the same name as the original file with a .PDF extension into the same directory you navigated to.



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

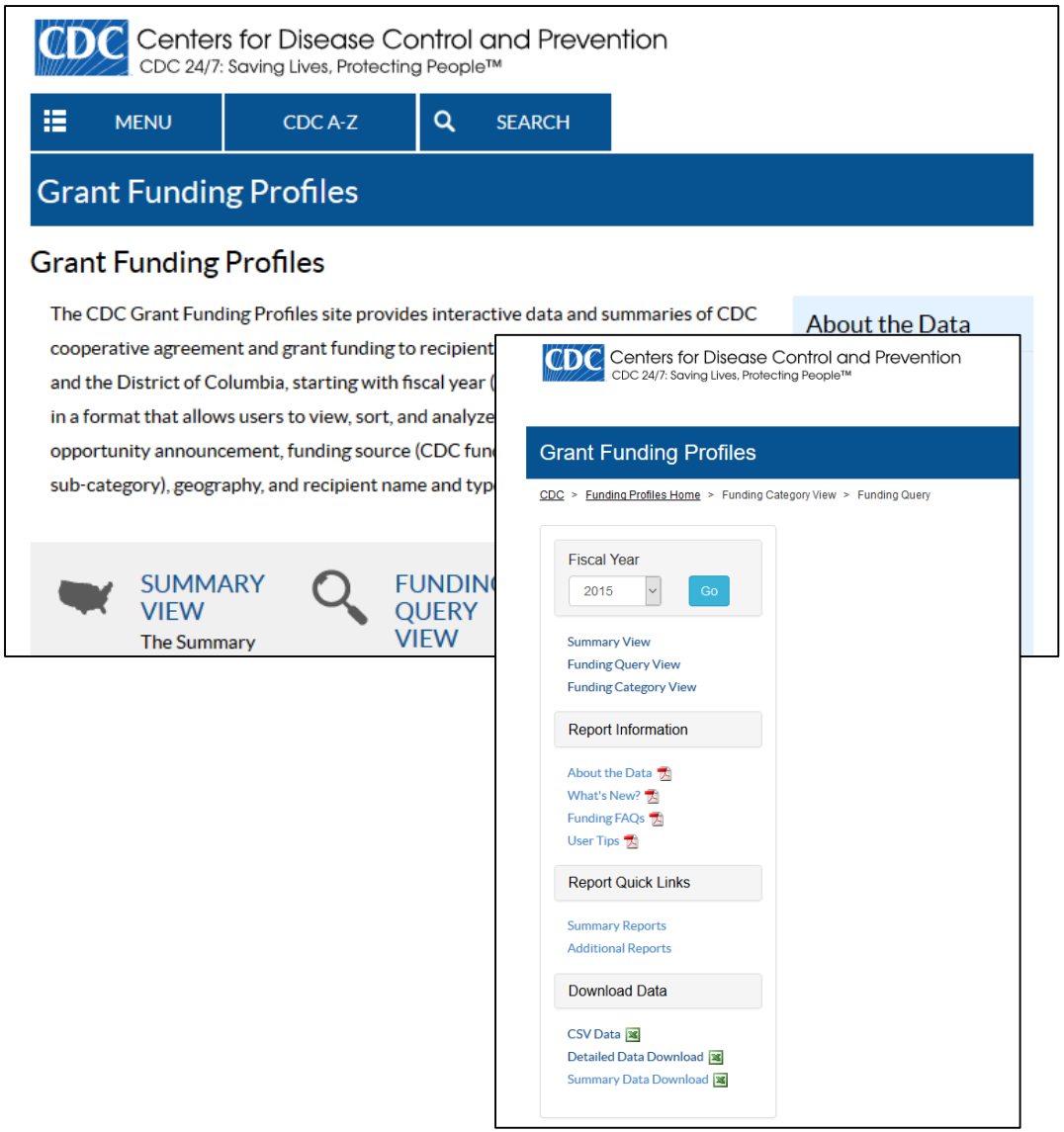
Area

5
2
1
0
5.21 Year(s)

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.

Burst Detection: CDC Grants Dataset

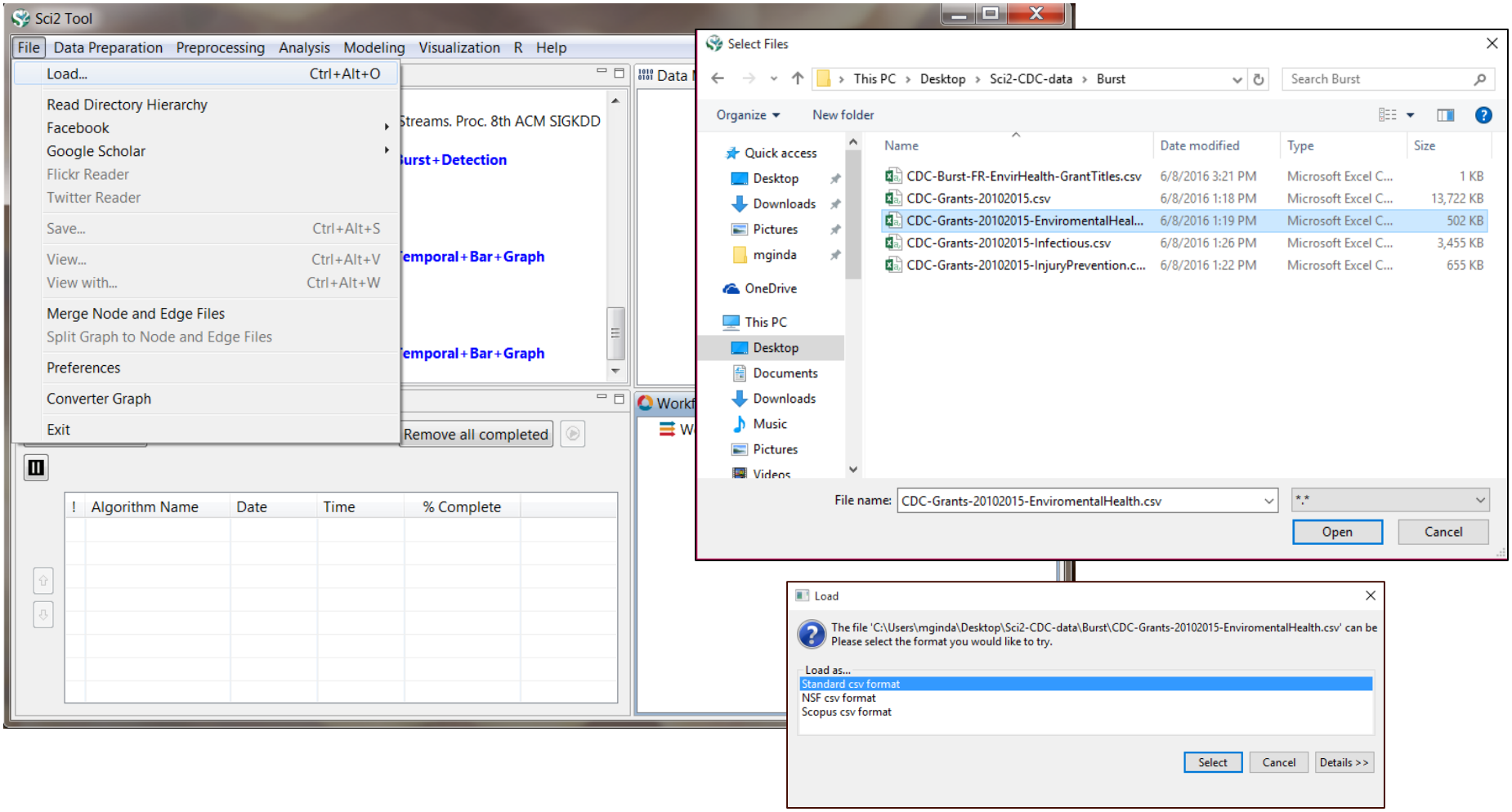


CDC Grant funding data was collected for each fiscal year between 2010-2015 from the Grant Funding Profiles page for CDC.

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.

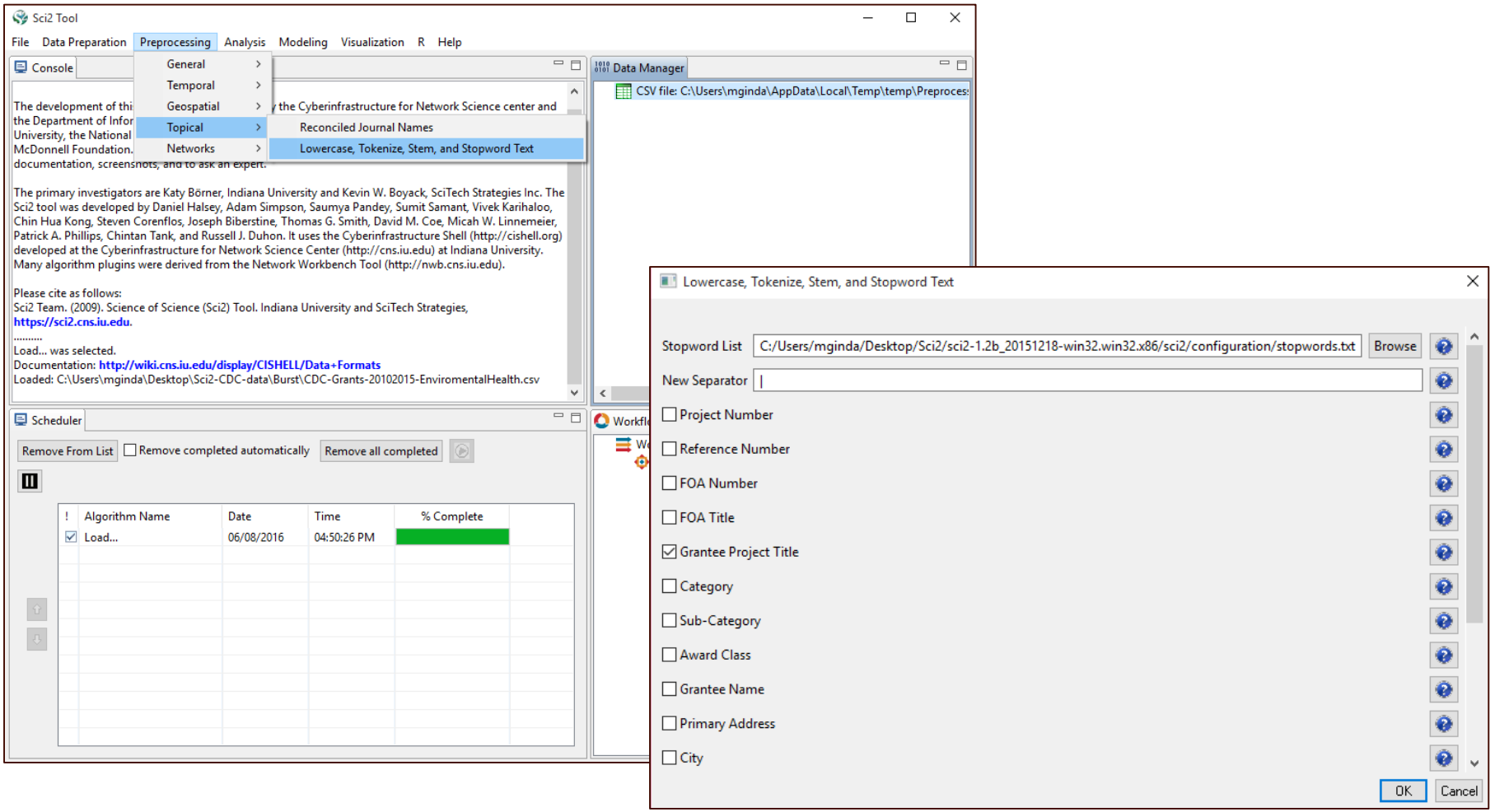
This data is available to download as in the CSV format.

Burst Detection: CDC Grants Dataset



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

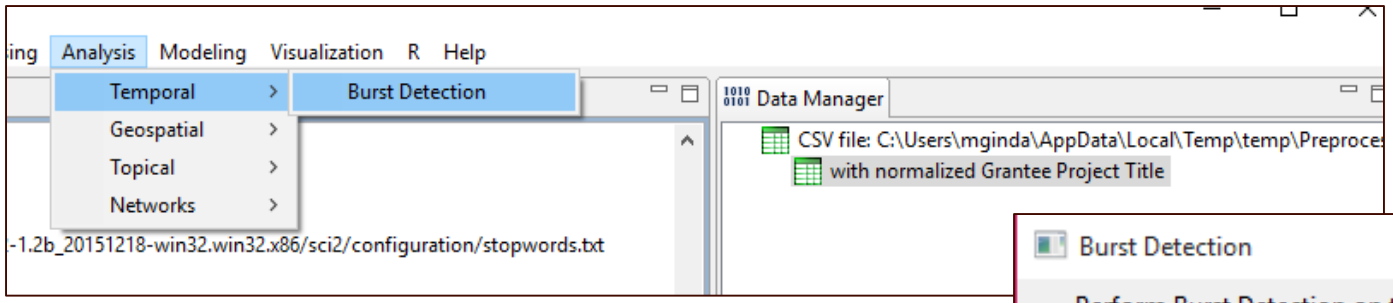
Burst Detection: CDC Grants Dataset

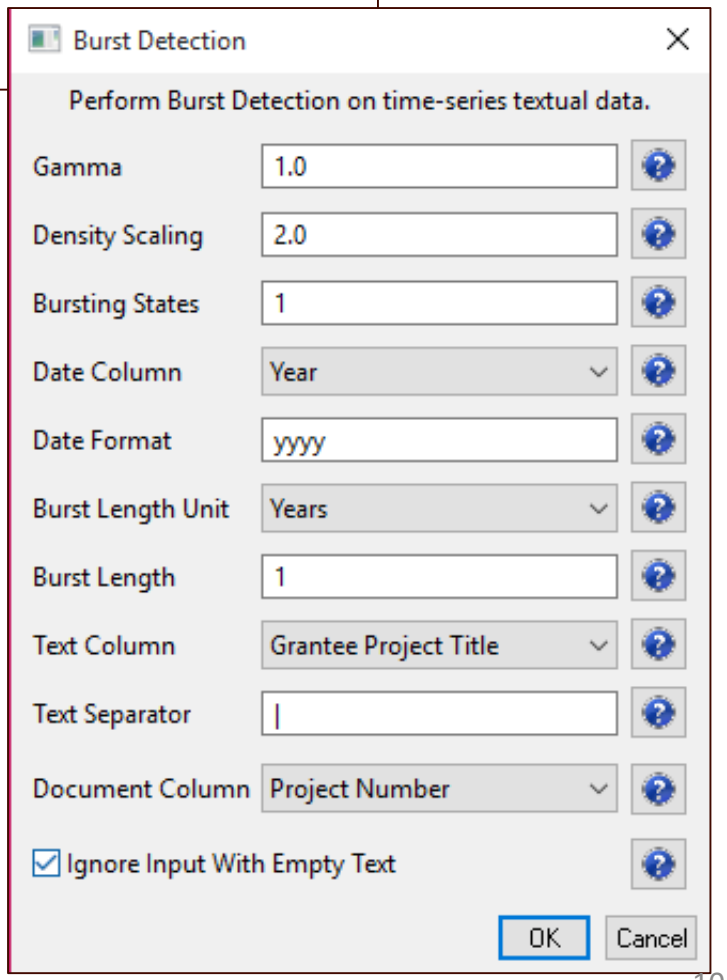


Select the loaded data file in the Data Manager and then navigate to *Preprocessing > Topical > Lowercase, Tokenize, Stem, and Stopword Text*

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

Burst Detection: CDC Grants Dataset





Highlight the table 'with normalized Grantee Project Title'

Select *Analysis > Temporal > Burst Detection*

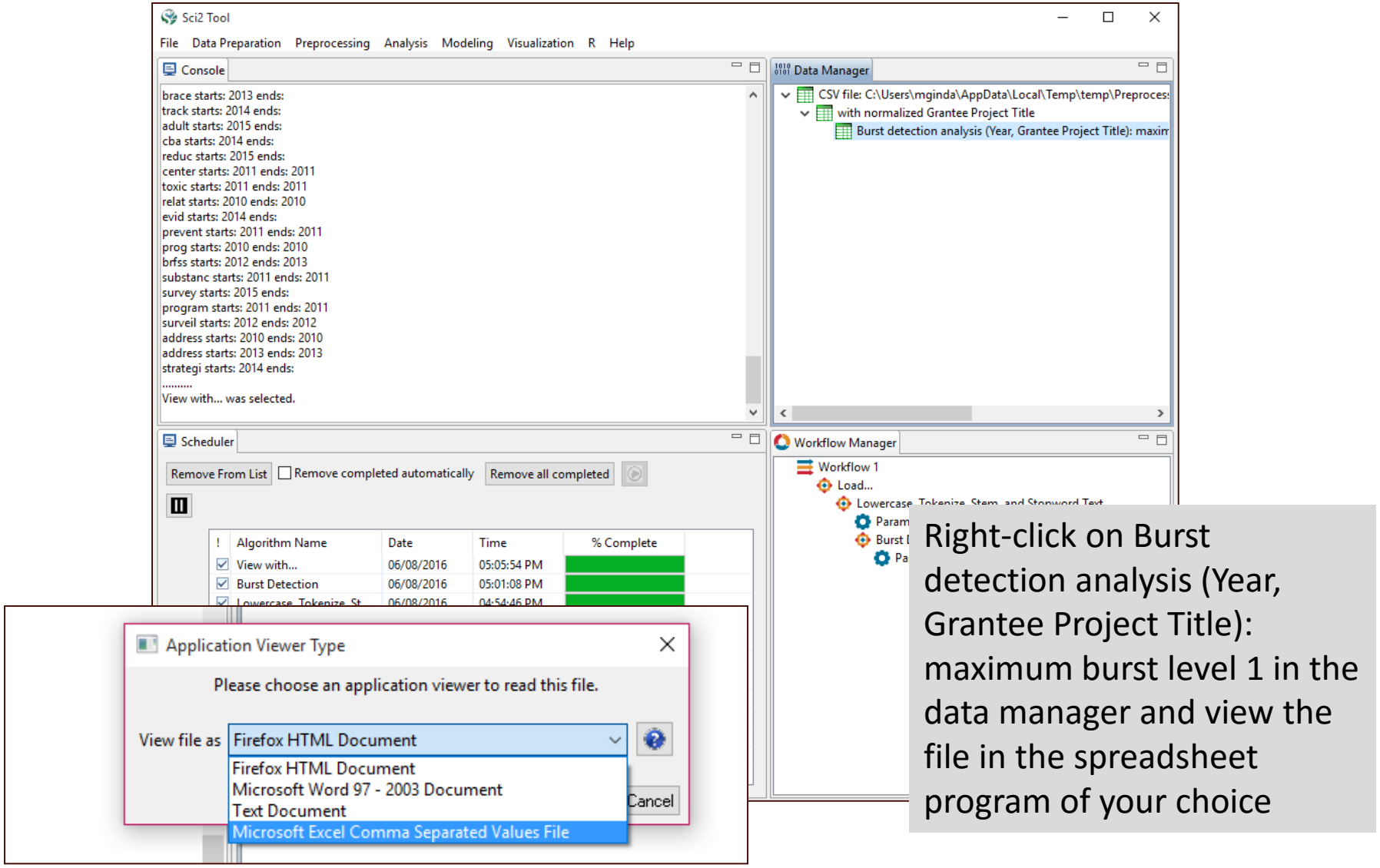
For our initial burst analysis we will use the base **Gamma** and **Density Scaling** Parameters.

For the **Date Dolumn** select "Year"

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

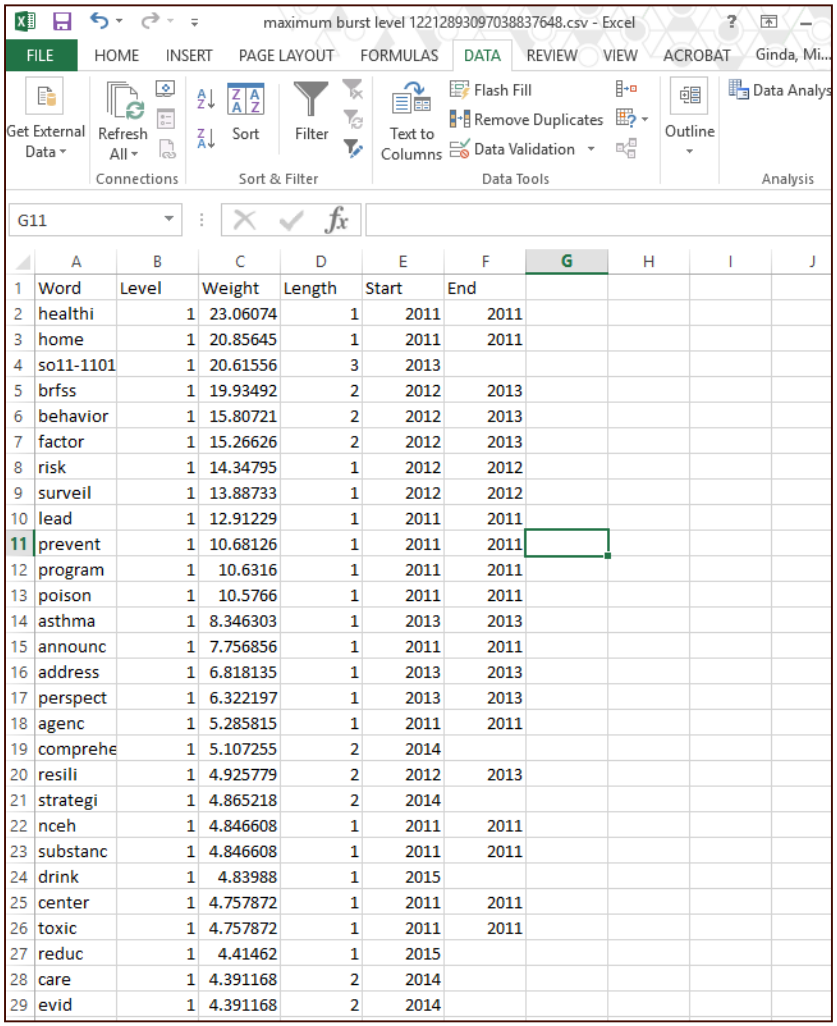
For **Document Column** select "Project Number"

Burst Detection: CDC Grants Dataset



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

Burst Detection: CDC Grants Dataset



	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					

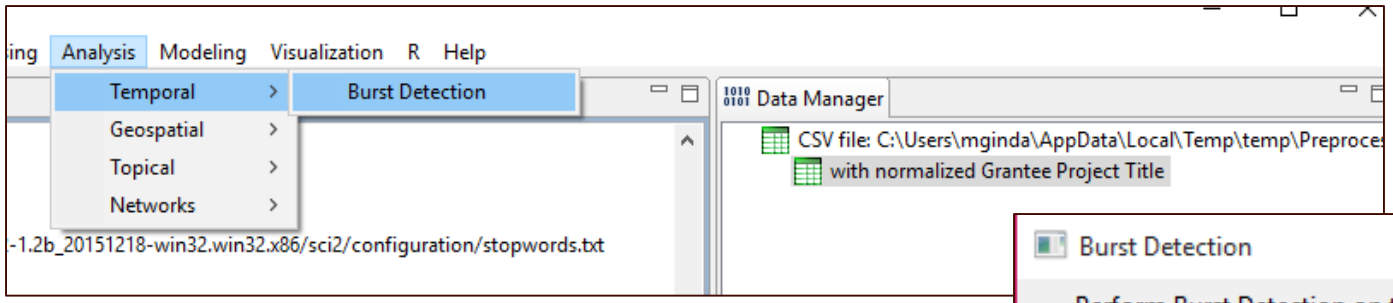
As before, the missing end dates indicate the continuation of a burst in a given data set.

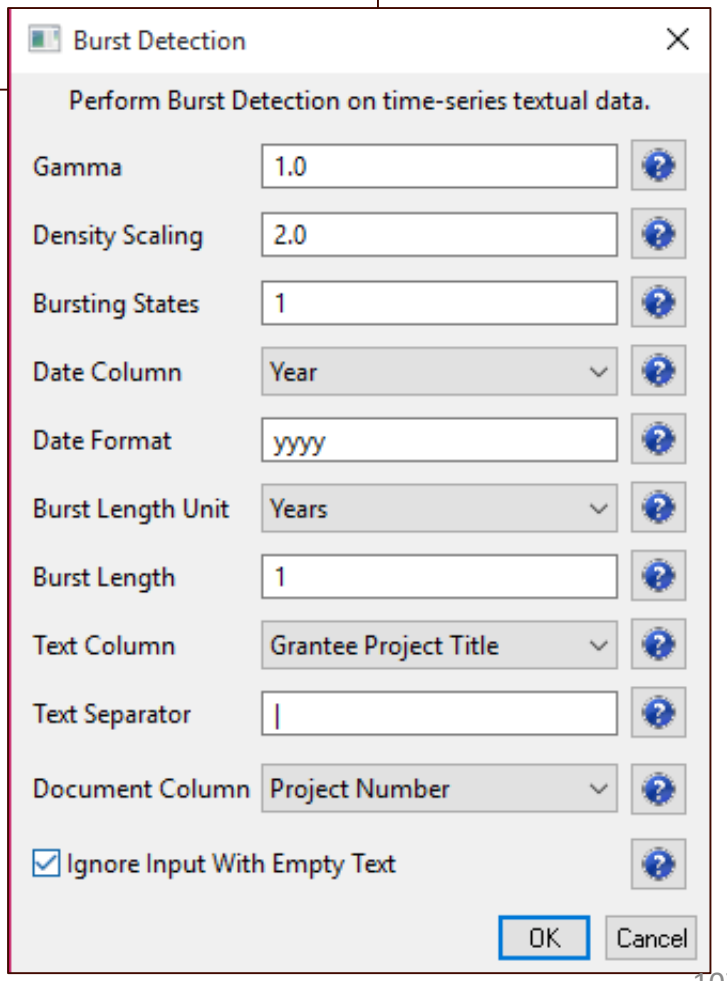
Before we add the End date of 2015 to those records missing a value, lets review the results.

What are the primary bursting terms?

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

Burst Detection: CDC Grants Dataset





Highlight the table 'with normalized Grantee Project Title'

Select *Analysis > Temporal > Burst Detection*

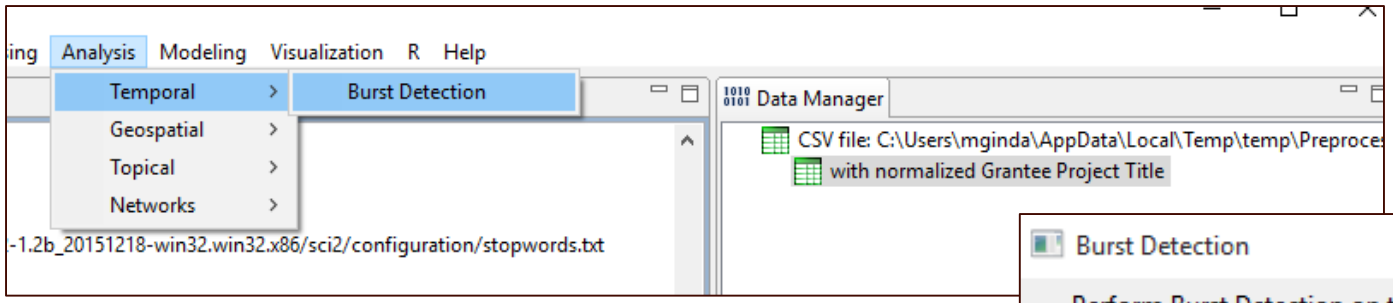
For our initial burst analysis we will use the base **Gamma** and **Density Scaling** Parameters.

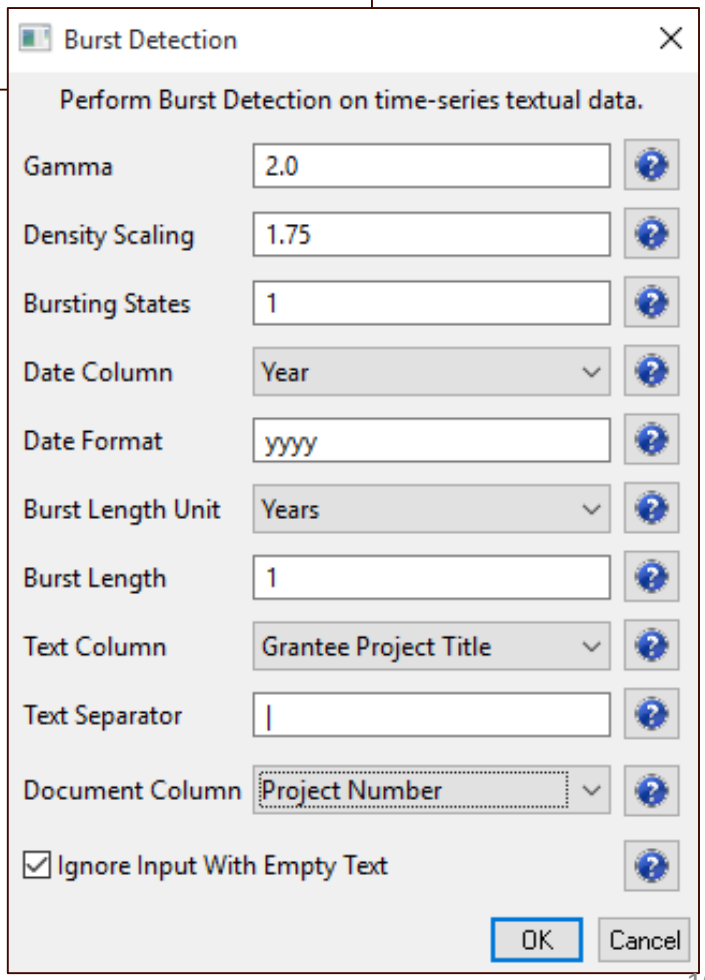
For the **Date Dolumn** select "Year"

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

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

Burst Detection: CDC Grants Dataset





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 Dolumn** select "Year"

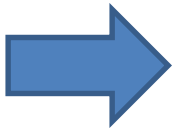
For **Text Column** select "Grant Project Title"

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

Temporal Analysis: Burst Detection

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.

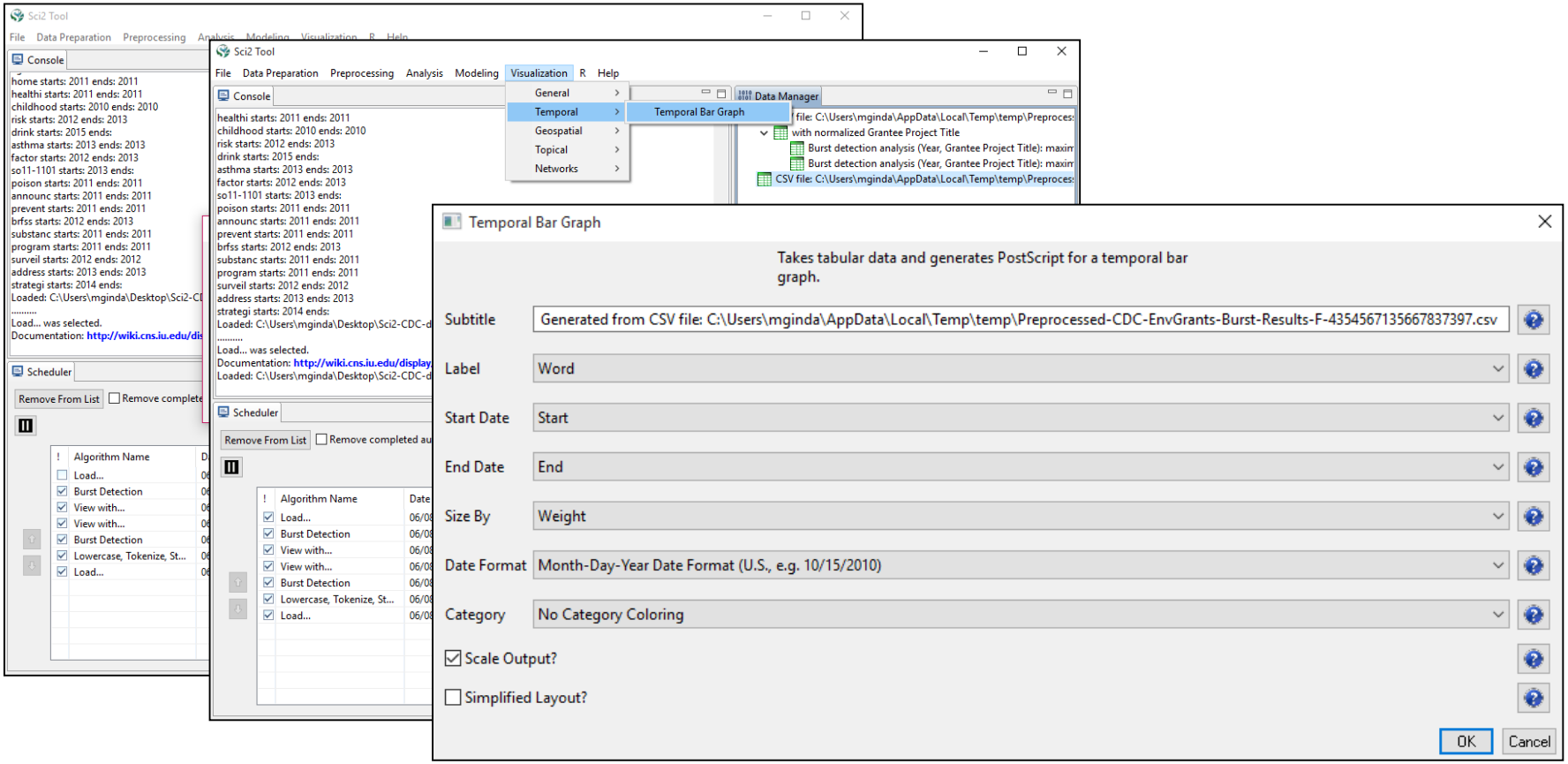
	A	B	C	D	E	F
1	Word	Level	Weight	Length	Start	End
2	perspect	1	5.43949	1	2013	2013
3	behavior	1	15.12261	2	2012	2013
4	resili	1	4.874477	4	2012	
5	comprehe	1	4.397597	2	2014	
6	lead	1	11.27075	1	2011	2011
7	nceh	1	3.953627	1	2011	2011
8	agenc	1	4.323344	1	2011	2011
9	home	1	17.00882	1	2011	2011
10	healthi	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	



	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	nceh	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

Burst Detection: CDC Grants Dataset

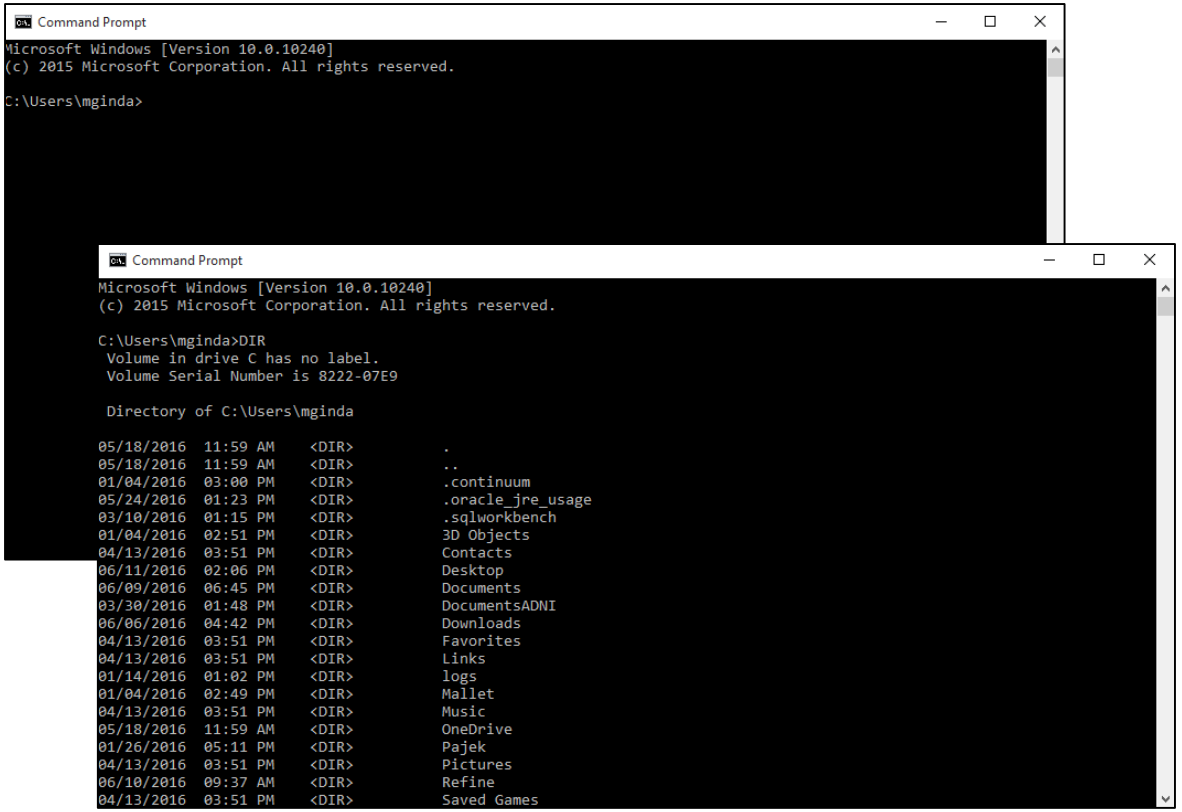


- Load updated Burst Detection result file into Sci2 as a standard CSV format.
- Select the newly loaded file in the data manager and the*Visualization > Temporal > Temporal Bar Graph* in the menu bar.
- Set the parameter values to those shown to the right

Sci2: Converting the Output PostScript Bursts to PDF

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.



Sci2: Converting the Output PostScript Bursts to PDF

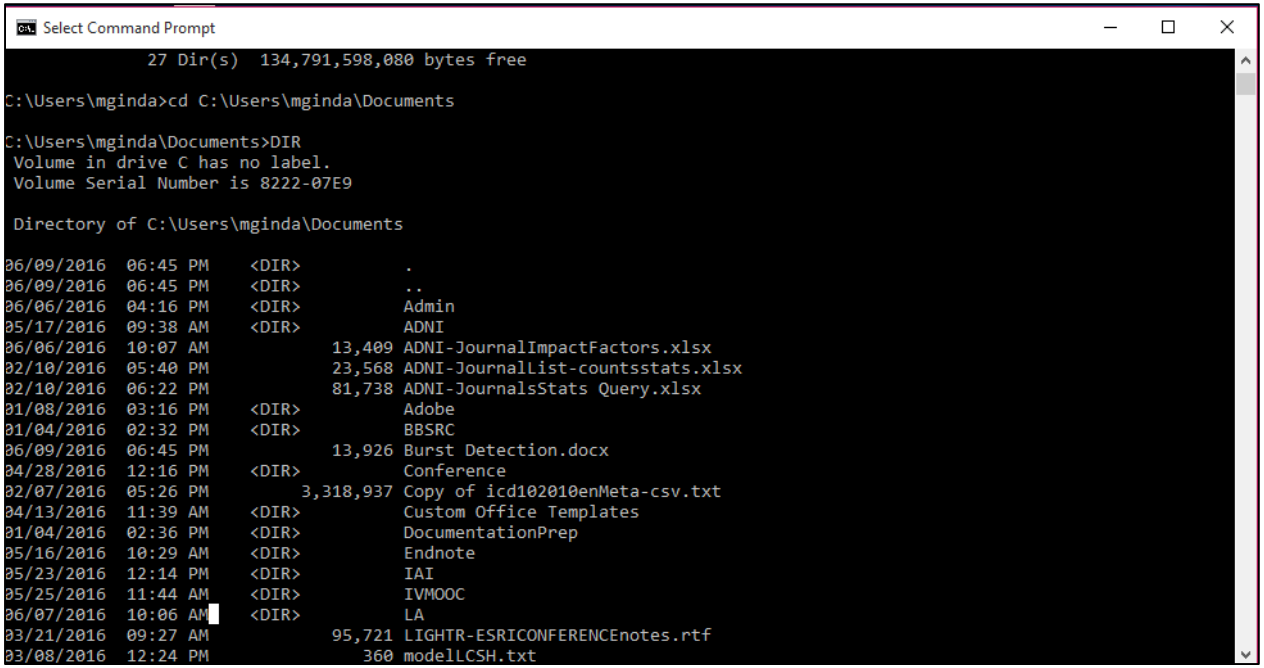
A common place to save a file may be the Desktop or the Documents folder.

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

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

or

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

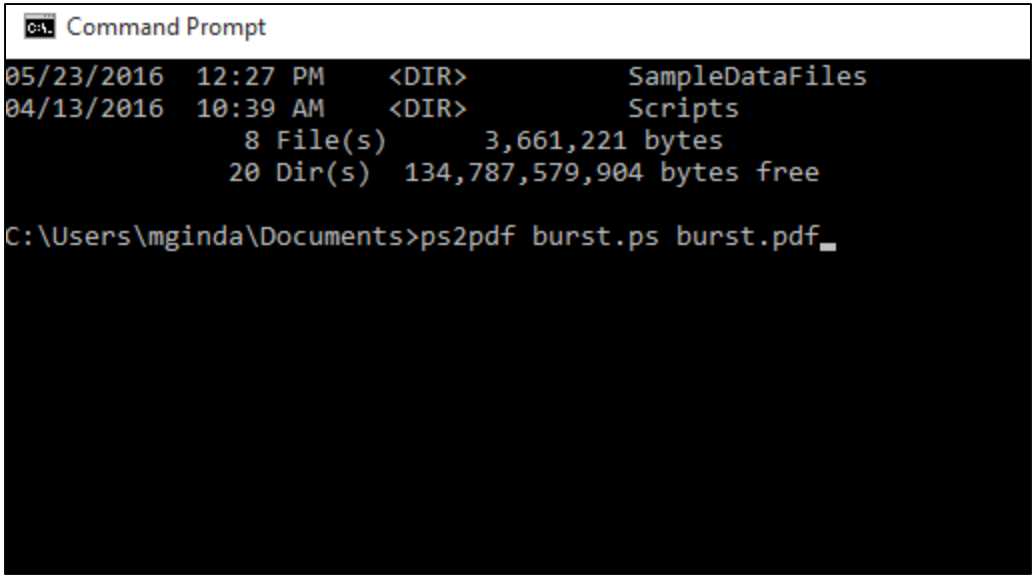


Sci2: Converting the Output PostScript Bursts to PDF

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
```

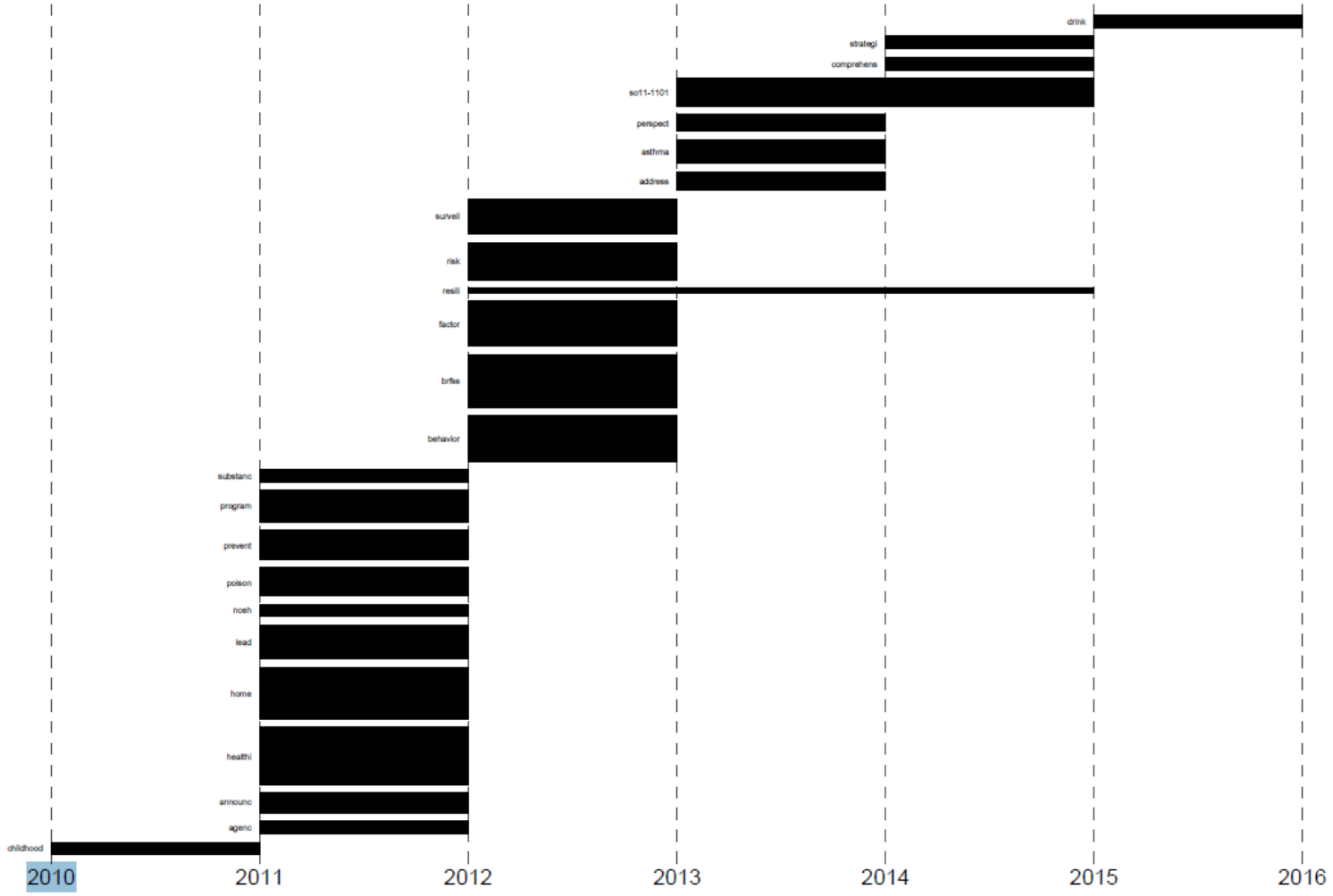
In you leave off the output file name, the resulting PDF file will have the same name as the original file with a .PDF extension into the same directory you navigated to.



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

Area

7
2
1
0
0.52 Year(s)

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?

Tutorial Overview

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?

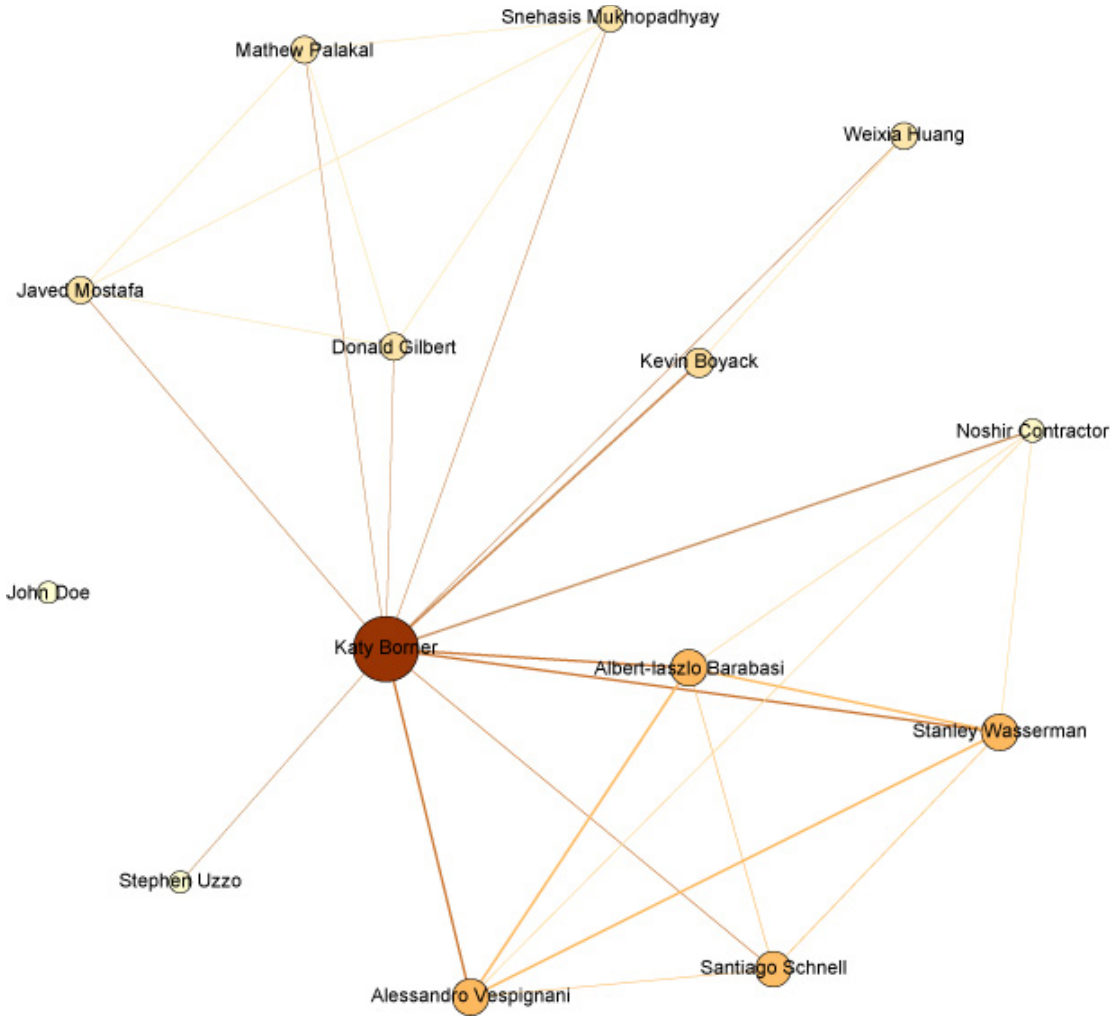
- Graph – network visualized
- Nodes
- Edges
- Components

Representations

- Matrices
- Graphs
- Edge and Node Lists

Data Formats

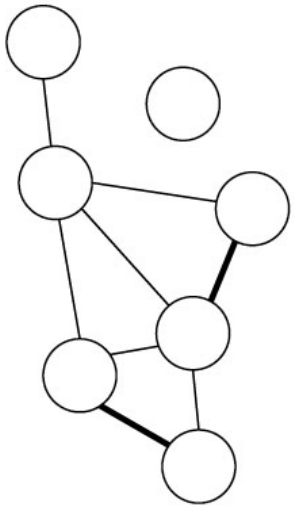
- Tabular
- XML
- Text
- JSON



Introduction to Network Analysis

General types of networks

Undirected Networks



Nodes:

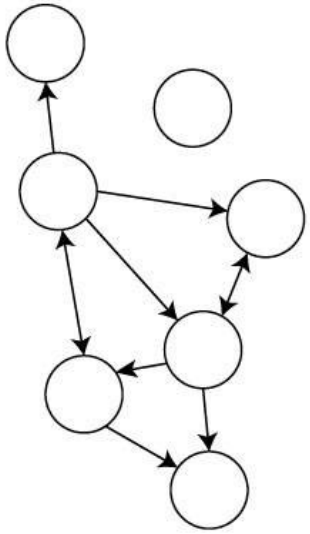
Edges:

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

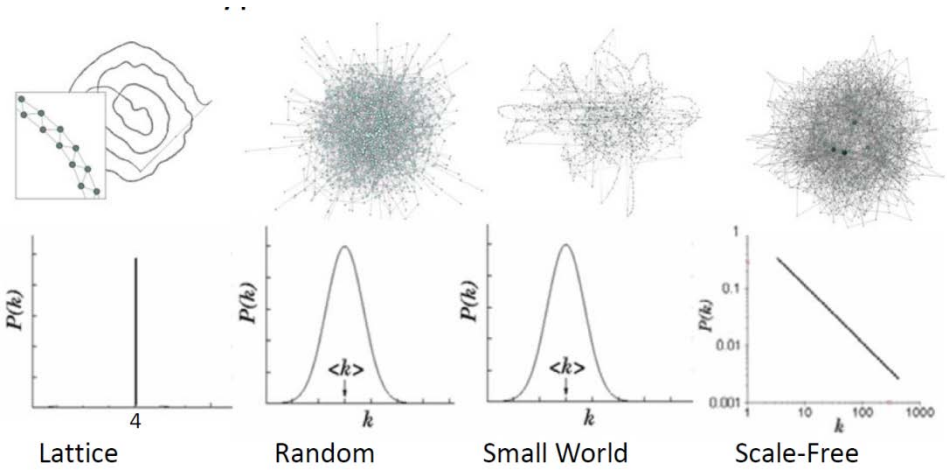
Graph Features

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

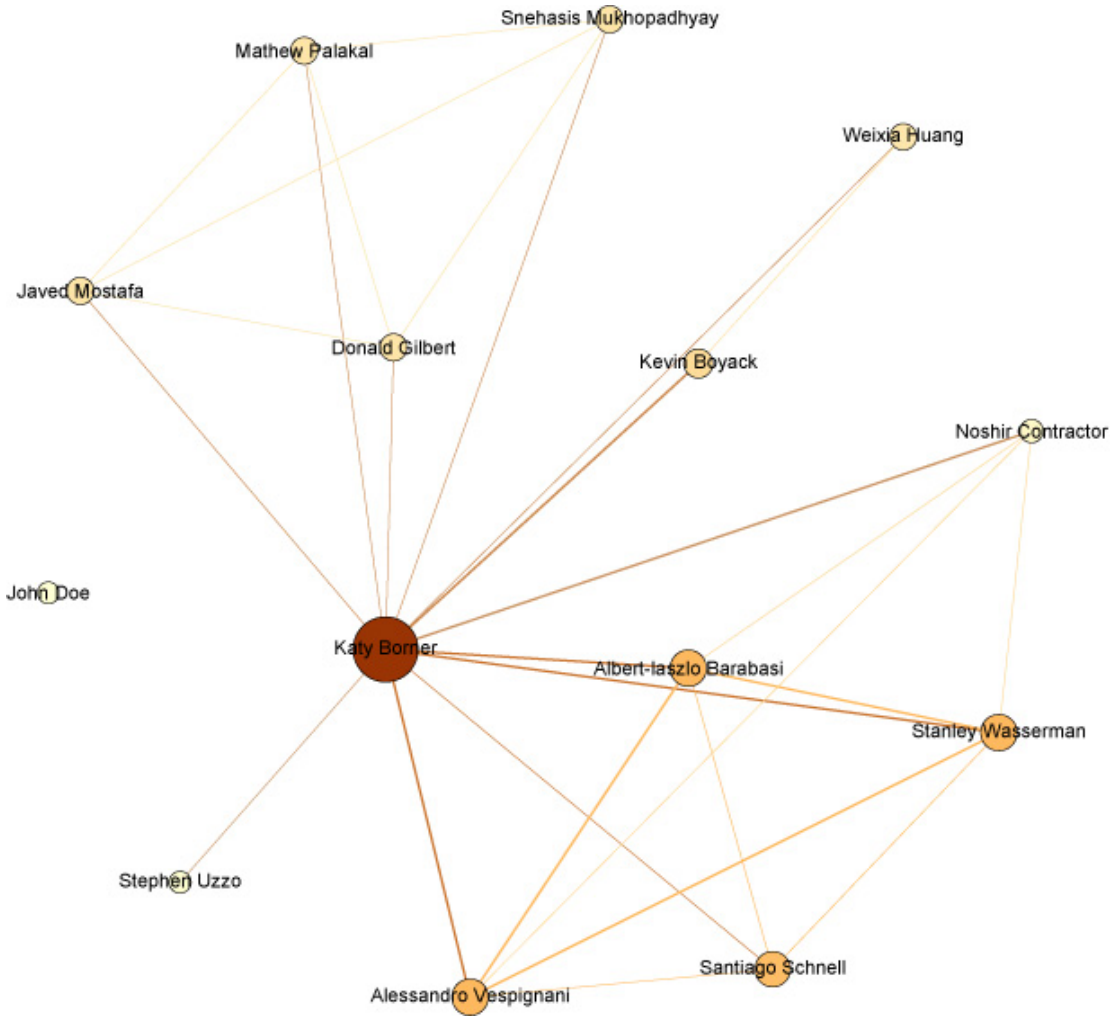
Measurements

- 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



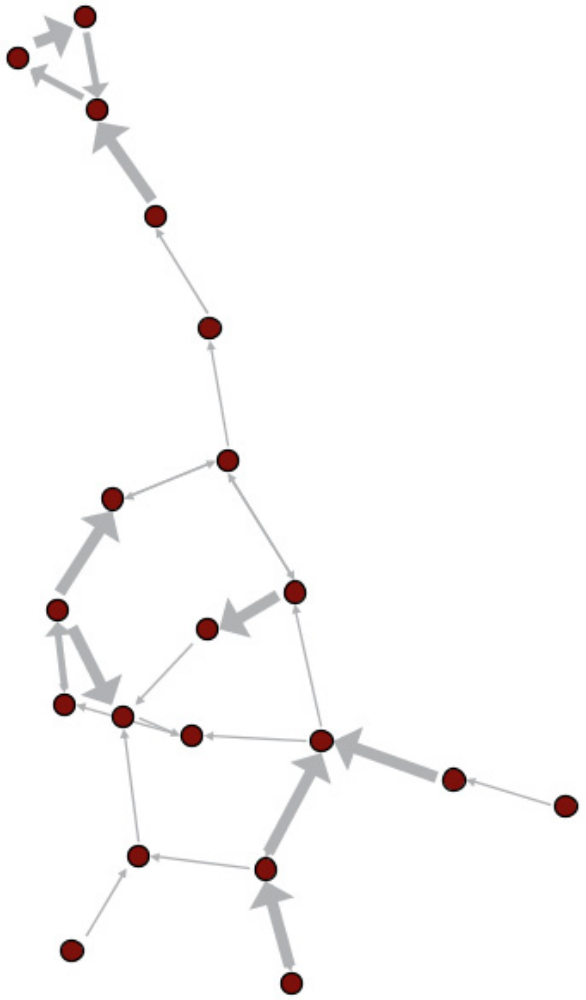
Graph Metrics - Edges

• 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



Tutorial Overview

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

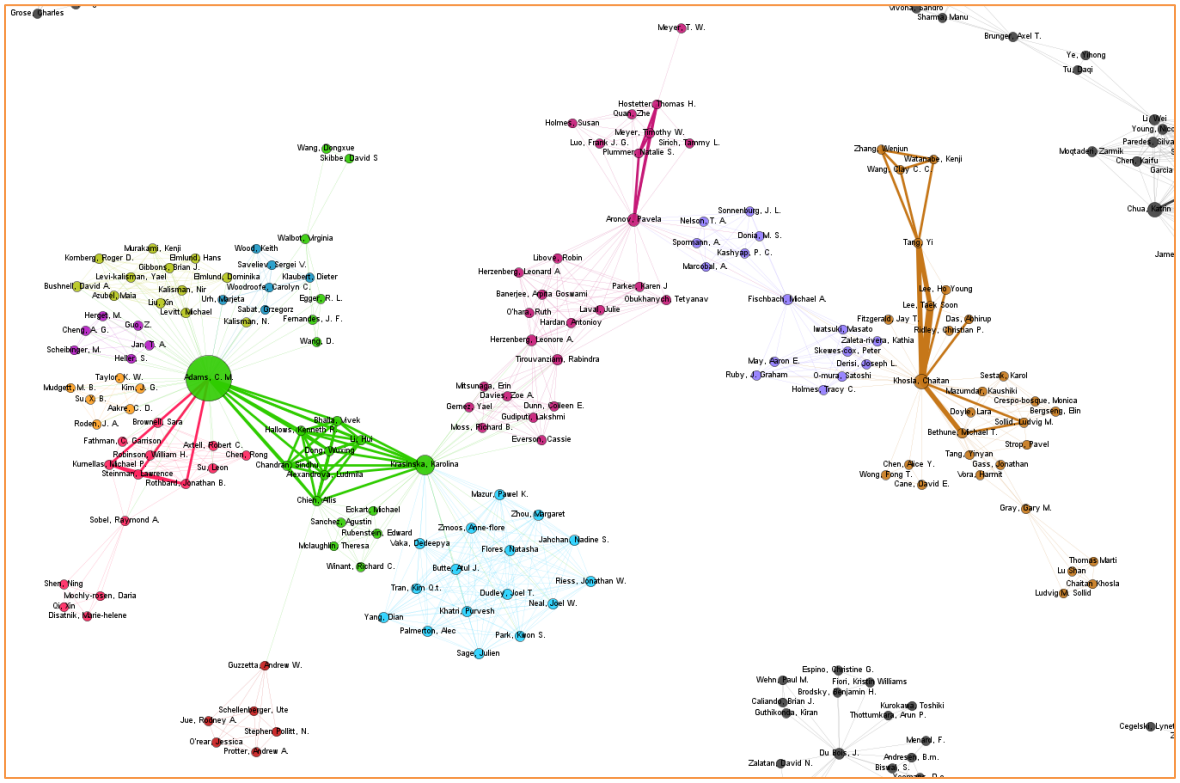
11:00 Topical/Temporal Analysis: Burst Detection
- Overview of burst analysis and introductory workflow
- Burst Detection with CDC Grants

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

Co-Author Network Analysis

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

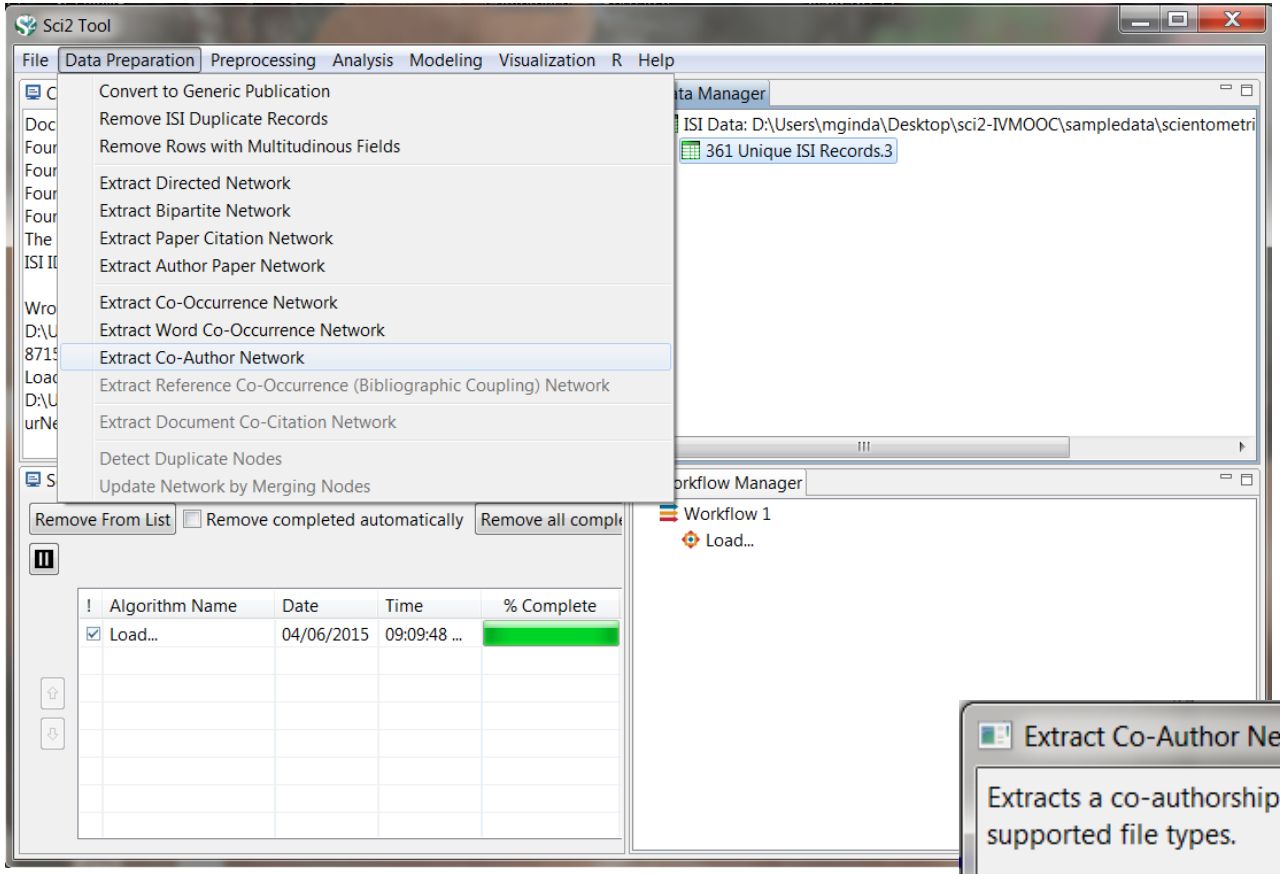


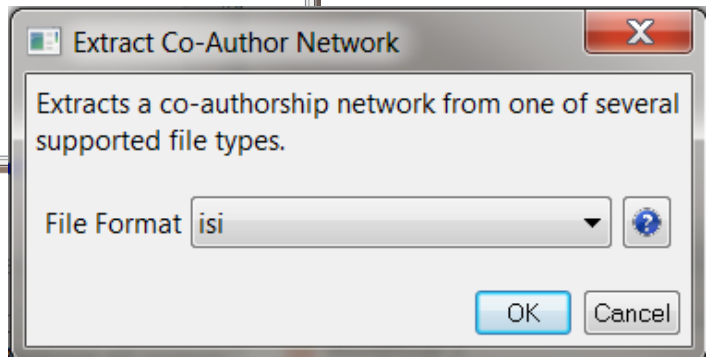
This is a visualization of a citation dataset from a researcher that administers a Core research facility at Stanford University.

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?

Co-Author Network Analysis



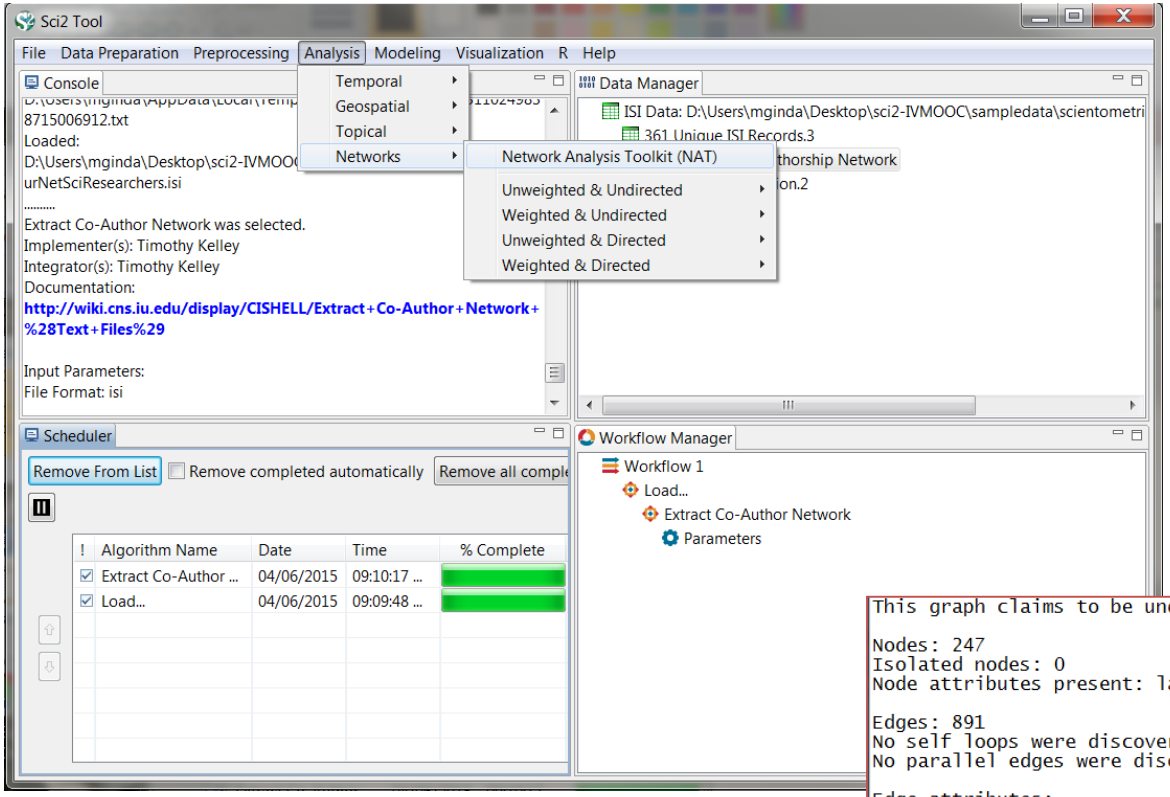


Load *FourNetSciResearchers.isi* located in Sci2 Directory

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.

Co-Author Network Analysis



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

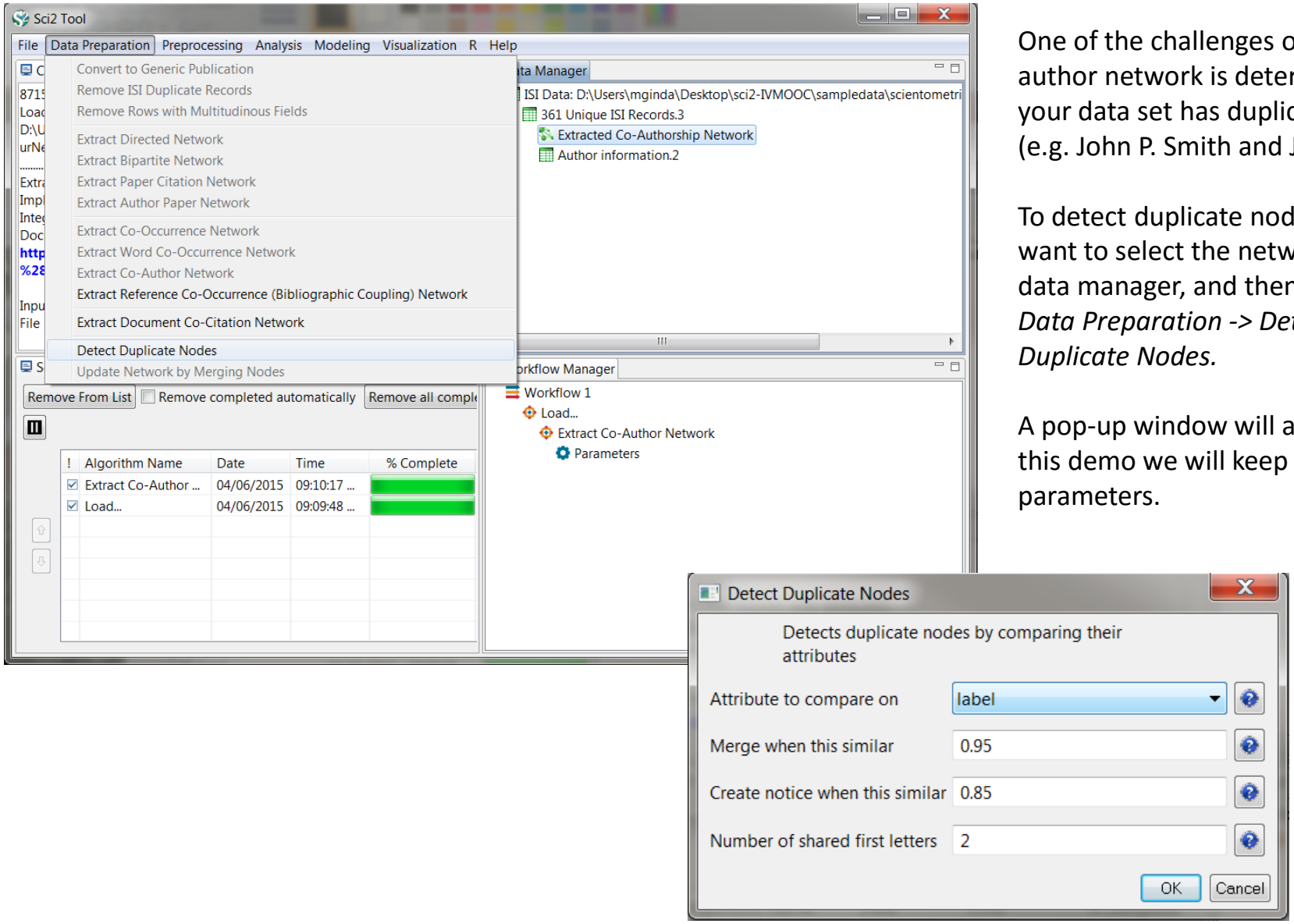
Sci2 has a built-in analysis toolkit to perform these basic statistics.

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
```

Co-Author Network Analysis

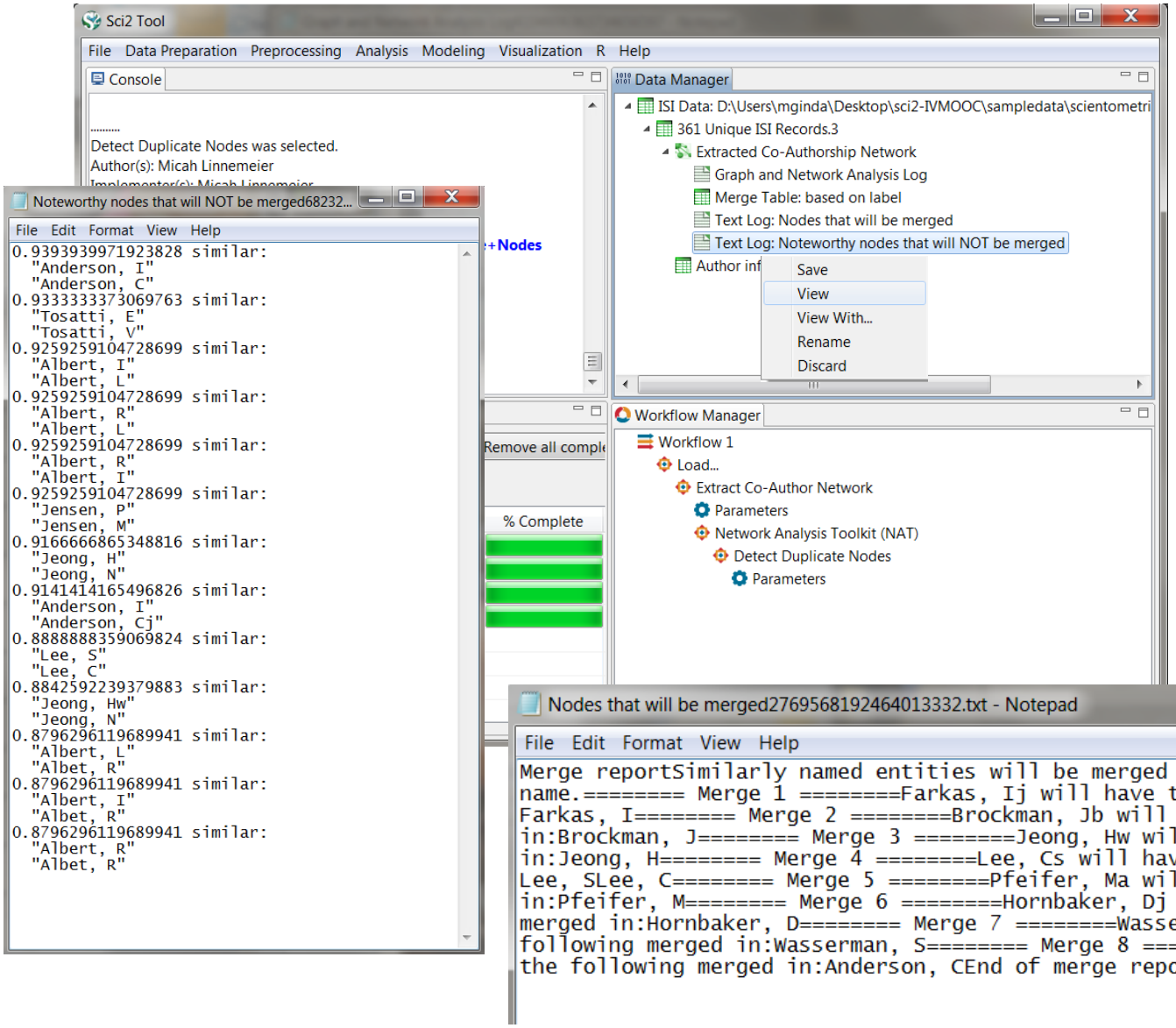


One of the challenges of a co-author network is determining if your data set has duplicate names (e.g. John P. Smith and J P Smith).

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.

Co-Author Network Analysis

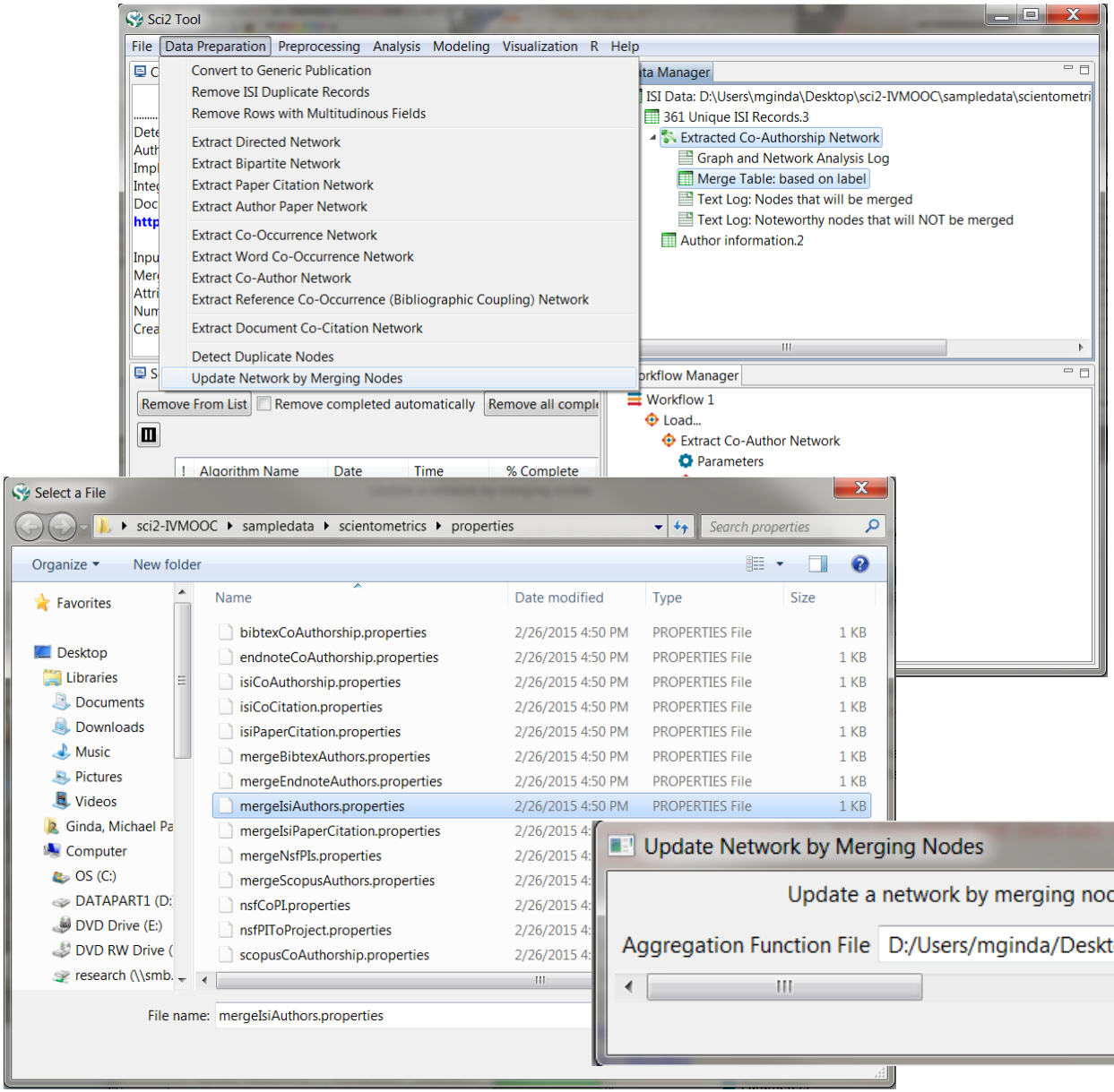


Let's look at the output file *Text Log: Noteworthy nodes that will NOT be merged.*

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.

Co-Author Network Analysis



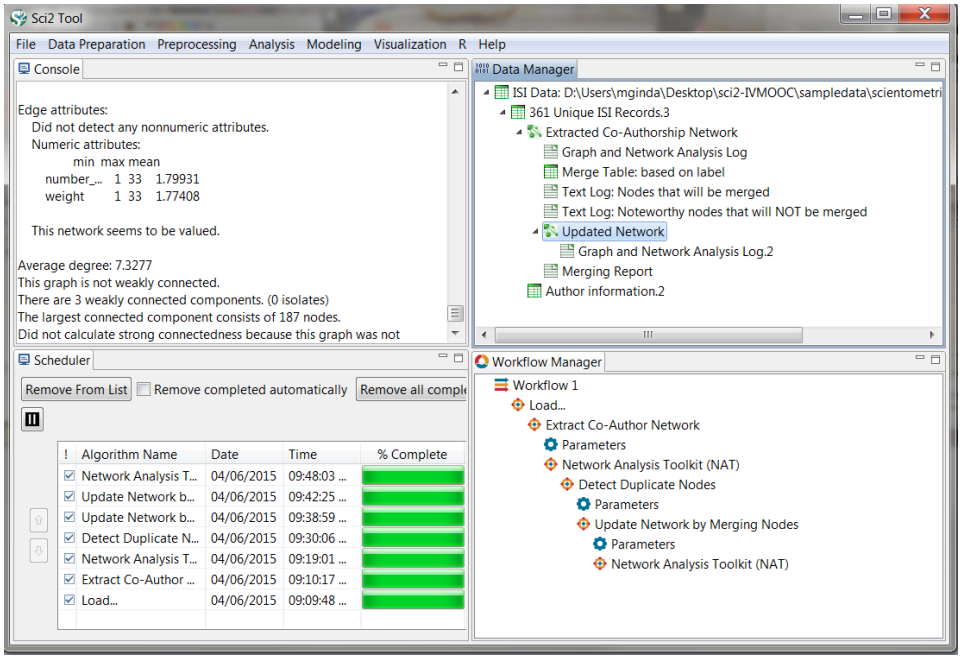
After we've identifying our duplicate nodes, we need to merge these duplicates.

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 *MergeIsiAuthors.properties*

Select open, and then OK.

Co-Author Network Analysis



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?

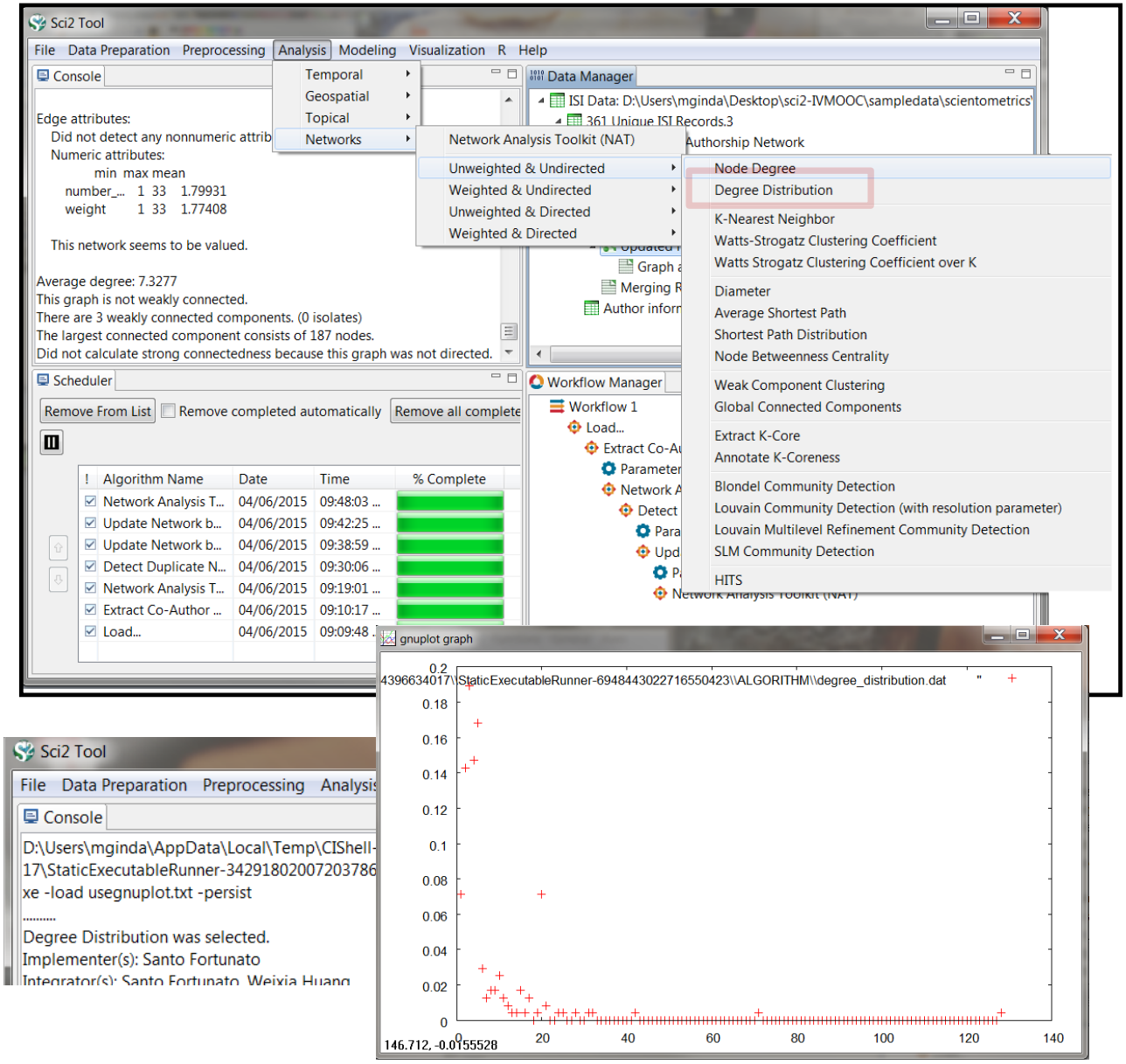
Original Network

```
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
```

Revised Network

```
This graph claims to be undirected.

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
```

Co-Author Network Analysis



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

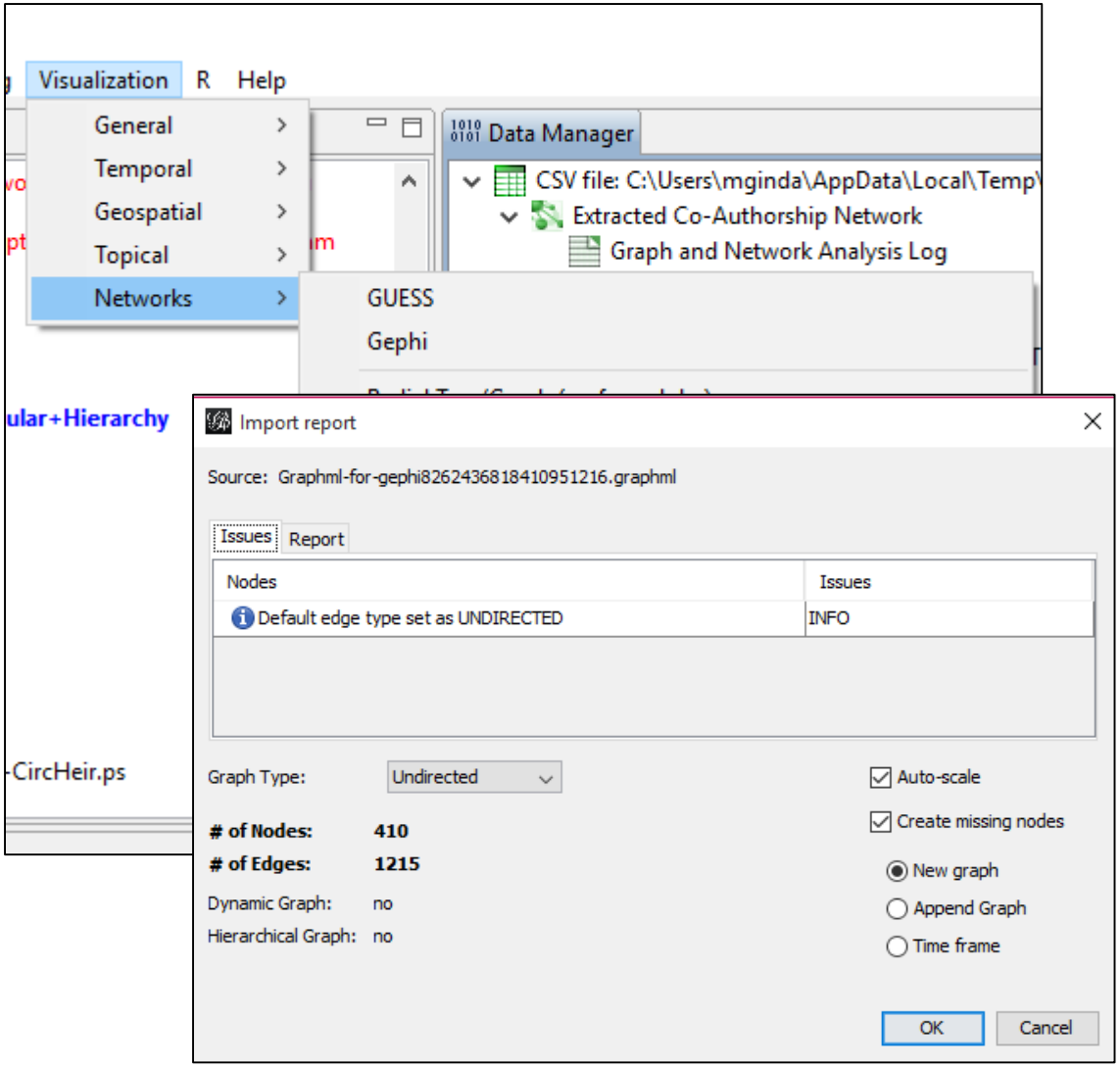
Select the updated network in the data manager, and then select in the menu *Analysis -> Networks-> Unweighted & Undirected -> Node Degree*

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*

Co-Author Network Visualization in Gephi



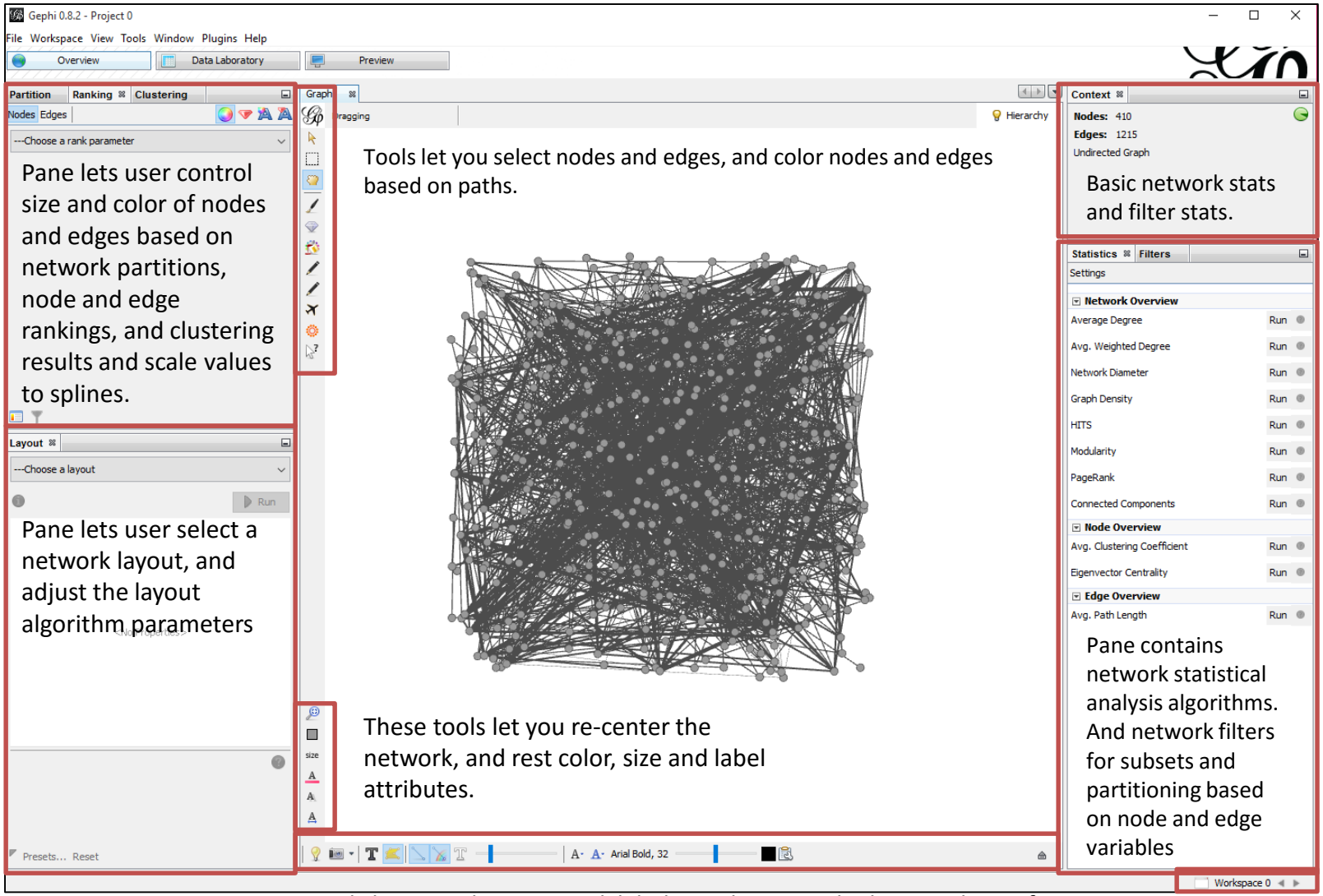
Next we will visualize the network in Gephi.

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

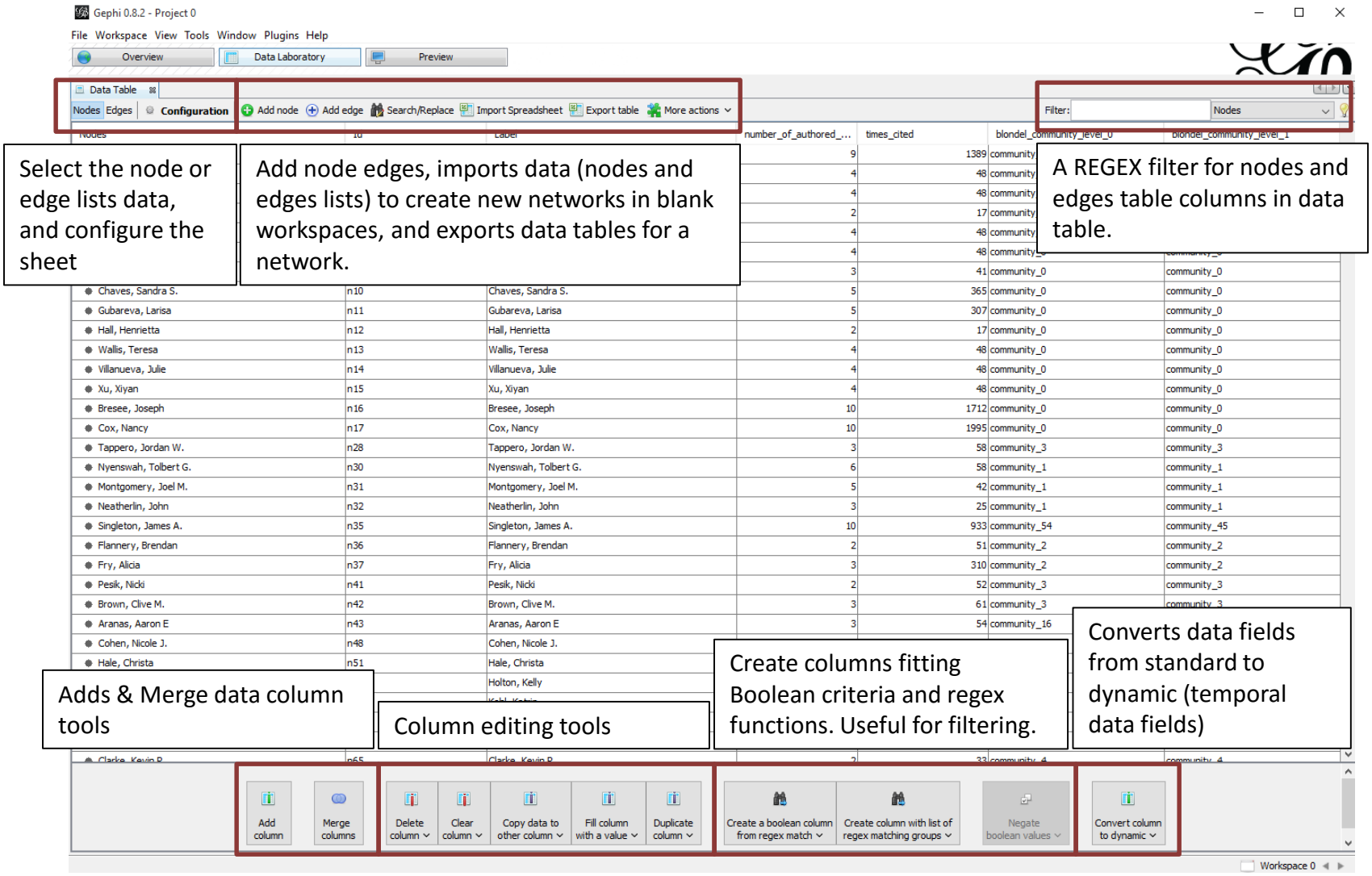
The algorithm is a bridge that passes the network data to Gephi. The program will automatically start. The tool produces an Import Report. It lets you select the network type, gives load errors, etc.

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

Overview of Gephi: Overview Pane



Overview of Gephi: Data Laboratory



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.

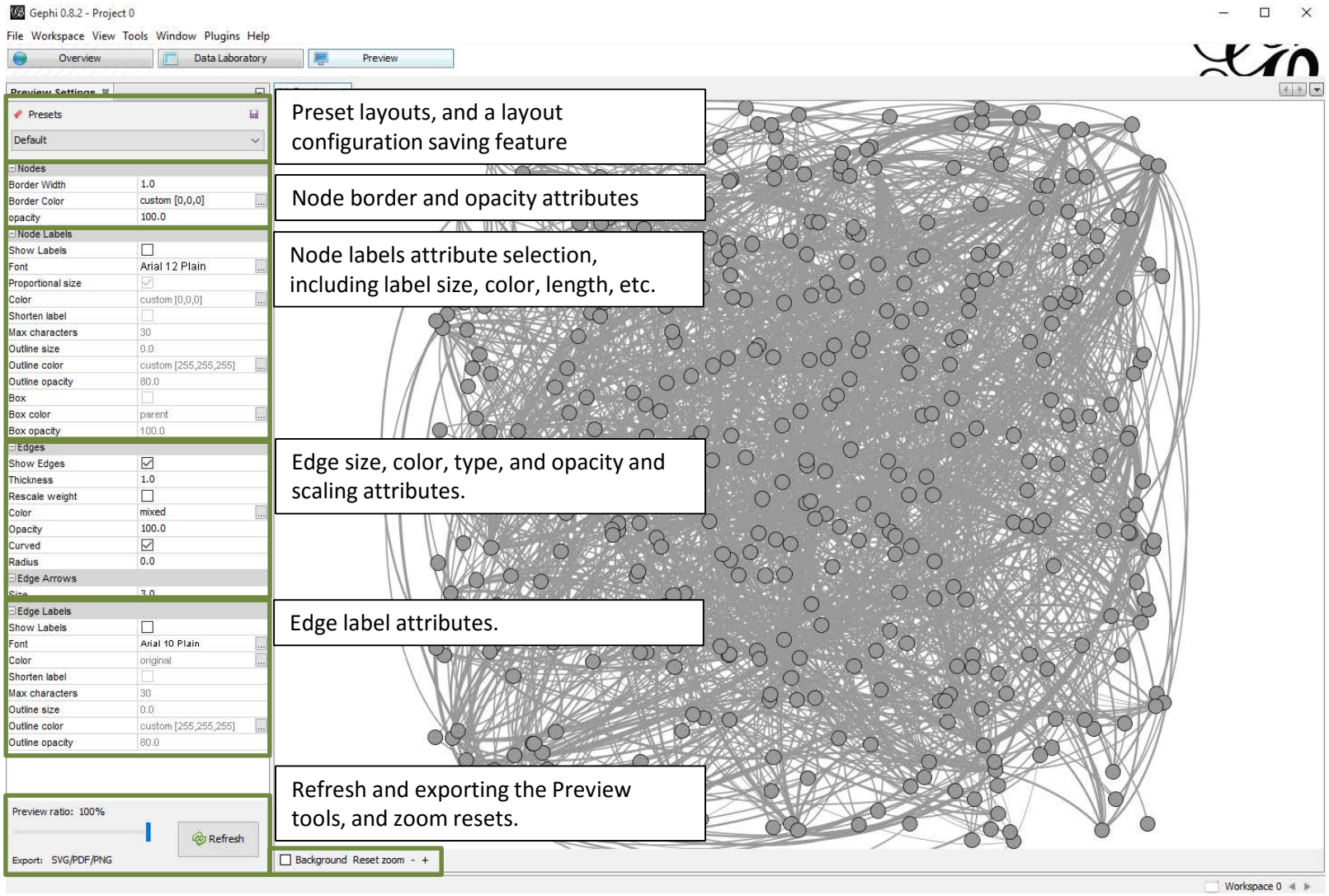
Adds & Merge data column tools

Column editing tools

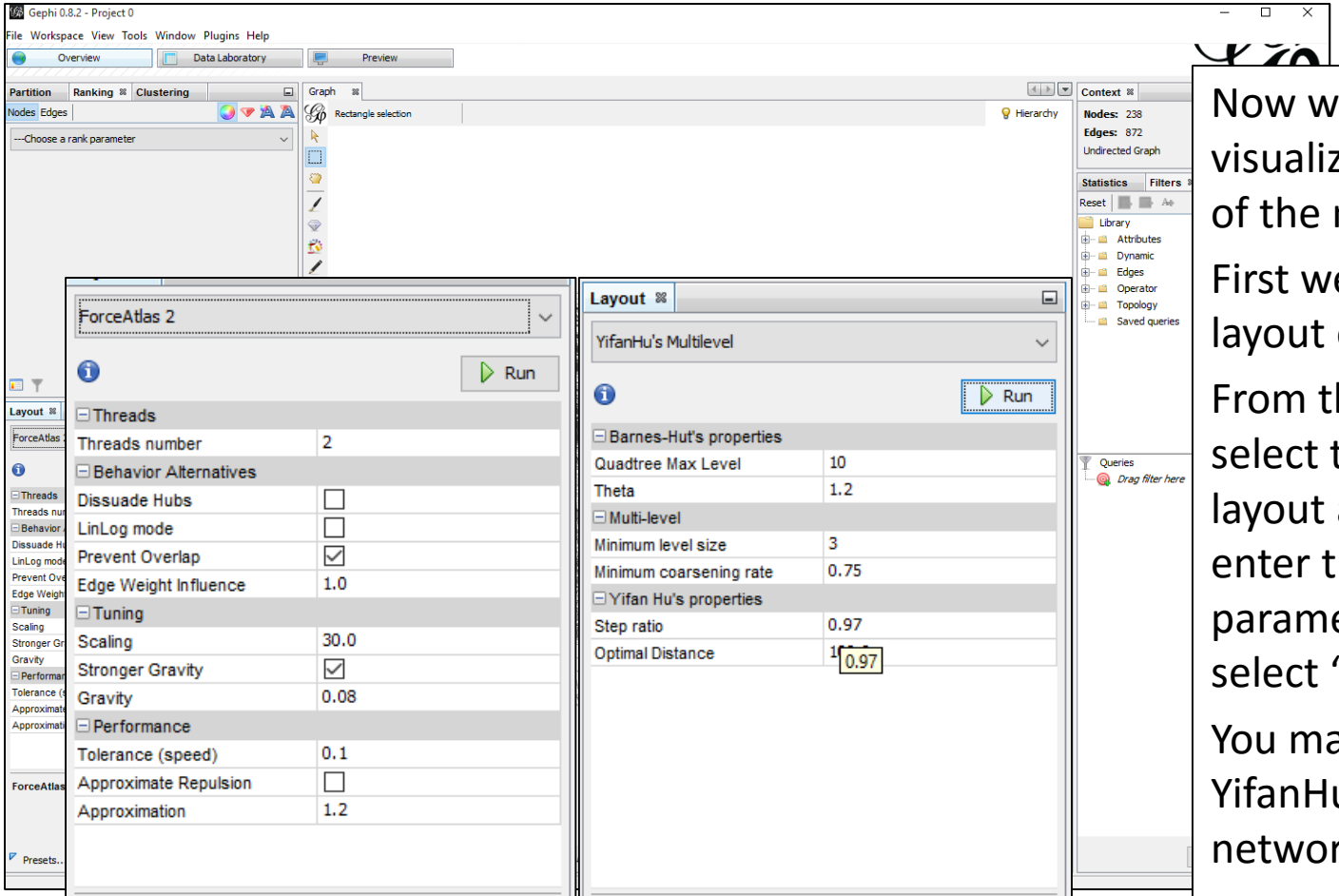
Create columns fitting Boolean criteria and regex functions. Useful for filtering.

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

Overview of Gephi: Preview



Gephi: Initial Layout of Network



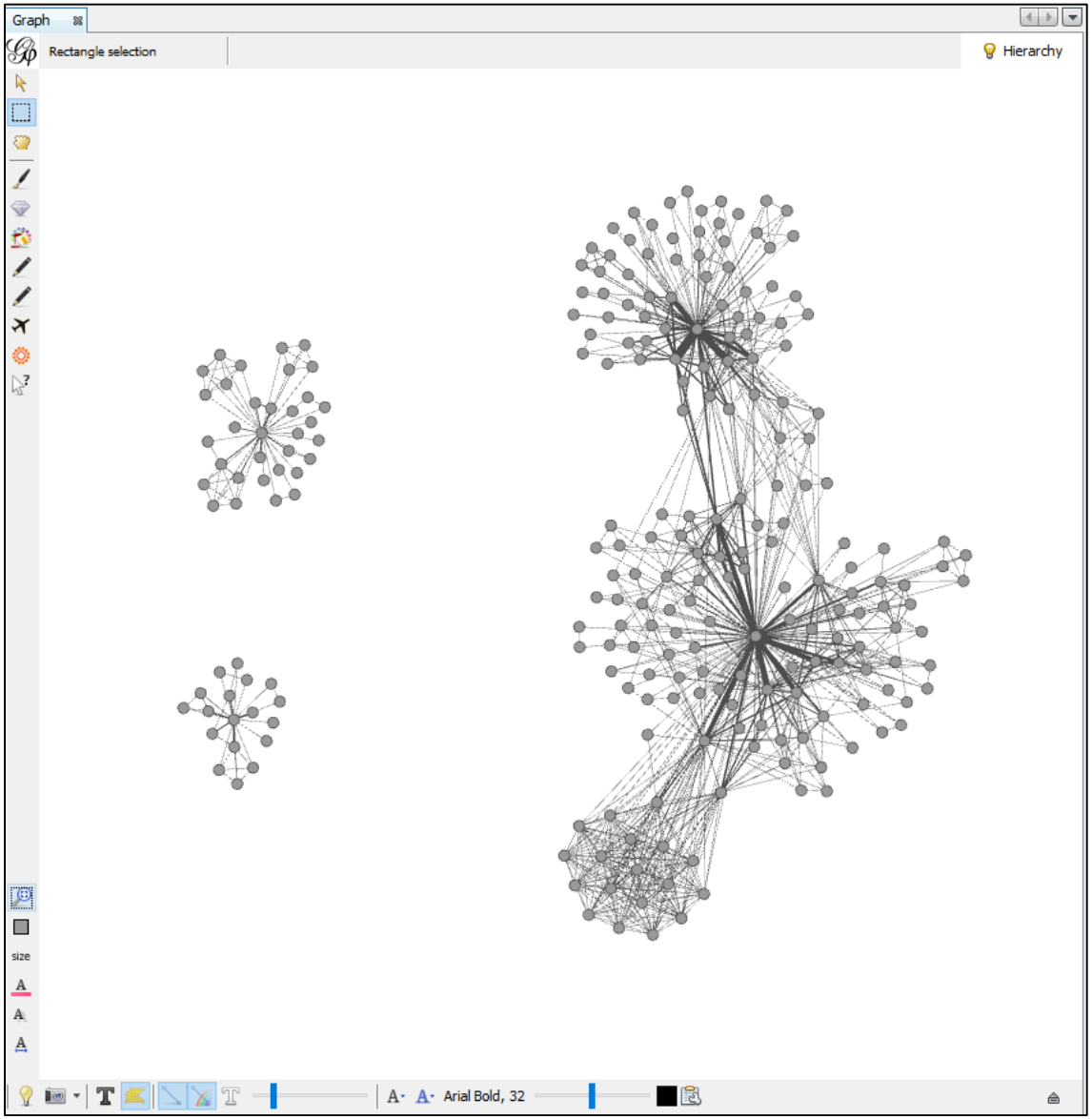
Now we can start an visualization and analysis of the network.

First we will adjust the layout of the network.

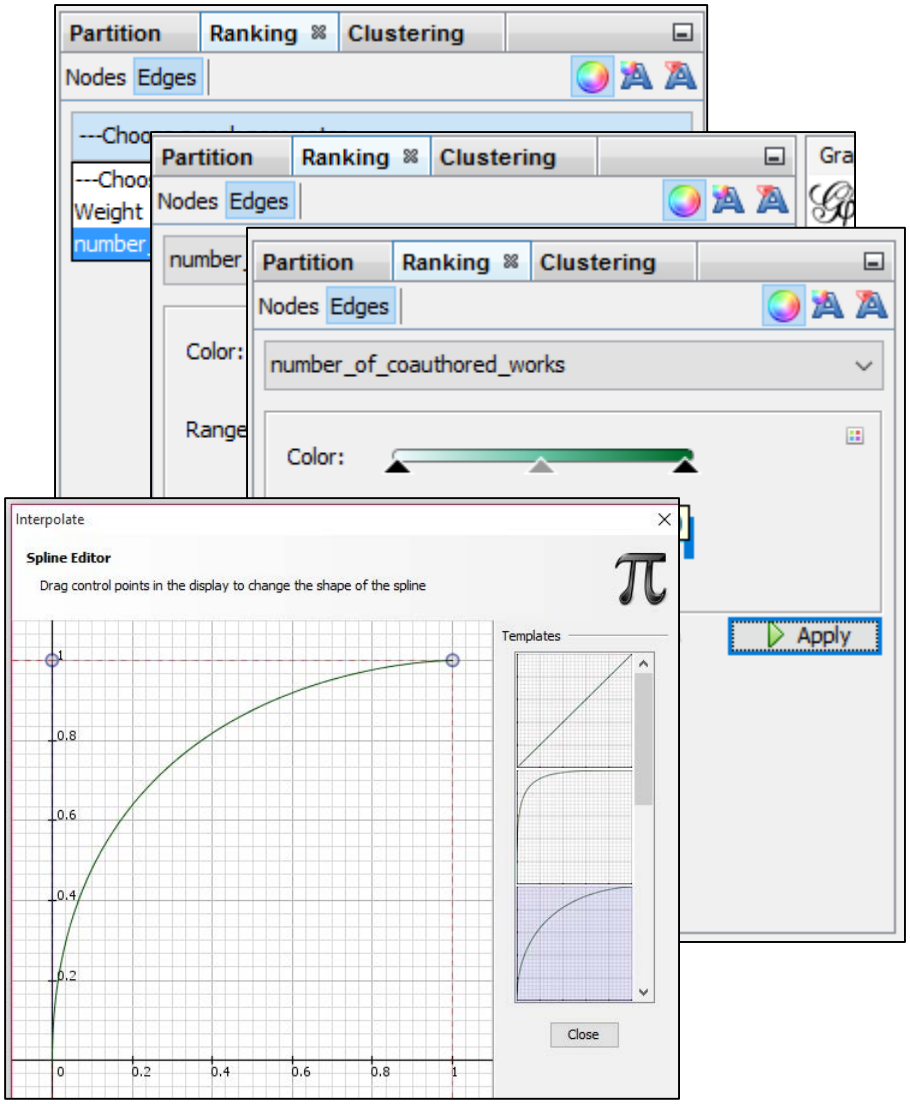
From the layout pane, select the “ForceAtlas2” layout algorithm and enter the following parameters, and then select “Run”.

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

Gephi: Initial Layout of Network



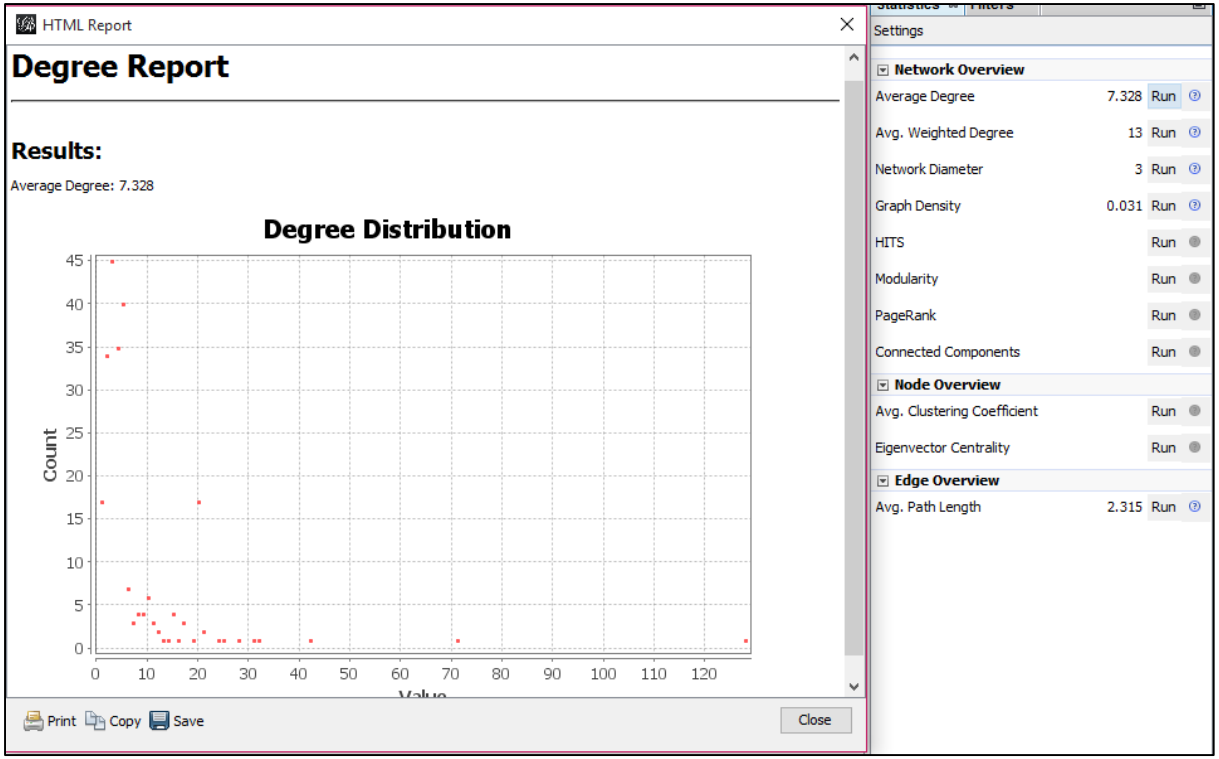
Gephi: Edge Color



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

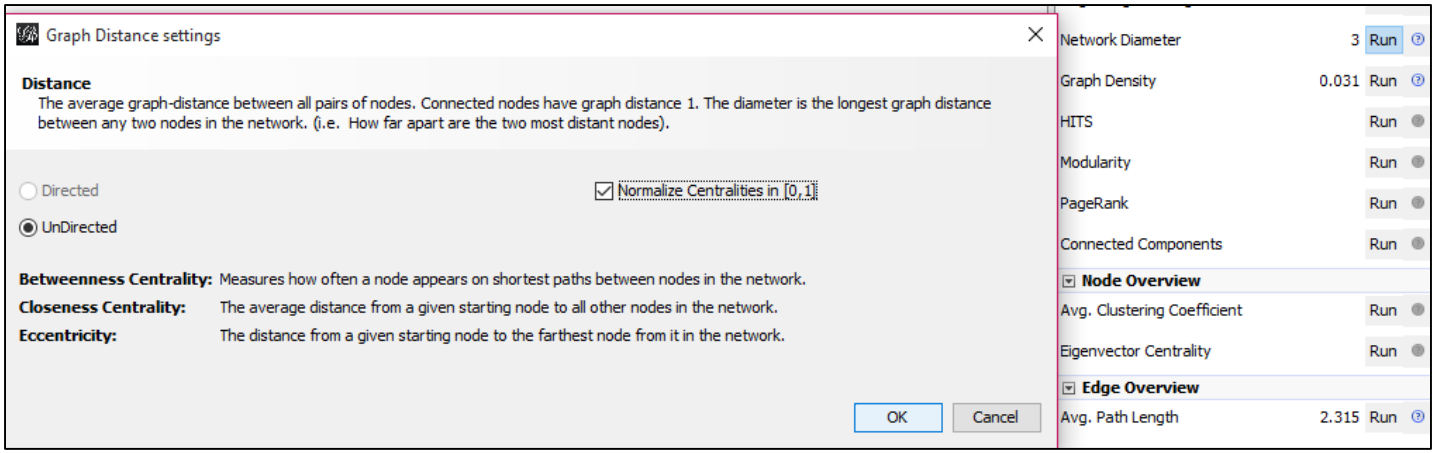
- In the drop down menu, select “number_of_coauthored_works”.
- Select the small square in the right corner of the Color Range box. This lets us choose new color ranges for variables.
- You may also set the color range and values to apply the colors to, or adjust color scaling variables by adjusting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

Gephi: Network statistics

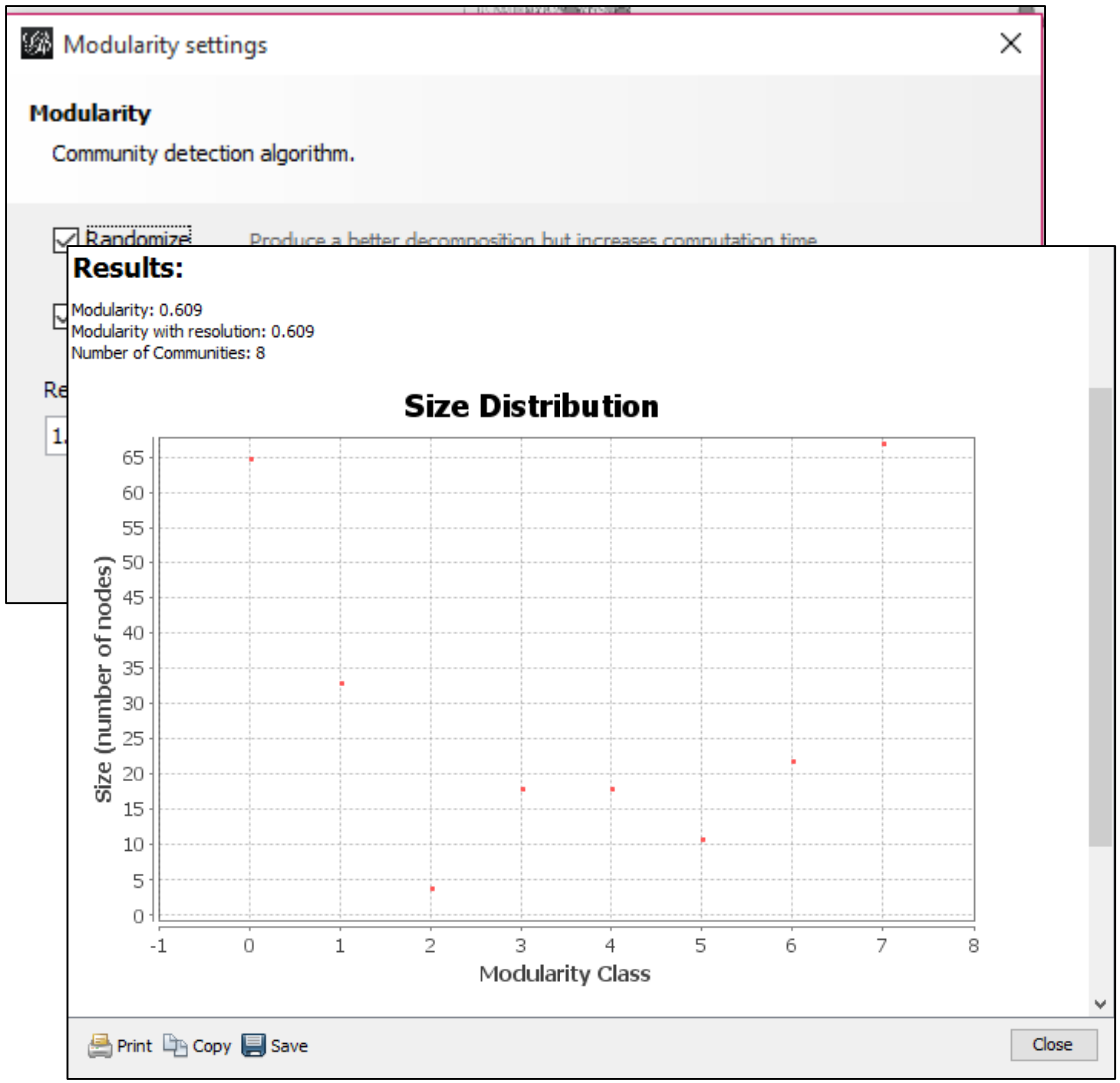


Gephi provides a variety of node and edge statistics to help understand the relationships, clustering, paths, centrality, and communities within a network.

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

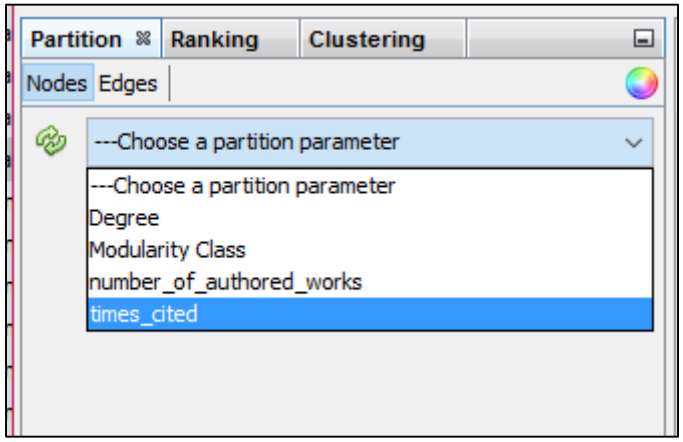


Gephi: Network statistics/Community Detection

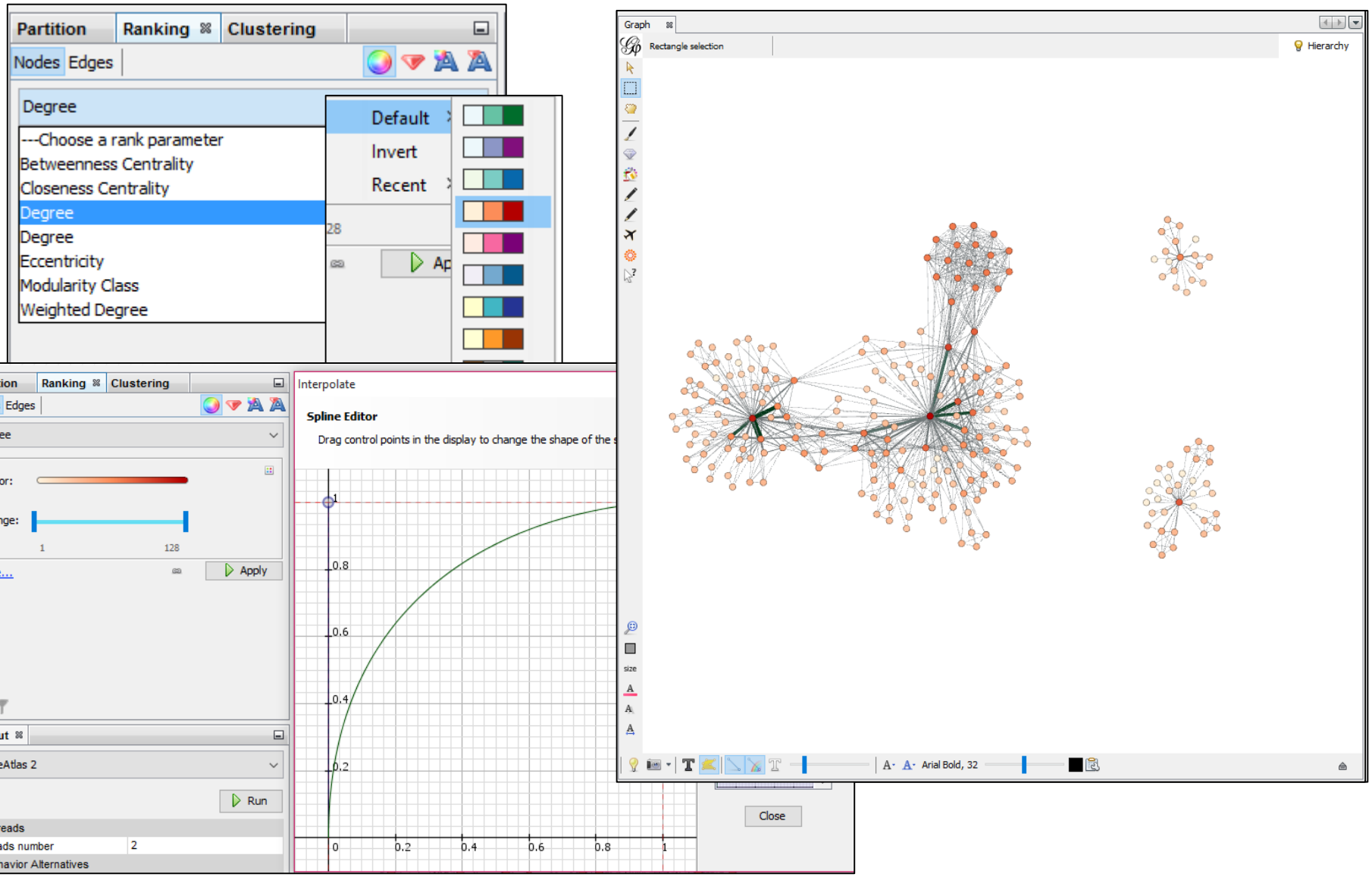


The modularity statistical algorithm calculates how the connectedness of a network, and the Blondel Communities that exist in the network. The communities are added as a partition to the nodes.

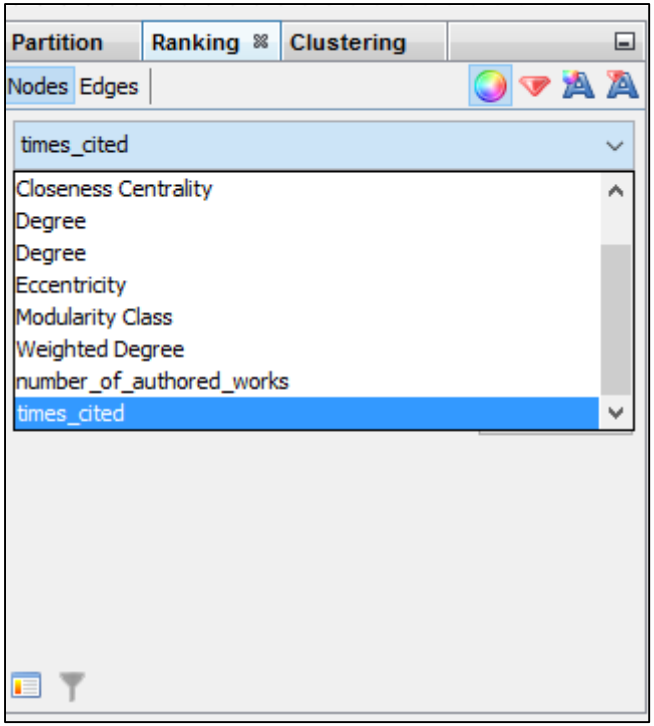
The modularity categories may be applied to the network from the Partitions window.

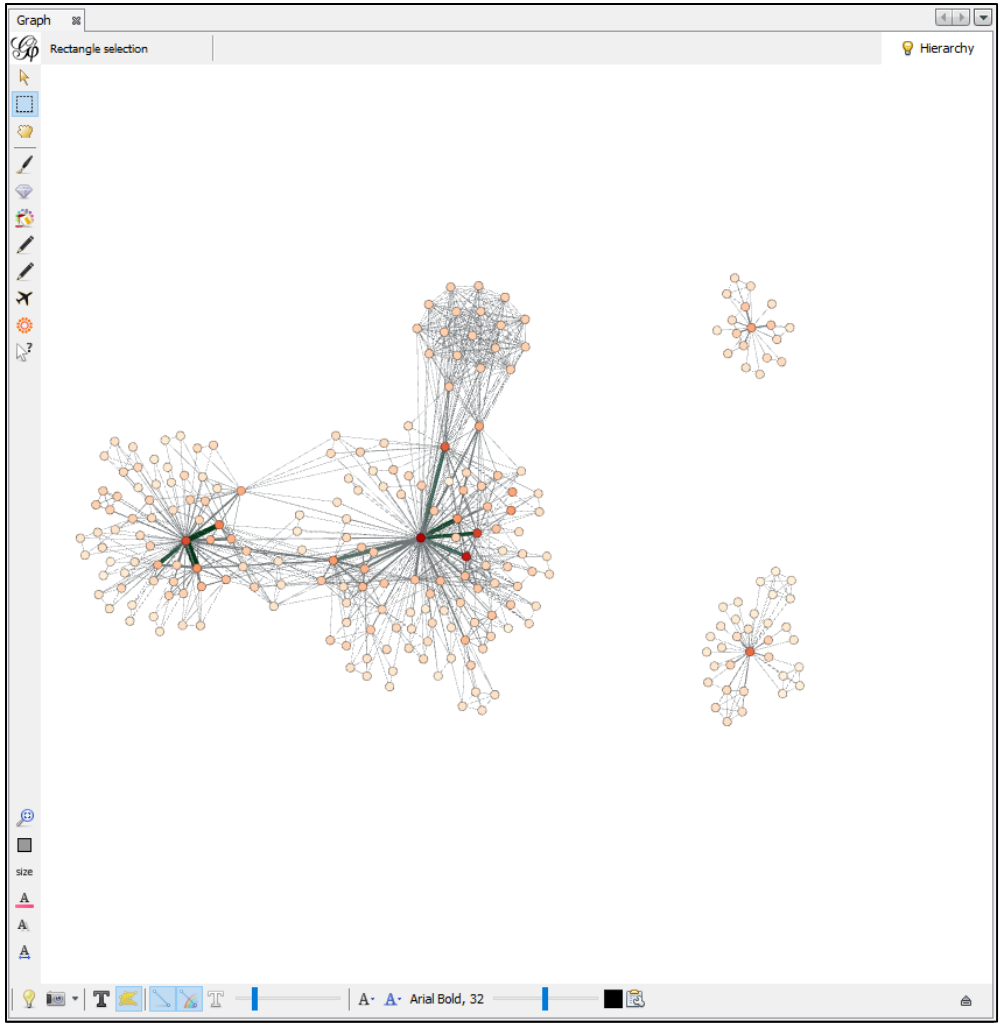


Gephi: Node Color Ranking & Scaling – Degree & Times Cited

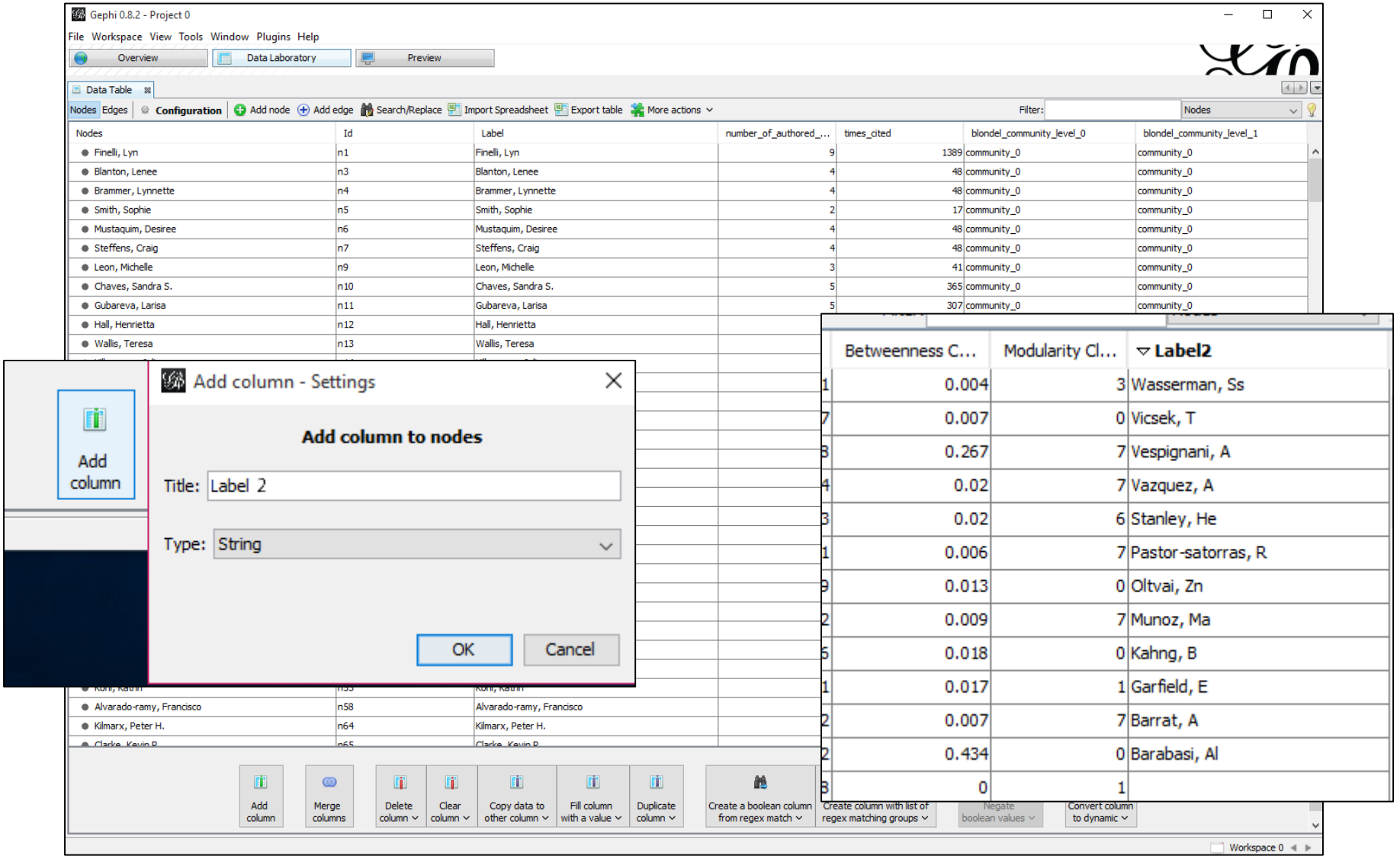


Gephi: Node Color Ranking & Scaling – Betweenness Centrality

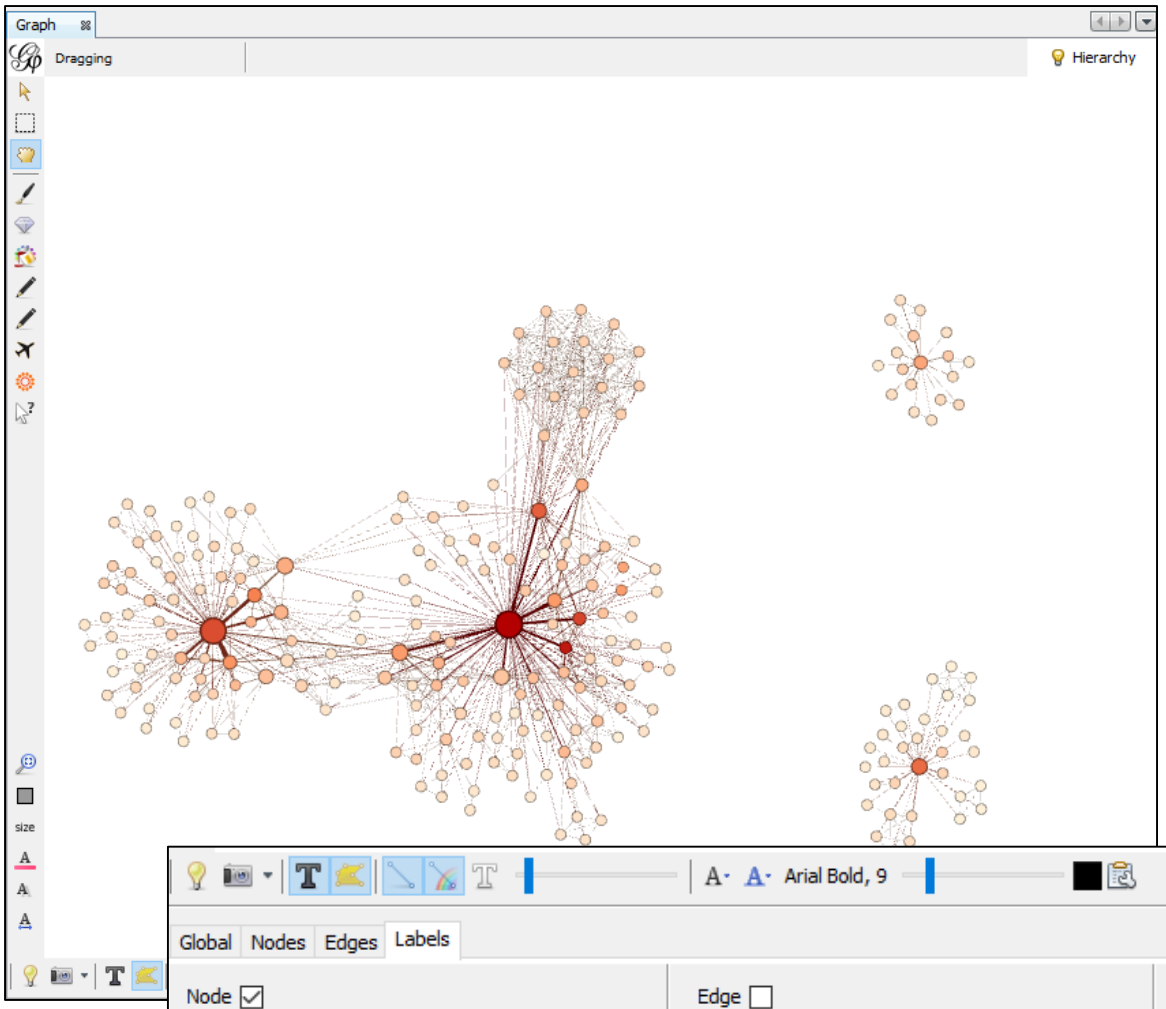


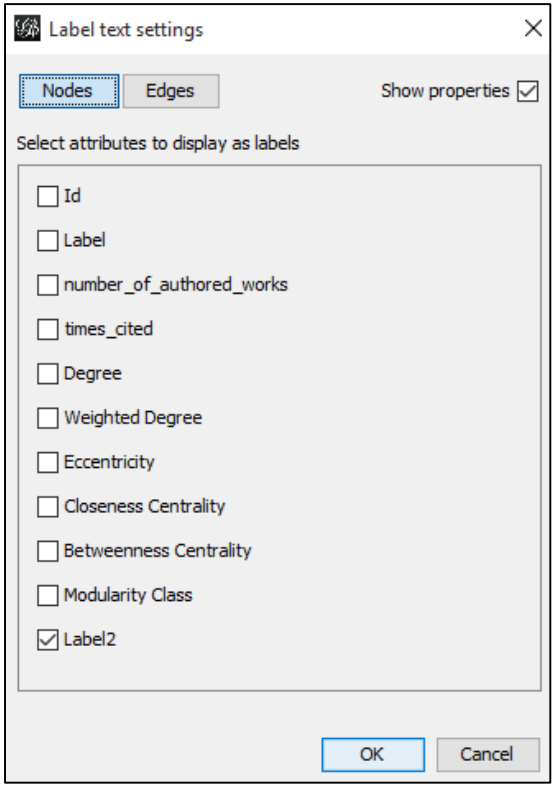


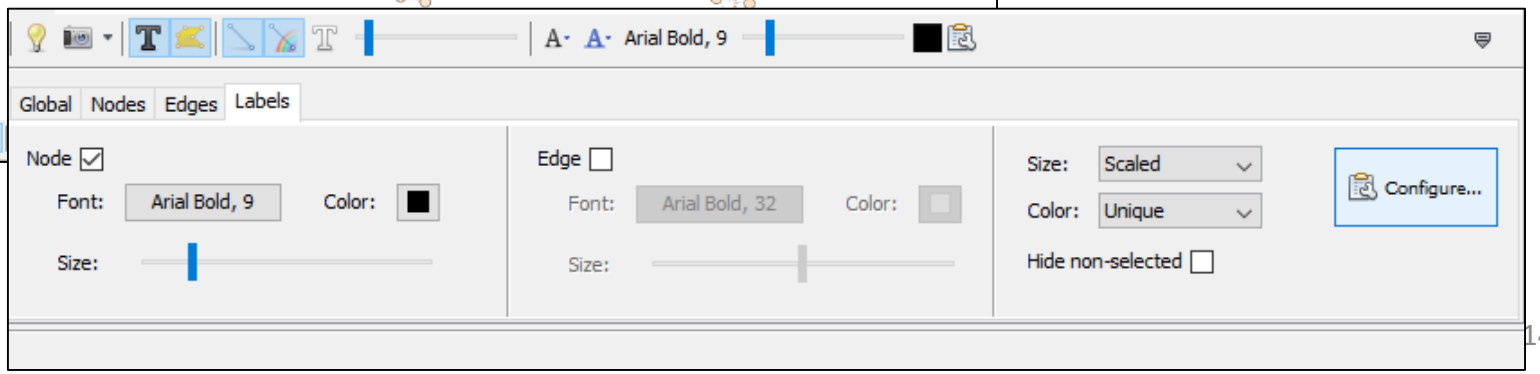
Gephi: Node Labels – Data Laboratory



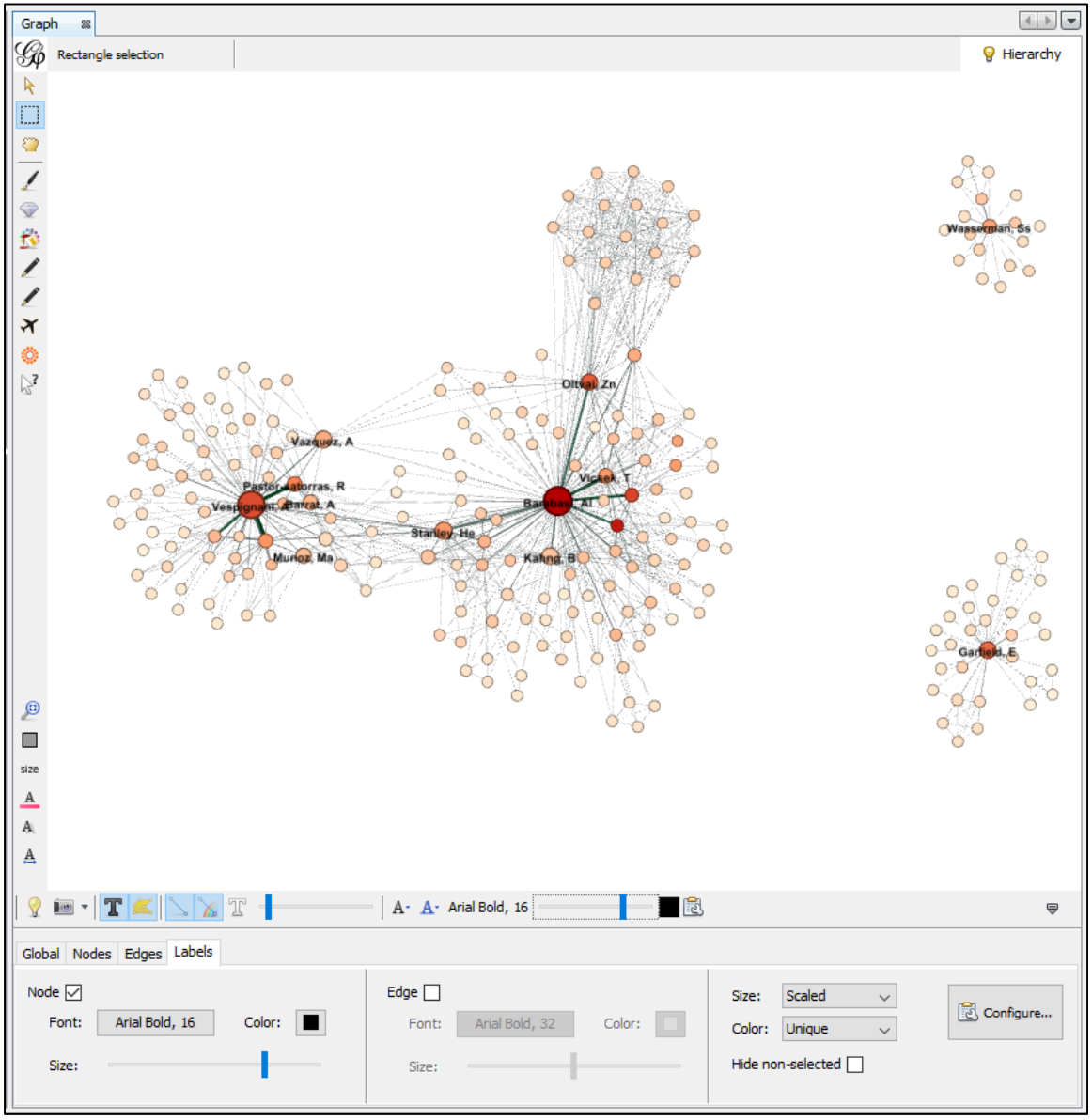
Gephi: Node Labels – Overview



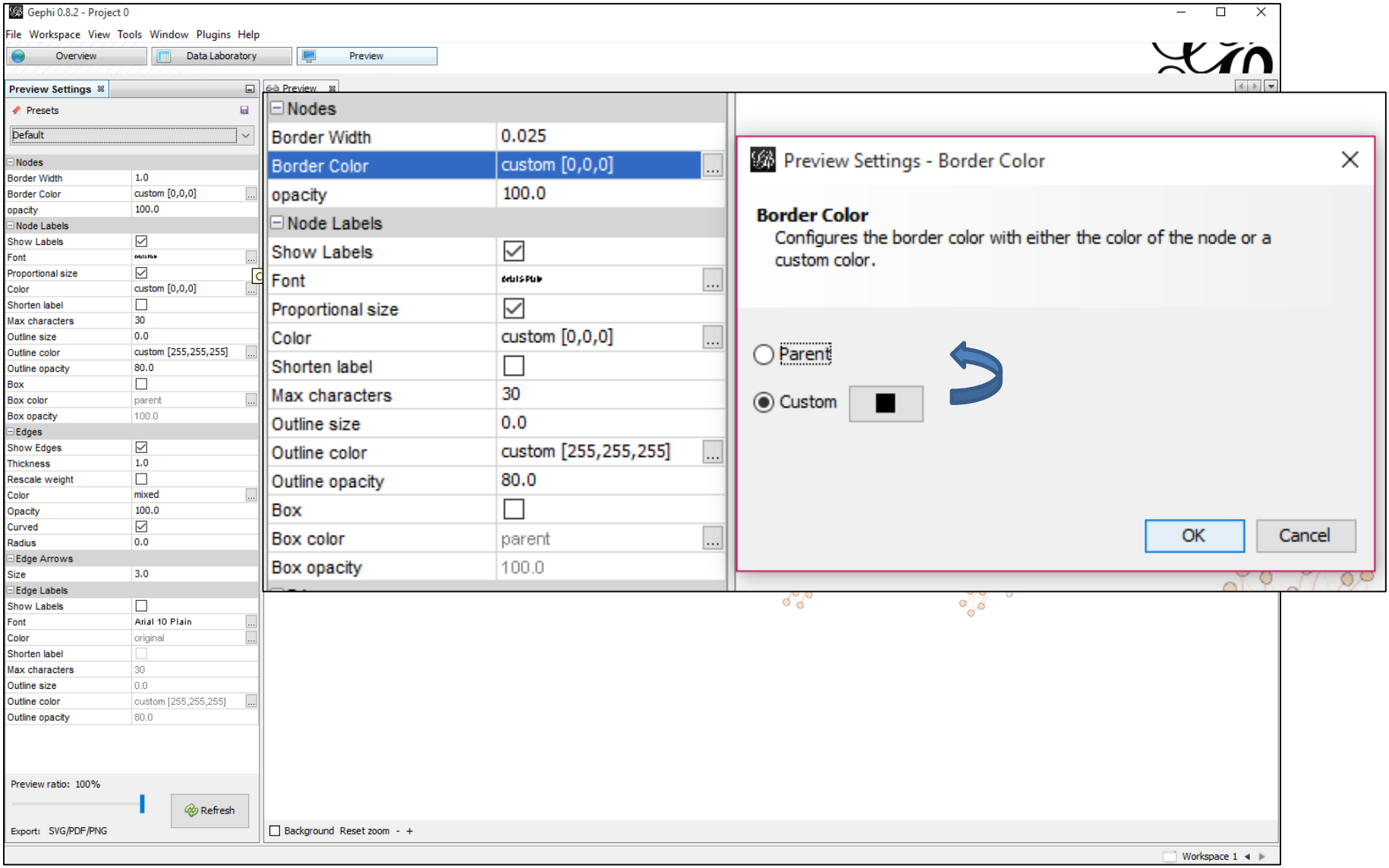




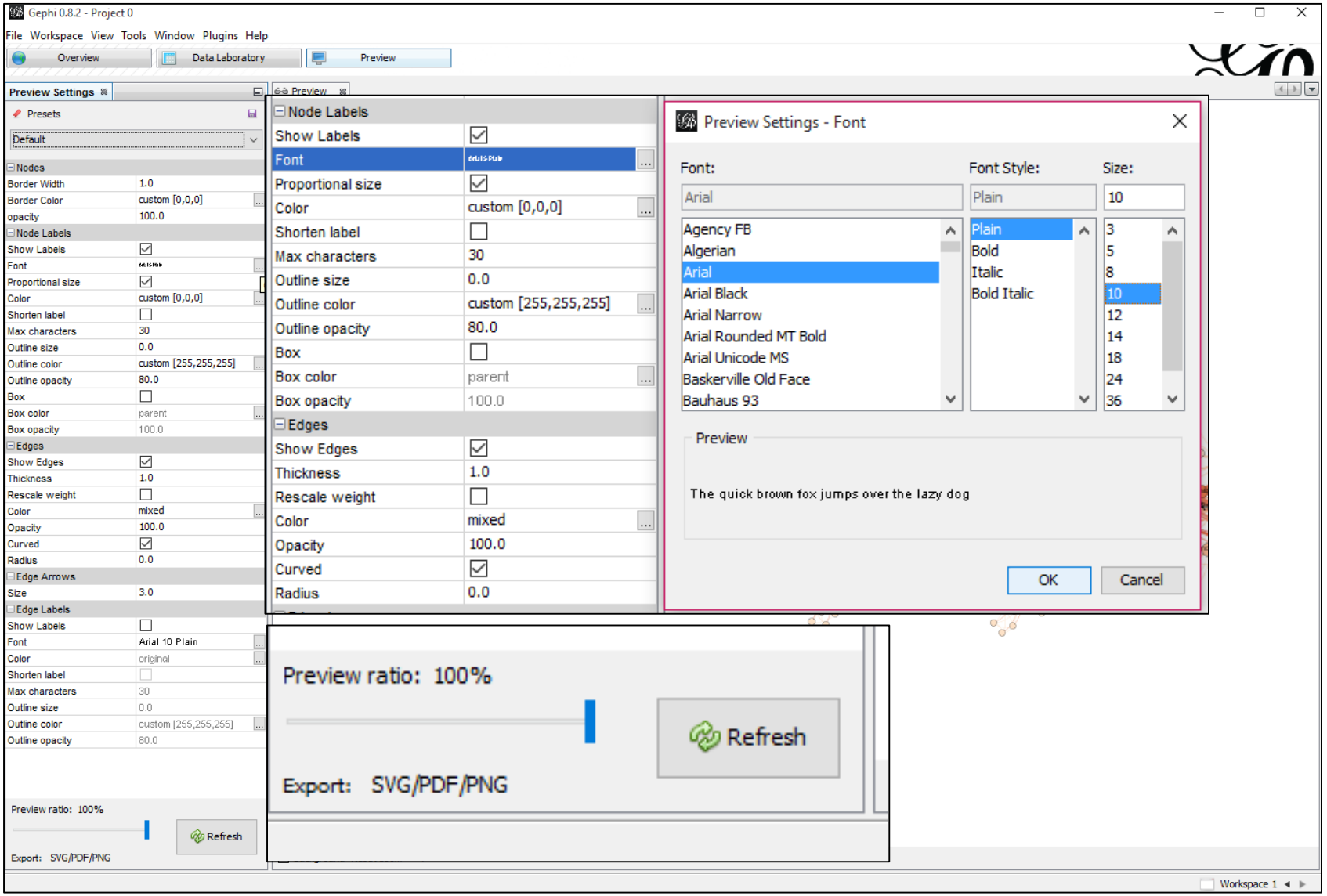
Gephi: Node Labels – Overview



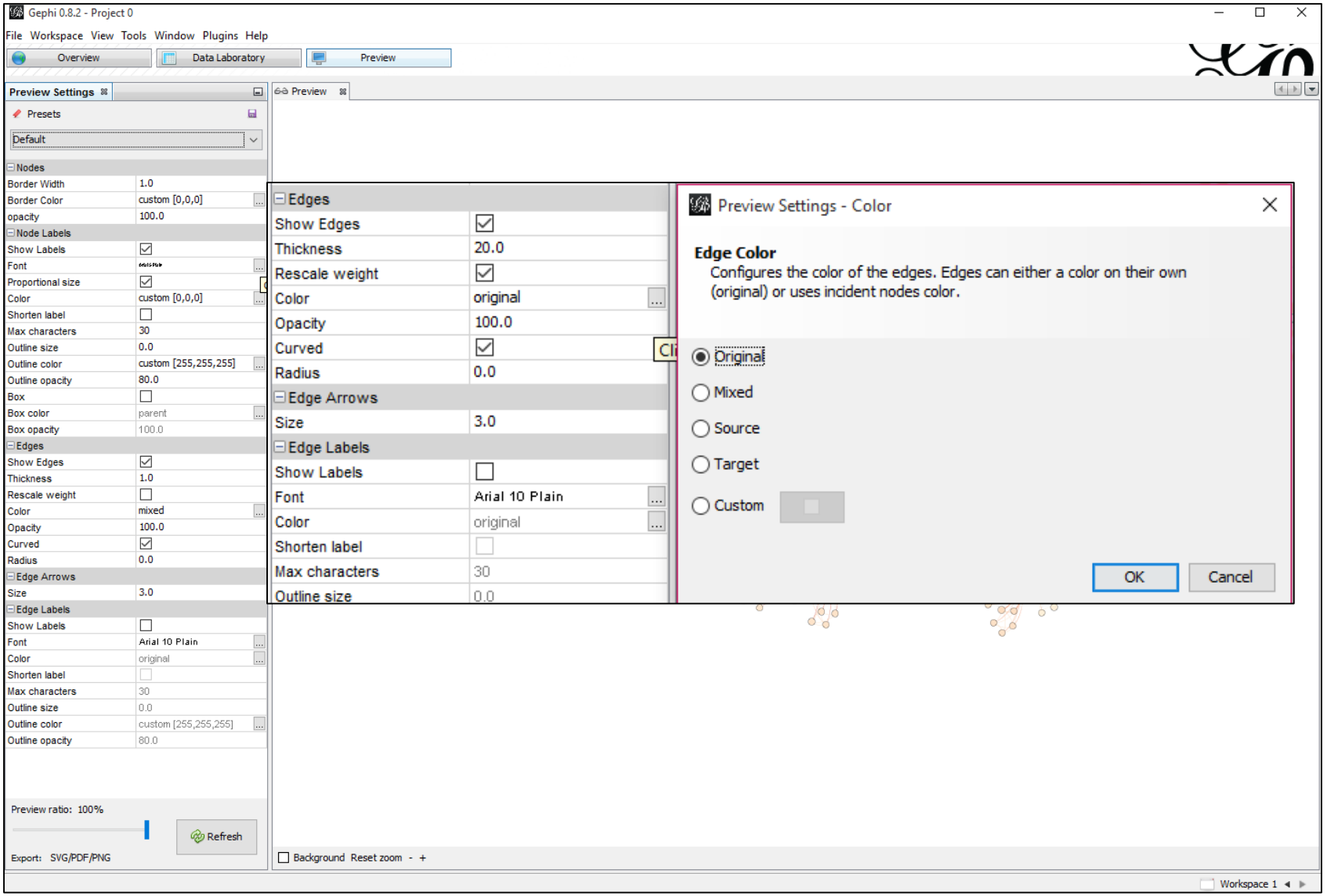
Gephi: Final visualization – Node Parameters



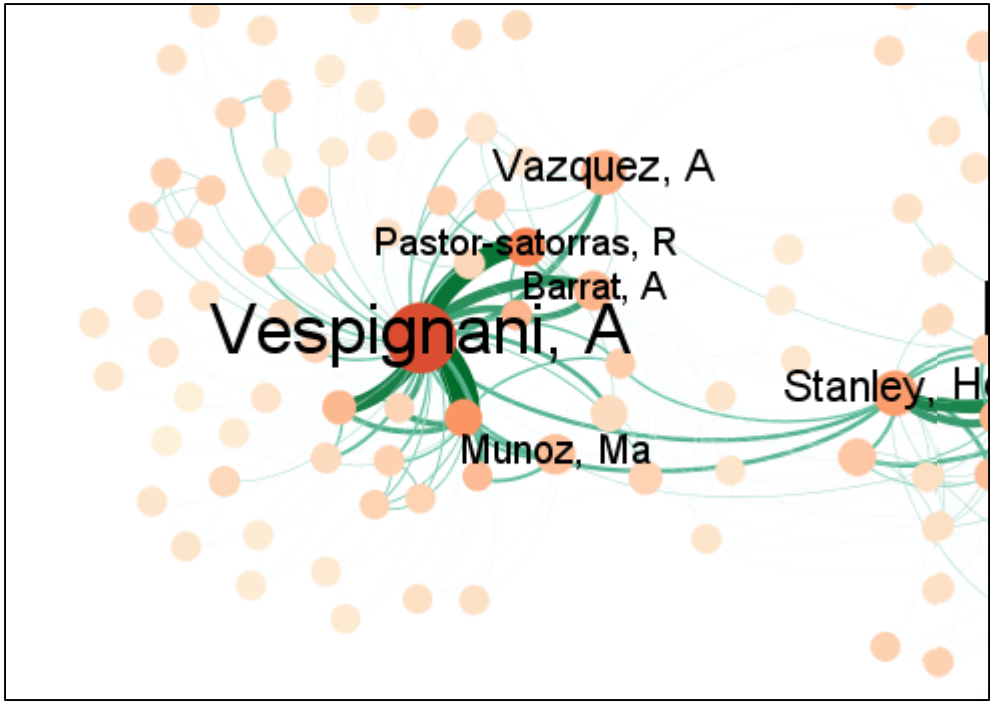
Gephi: Final visualization – Node Parameters



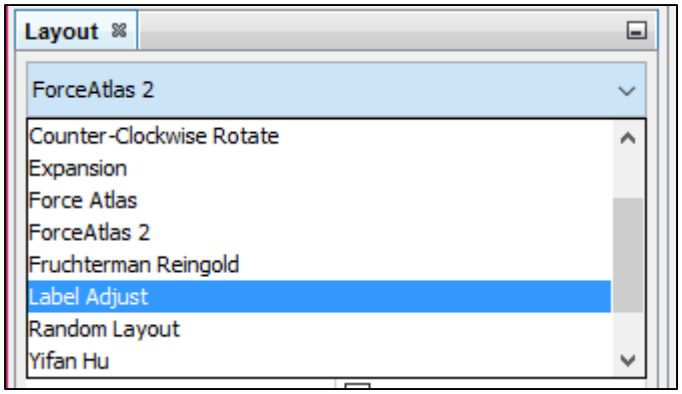
Gephi: Final visualization – Edge Color Parameters



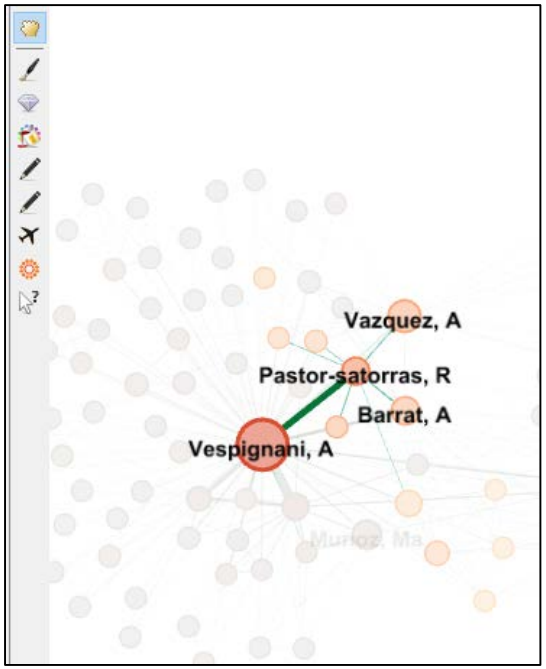
Gephi: Label Adjustments



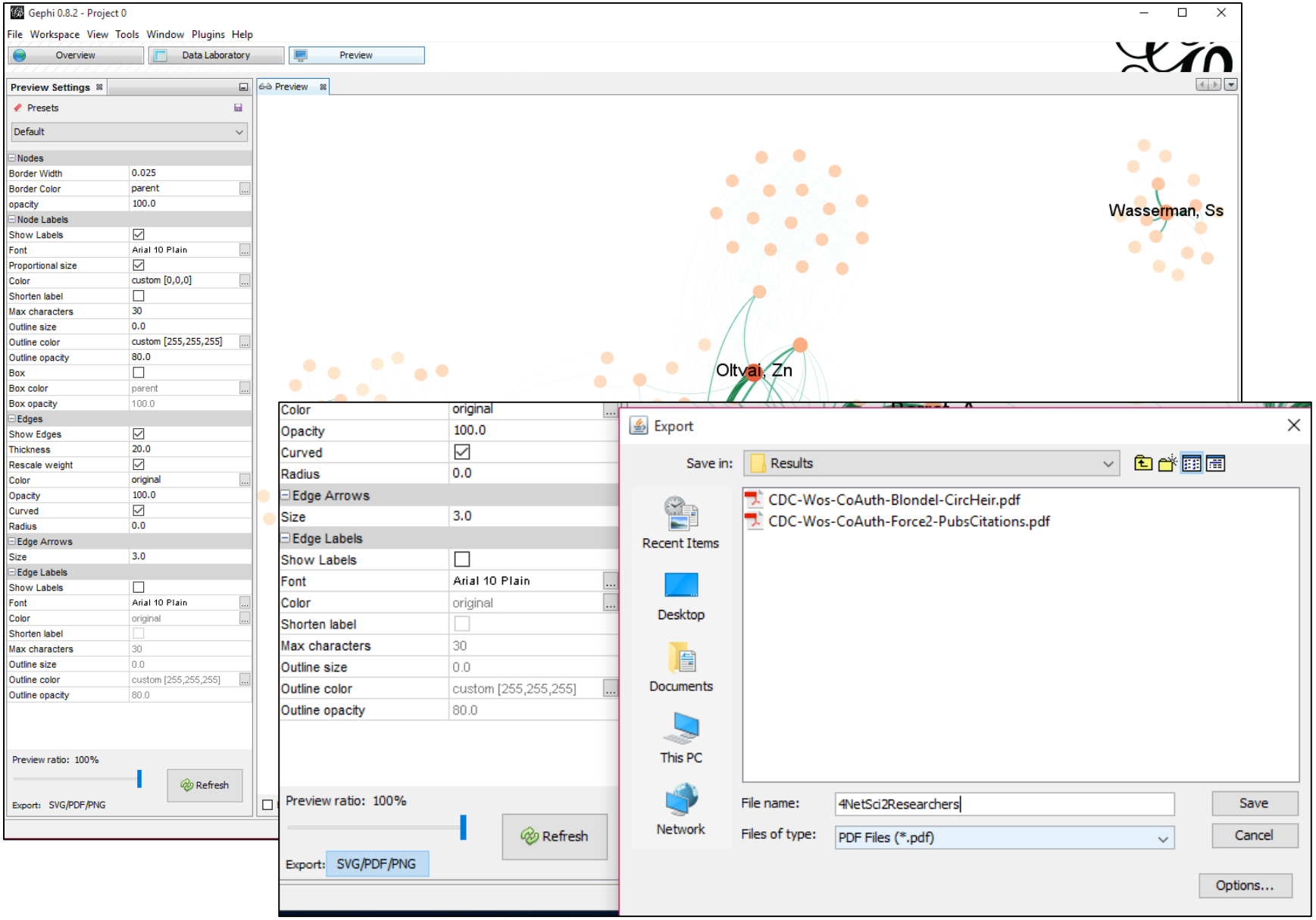
Layout algorithm



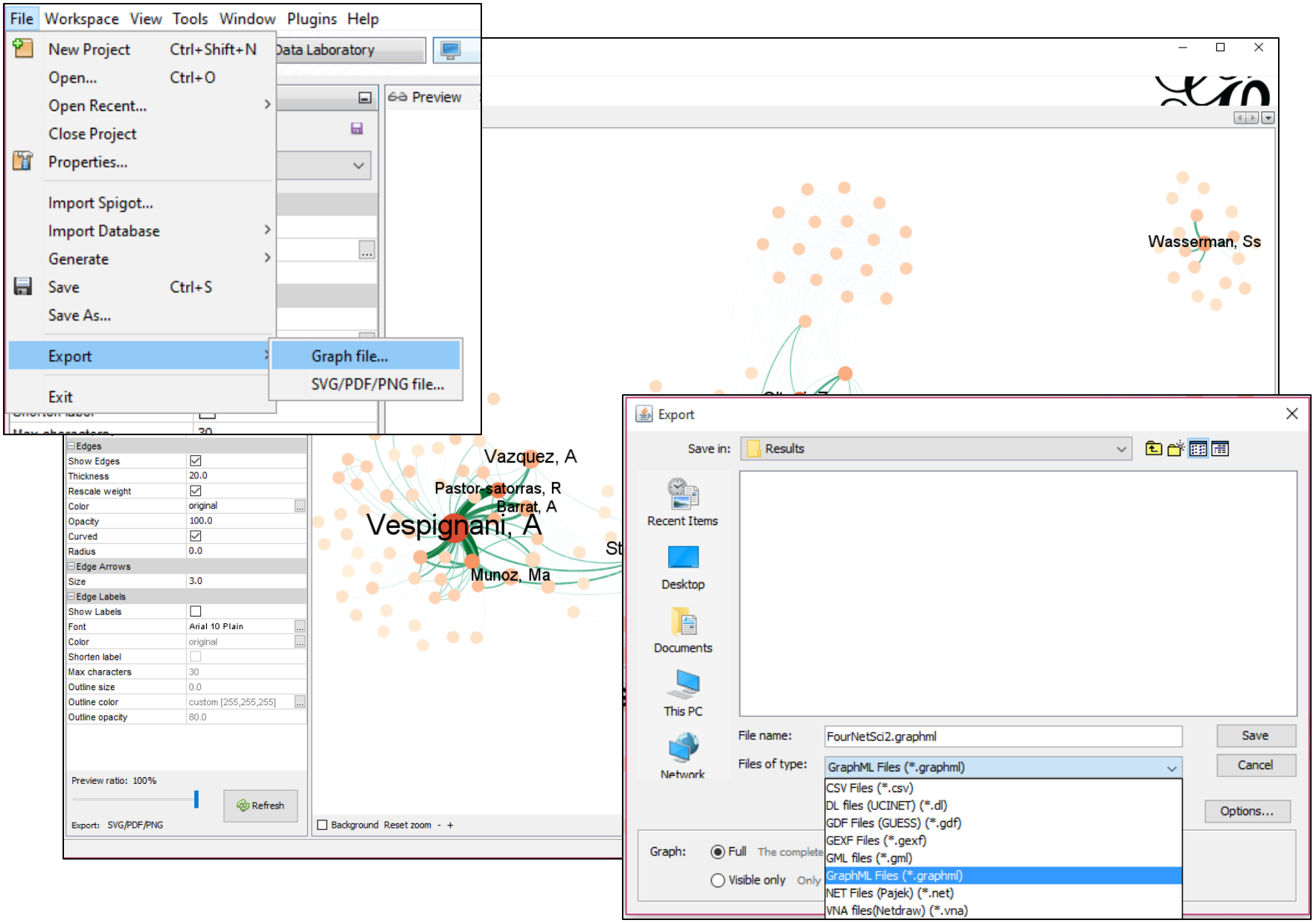
Manual adjustments



Gephi: Exporting the visualization and networks

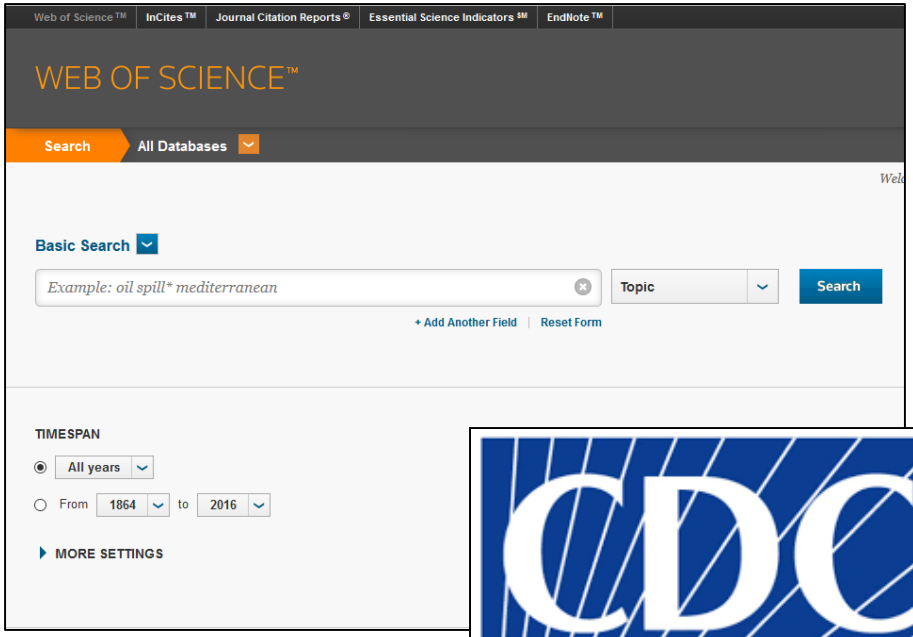


Gephi: Exporting the visualization and networks



Questions?

Co-Author Network Analysis: CDC Web of Science Publications – Top Cited Articles from 2004-2014



CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

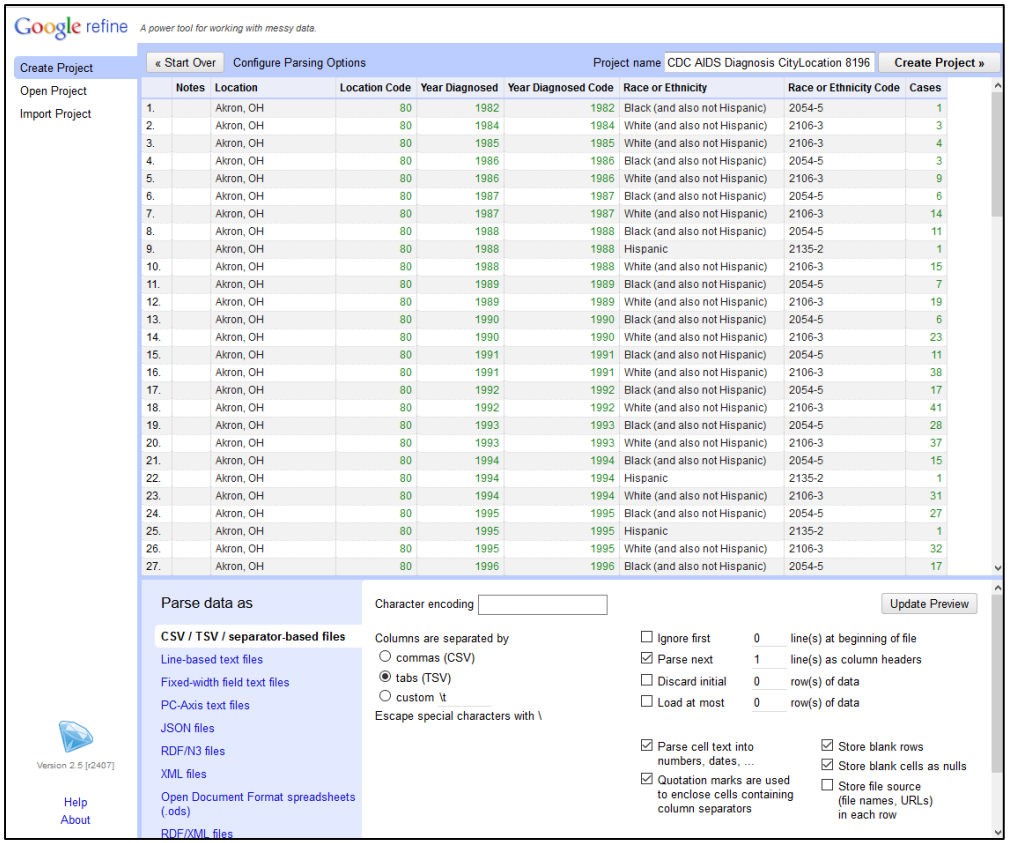
The data for this workflow was collected from **Web of Science,** which is a citation index database produced by Thomson Reuters.

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.

Open Refine: CDC Web of Science Publications – Top Cited Articles from 2004-2014



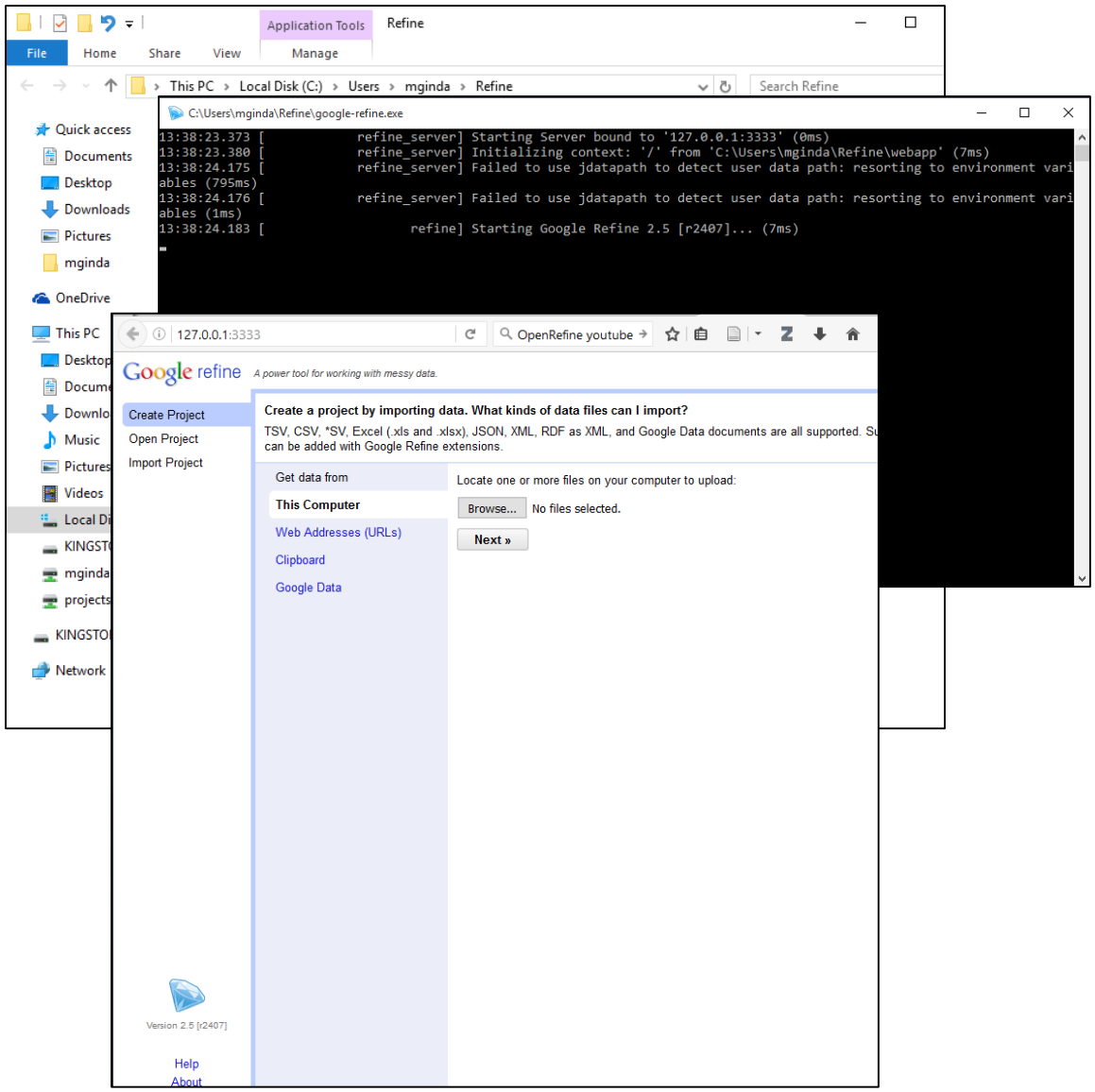
	Notes	Location	Location Code	Year Diagnosed	Year Diagnosed Code	Race or Ethnicity	Race or Ethnicity Code	Cases
1.		Akron, OH	80	1982	1982	Black (and also not Hispanic)	2054-5	1
2.		Akron, OH	80	1984	1984	White (and also not Hispanic)	2106-3	3
3.		Akron, OH	80	1985	1985	White (and also not Hispanic)	2106-3	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

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 Googles 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 the Google geocoder.

Open Refine: CDC Web of Science Publications – Top Cited Articles from 2004-2014



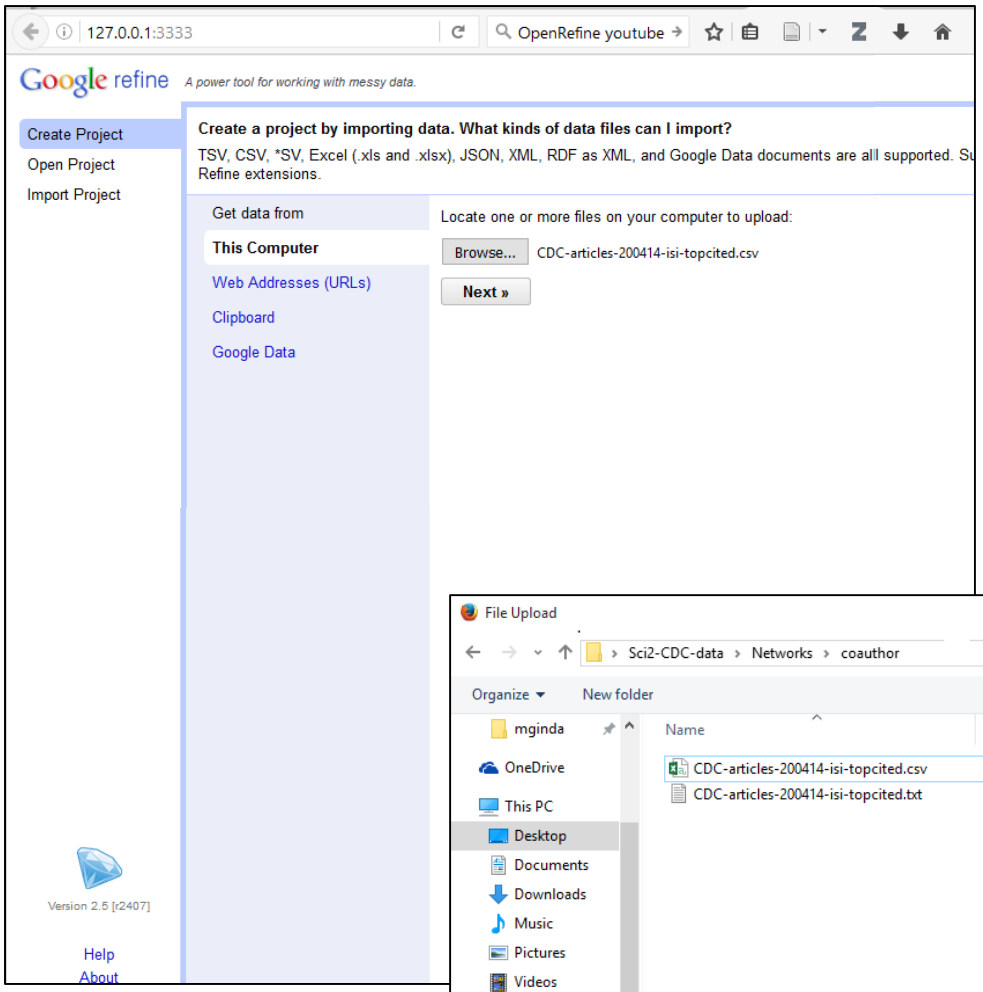
First we need to start OpenRefine. Navigate to the Desktop folder Refine.

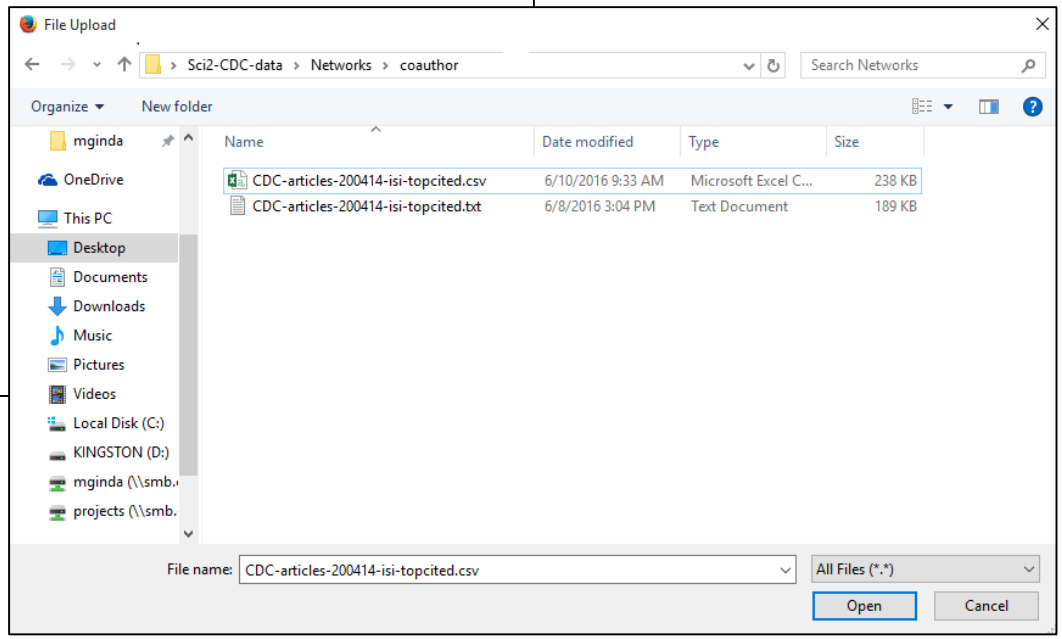
In the directory, select the executatable file “google-refine.exe”.

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

Open Refine: CDC Web of Science Publications – Top Cited Articles from 2004-2014 – Creating a Project



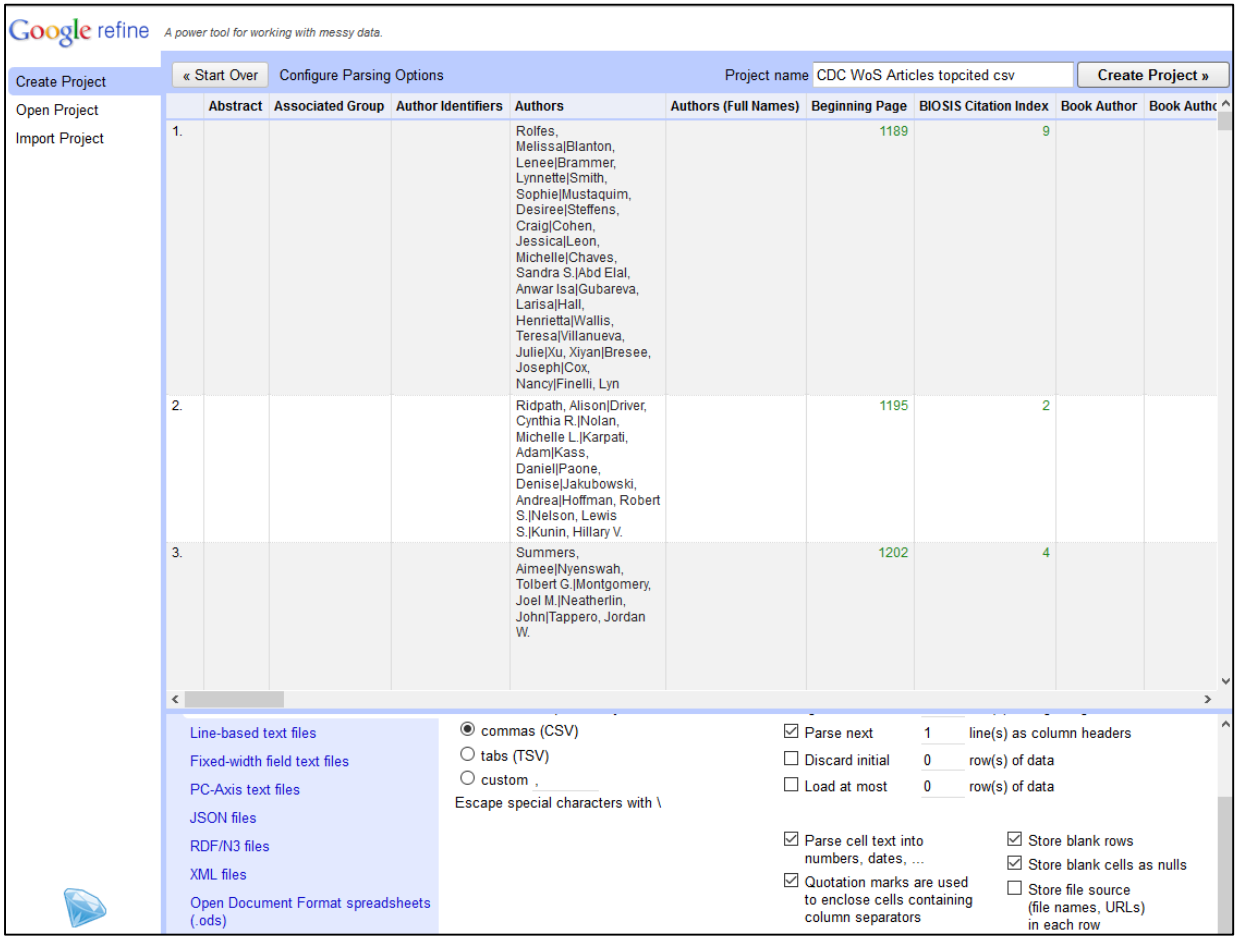


Once OpenRefine has started. You will want to create a project.

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.

Open Refine: CDC Web of Science Publications – Top Cited Articles from 2004-2014 – Creating a Project

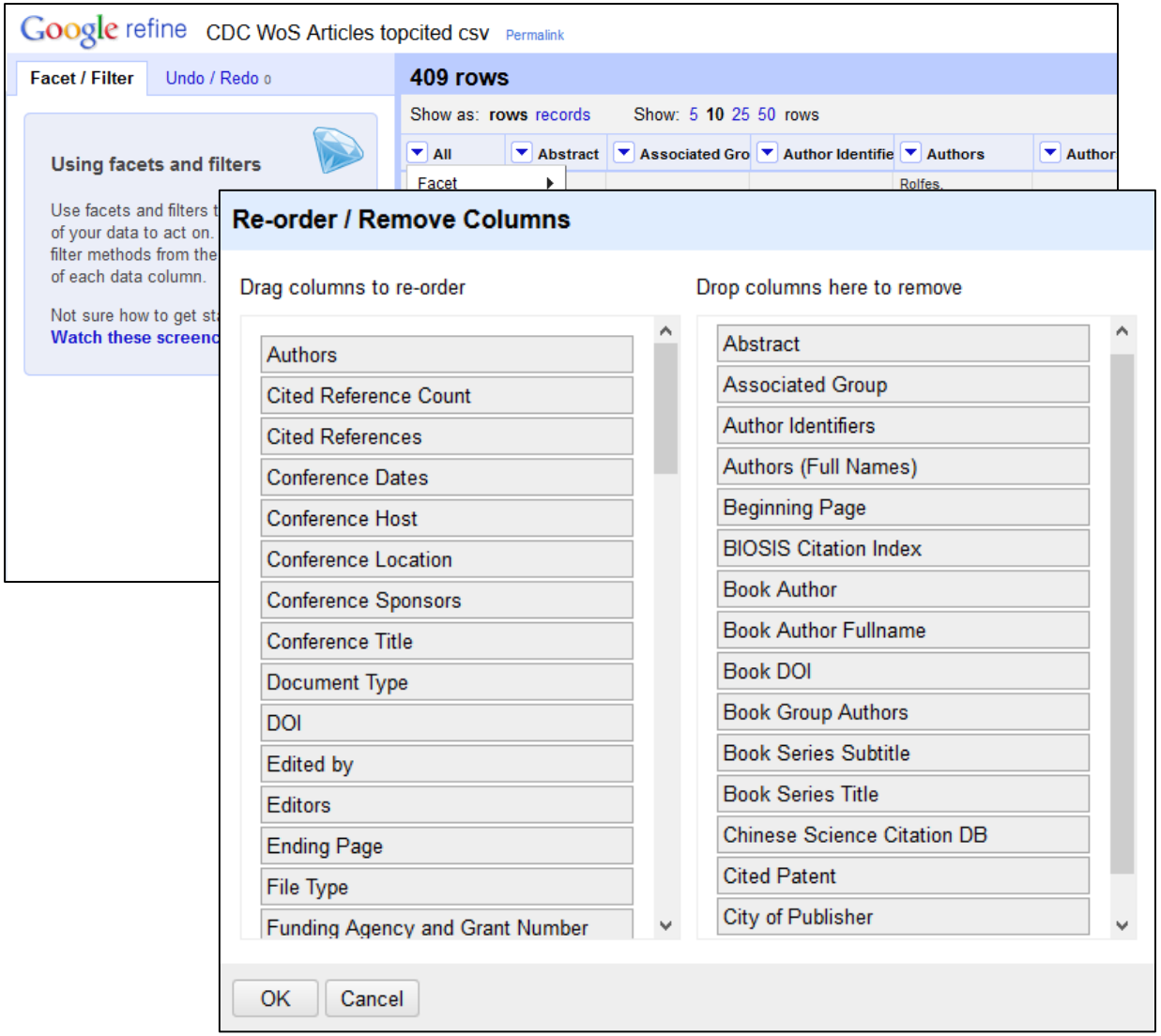


Once the data is loaded, it has to be parsed by OpenRefine. In this case, the data is in TSV format, and has column headers for the data.

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.

Open Refine: CDC Web of Science Publications – Top Cited Articles from 2004-2014 – Editing Columns In Bulk



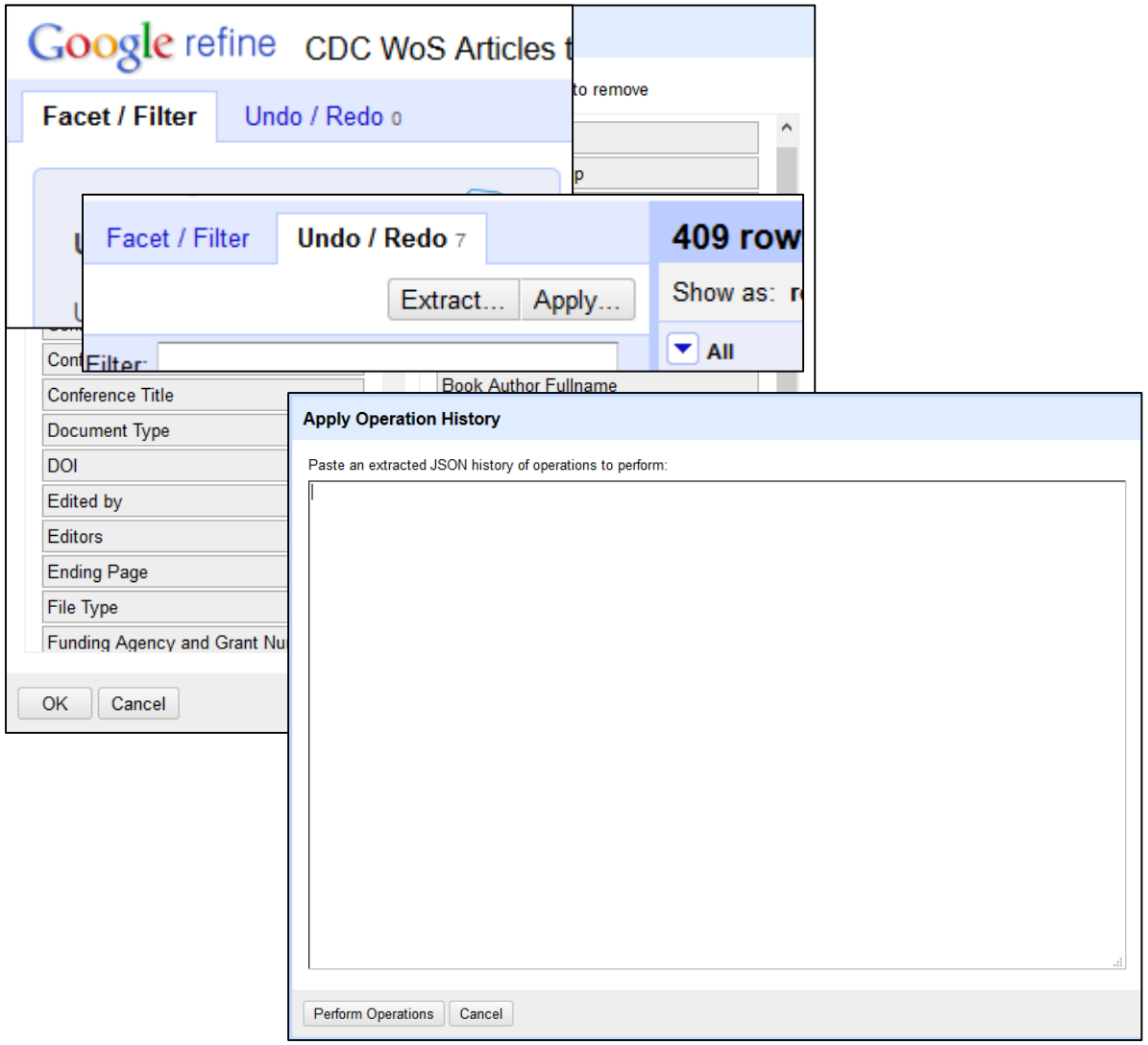
Data files downloaded from ISI in the plain text format, for which this CSV file was derived, contain many irrelevant columns to our analysis.

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.

Open Refine: CDC Web of Science Publications – Top Cited Articles from 2004-2014 – Editing Columns In Bulk

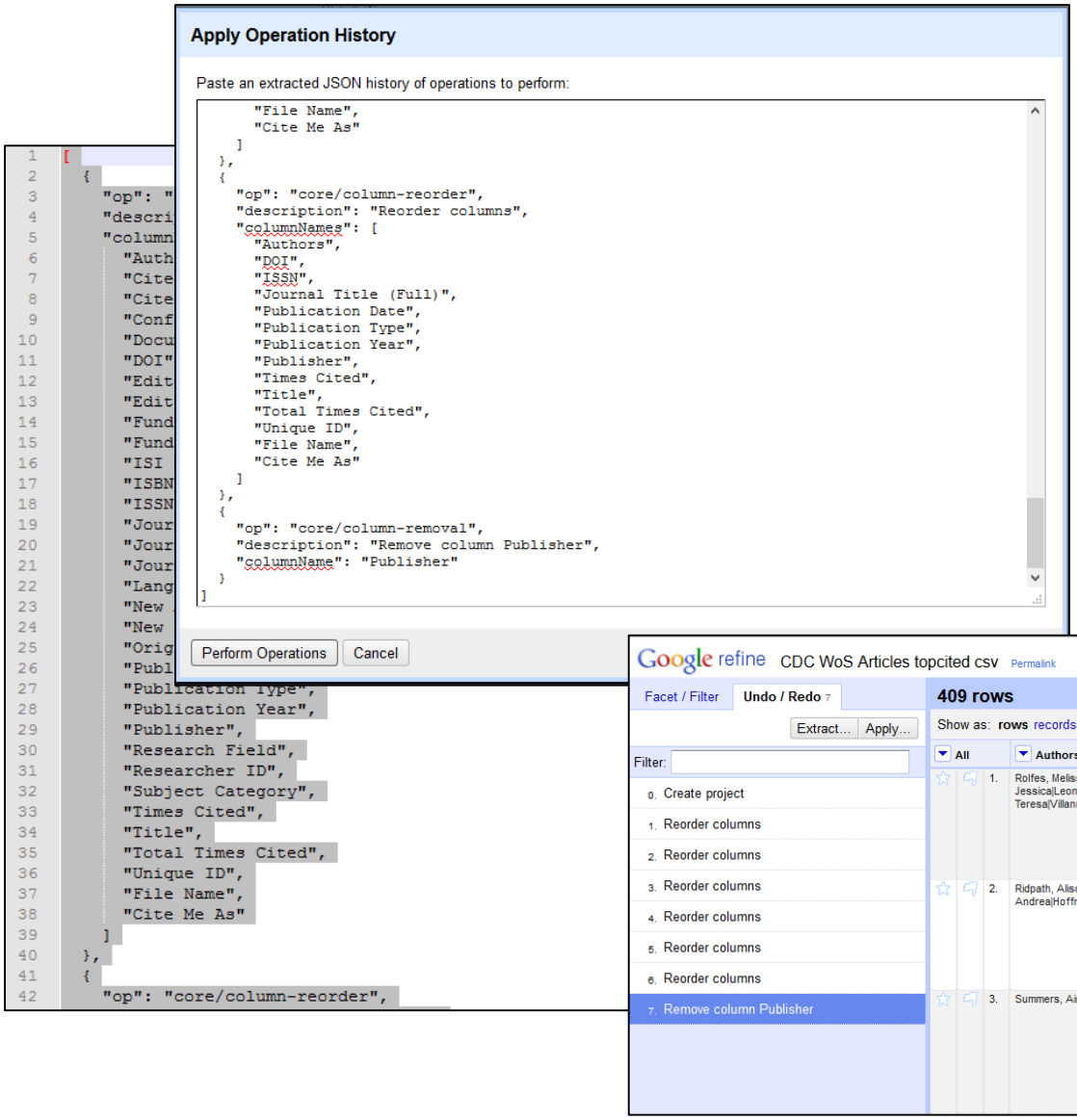


To speed up the process, in advance of the workshop, I have identified the all columns that need to be removed, and saved a JSON script that will allow you to duplicate this data preprocessing task.

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.

Open Refine: CDC Web of Science Publications – Top Cited Articles from 2004-2014 – Editing Columns In Bulk

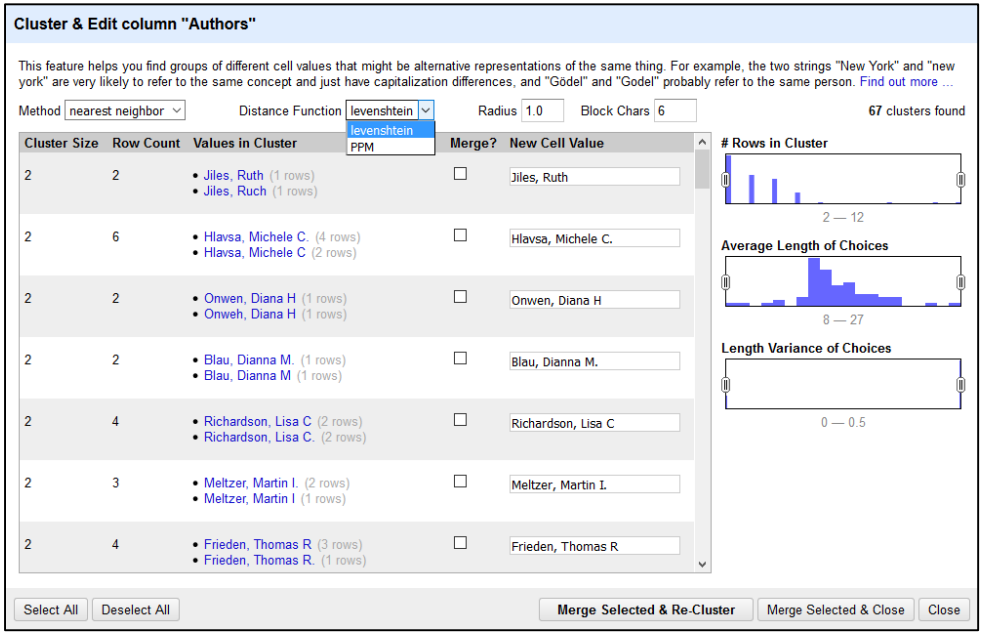


To apply the re-order tasks, navigate to **C:/…/Sci2-CDC-data > Networks > coauthor > refine > CDC-WoS-JSON-1-Preprocessing.txt**, and open the file in Notepad.

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

Then select the **Perform Operations** button.

Open Refine: CDC Web of Science Publications – Top Cited Articles from 2004-2014 – Clustering and Editing



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."

Open Refine: CDC Web of Science Publications – Top Cited Articles from 2004-2014 – Clustering and Editing

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.

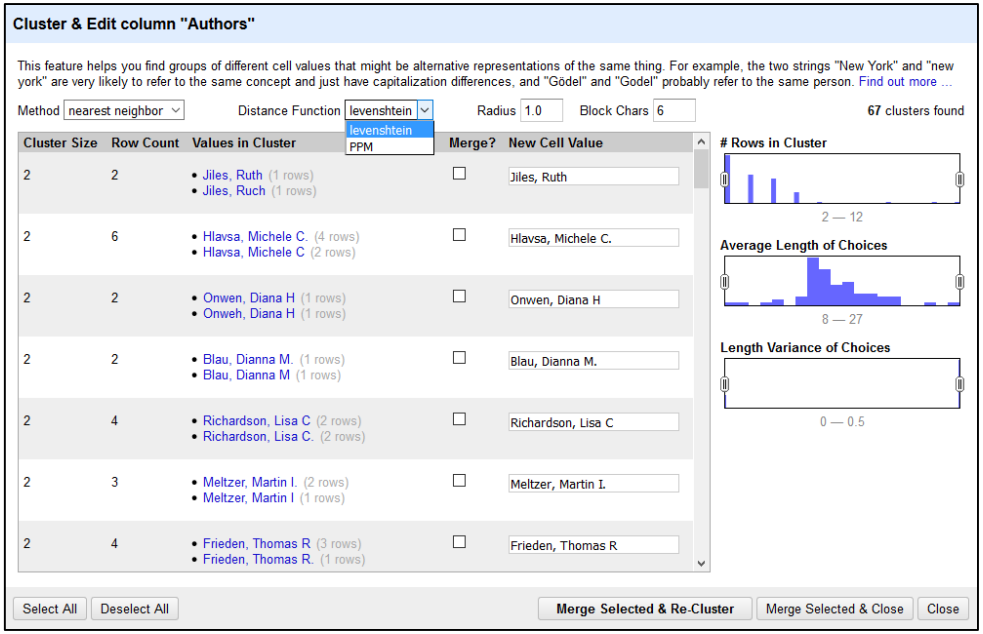
OpenRefine notes in its documentation:

> "…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."

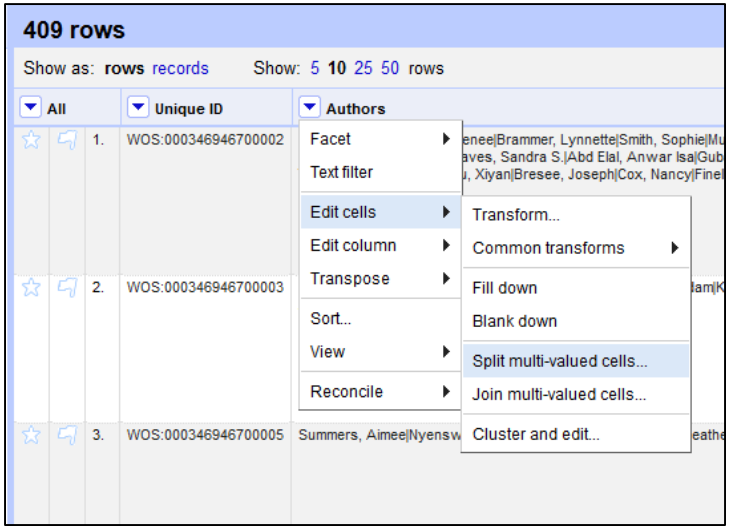
The next, we'll walk through preparing the data for clustering and editing a text field.

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.

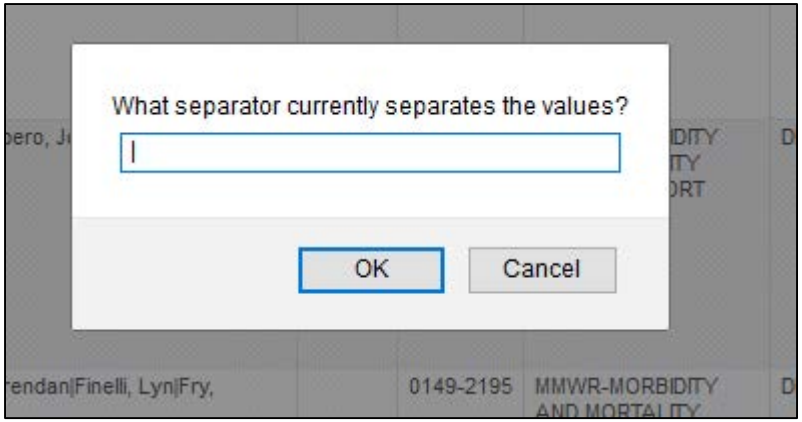


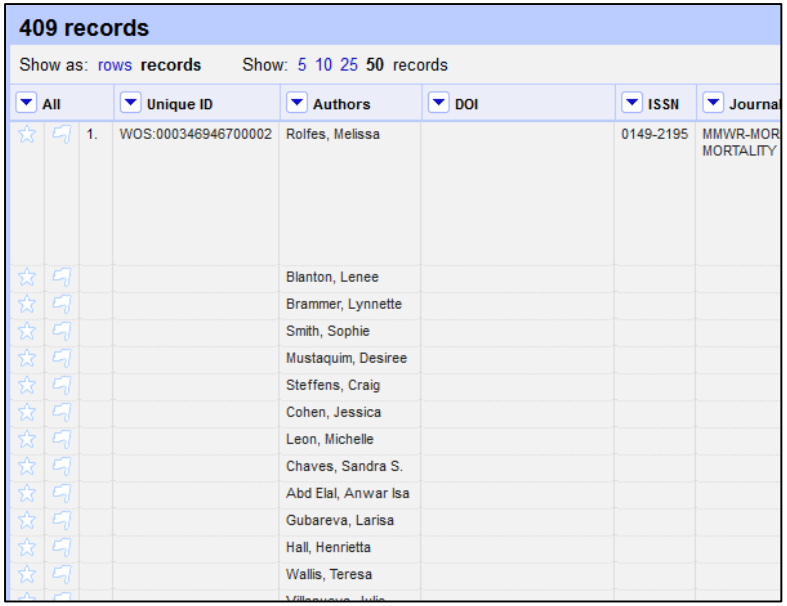
Open Refine: CDC Web of Science Publications – Top Cited Articles from 2004-2014 – Clustering and Editing



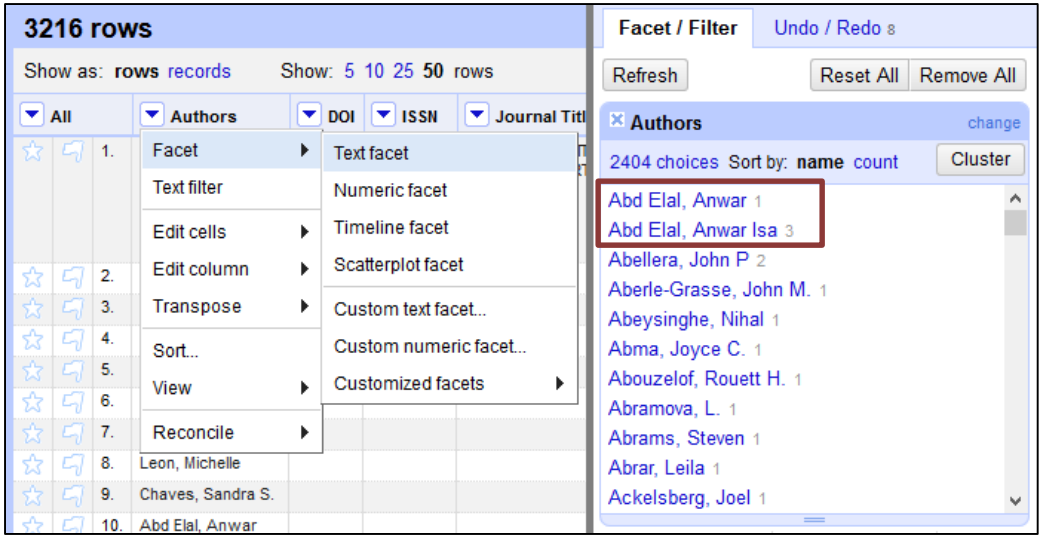
In OpenRefine, first navigate to the **Authors** column drop down menu **Edit cells > Split multi-valued cells…**

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





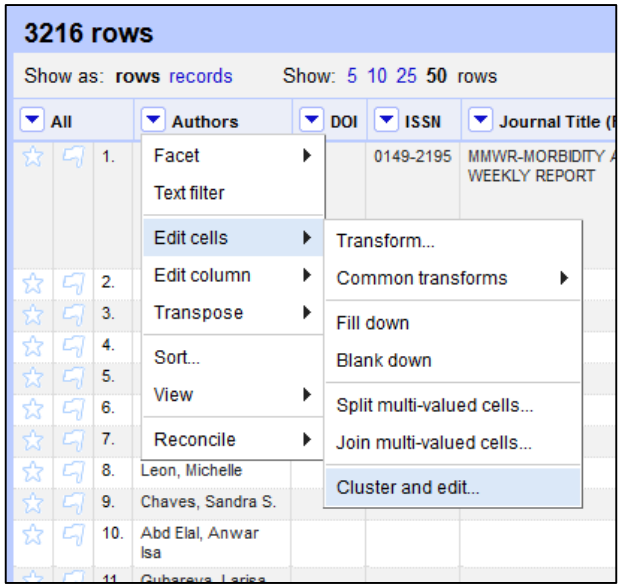
Open Refine: CDC Web of Science Publications – Top Cited Articles from 2004-2014 – Clustering and Editing



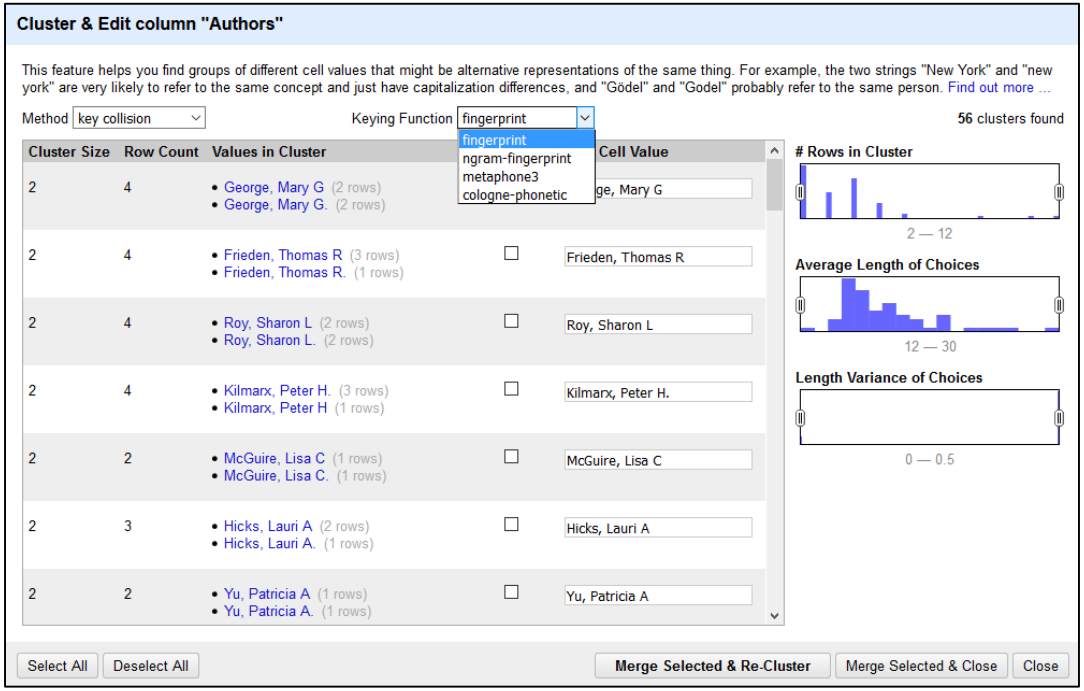
To track the effects of your merges, it is nice to have a Text Facet of the field you are editing.

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…**



Open Refine: Clustering and Editing – Key Collision



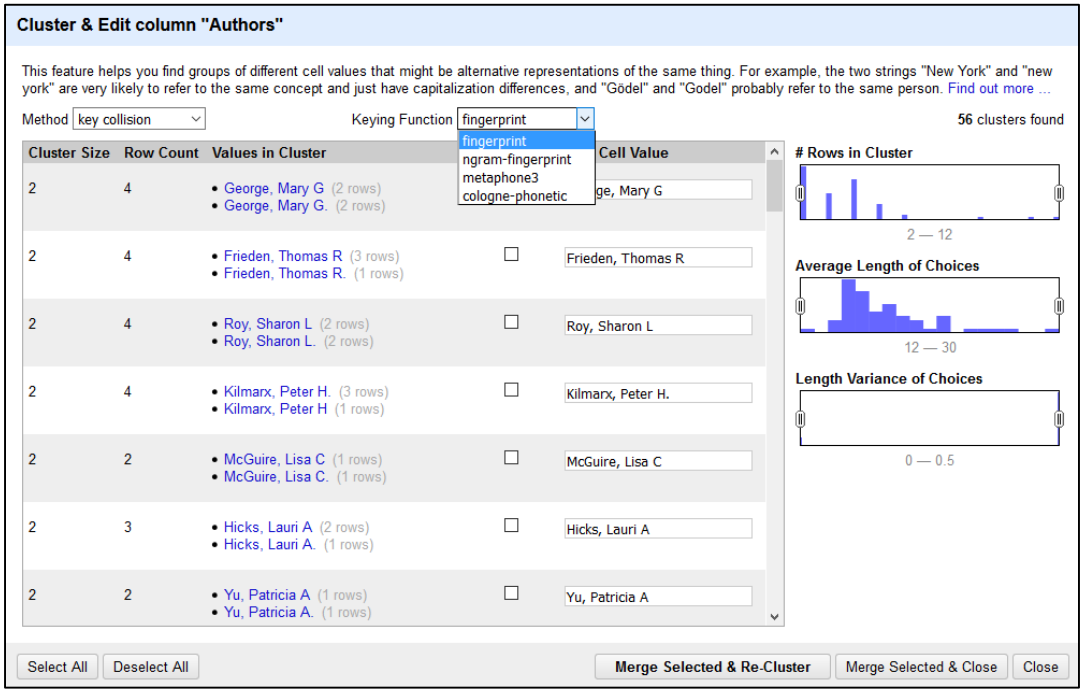
Key Collision

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.

Open Refine: Clustering and Editing – Key Collision



Fingerprint

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

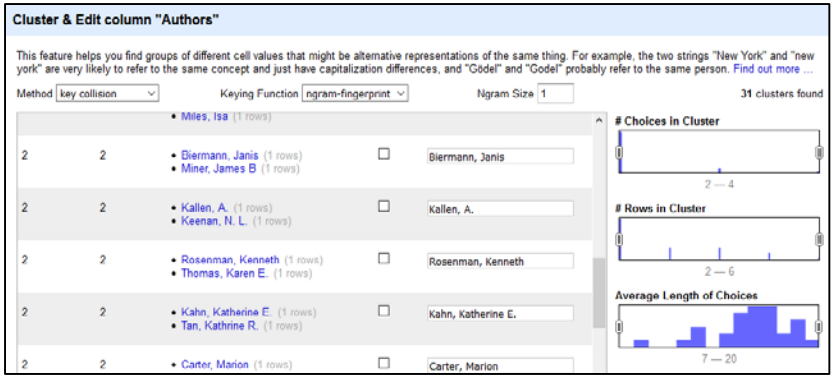
The process that generates the key from a string value is the following (note that the order of these operations is significant):

- 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")

Open Refine: Clustering and Editing – Key Collision

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

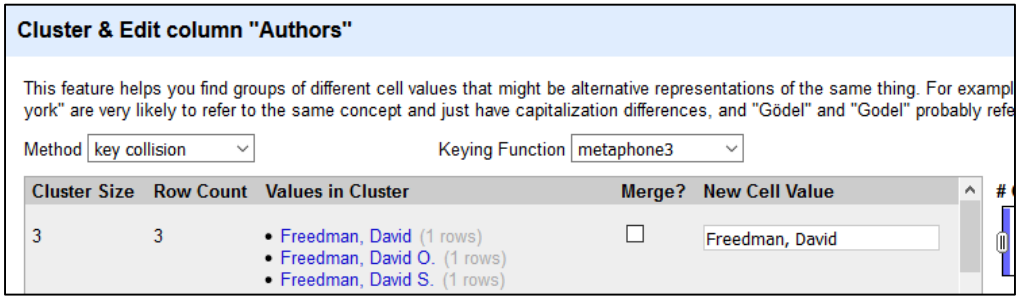
So, for example, the 2-gram fingerprint of "Paris" is "arispari" and the 1-gram fingerprint is "aiprs".

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.

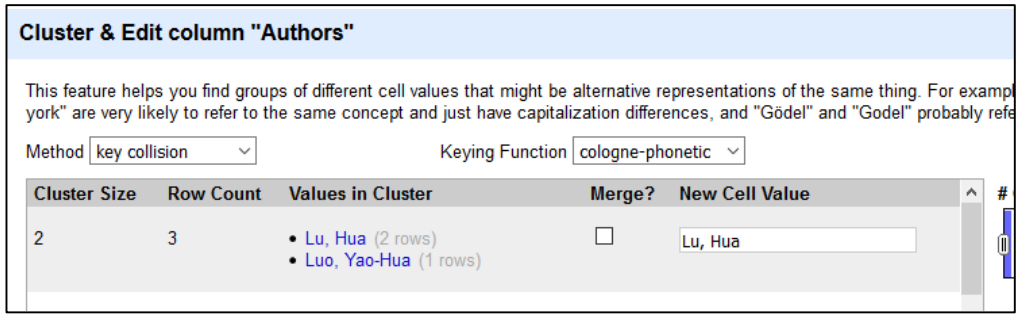
Open Refine: Clustering and Editing – Key Collision

Phonetic fingerprint

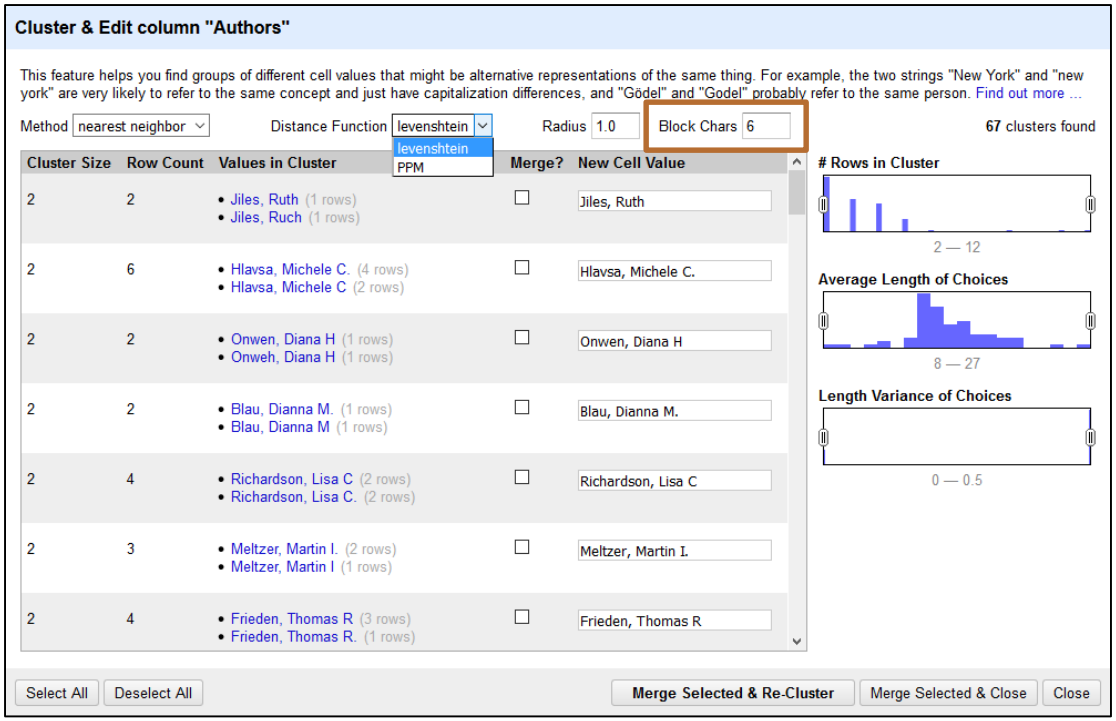
A third keying method uses a phonetic fingerprinting (specifically, Metaphone3 method for English and the Cologne phonetic keyer for German), which is a way to transform tokens into the way they are pronounced.



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.



Open Refine: Clustering and Editing – Nearest Neighbors



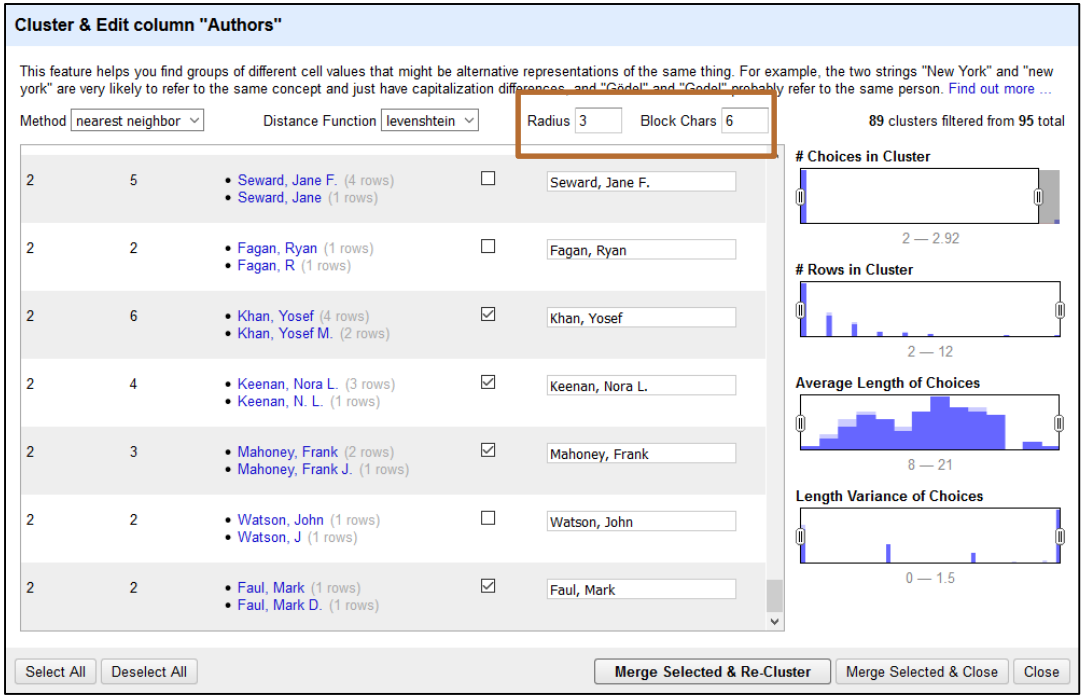
Nearest Neighbor

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.)

Open Refine: Clustering and Editing – Nearest Neighbors



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.

Open Refine: Clustering and Editing – Nearest Neighbors

PPM

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

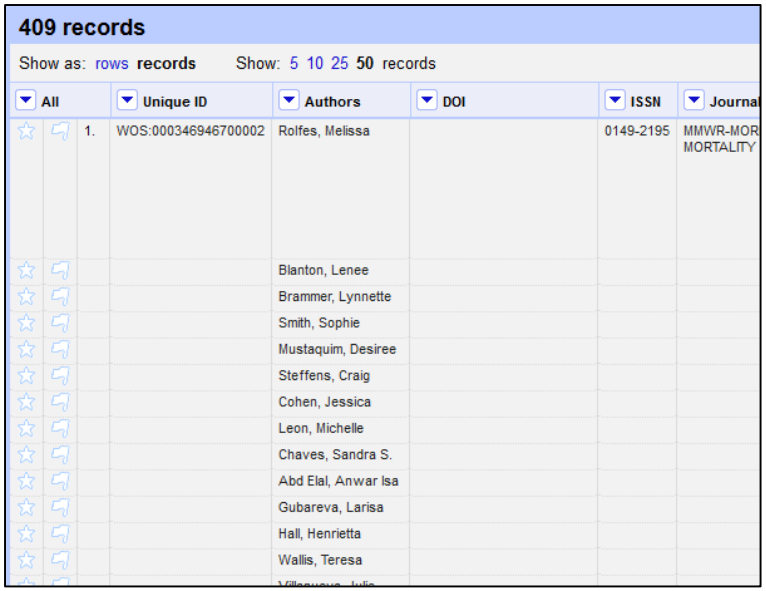
The algorithm text compressors estimate the information content of two strings to tell if they are identical.

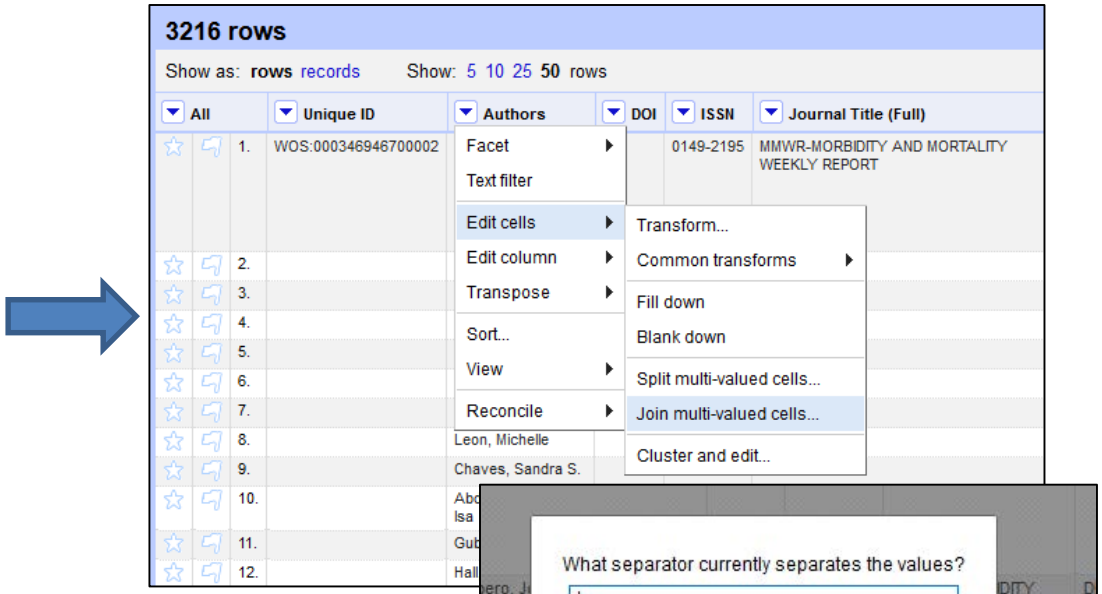
OpenRefine implements a normalized version:

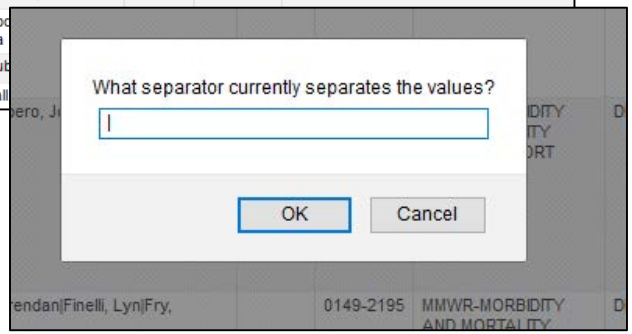
distance(A,B) = comp(A+B) + comp(B+A) / (comp(A+A) + comp(B+B))

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.

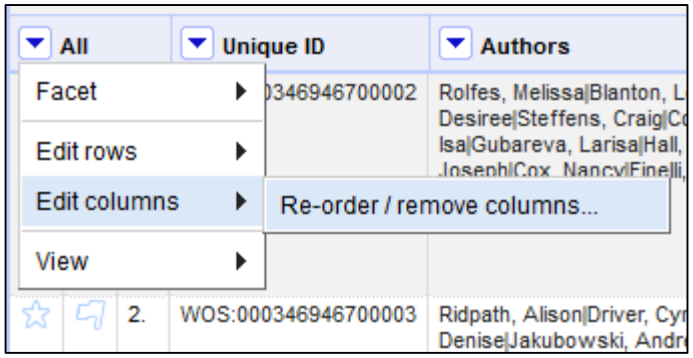
Open Refine: Post-Cluster & Editing data processing



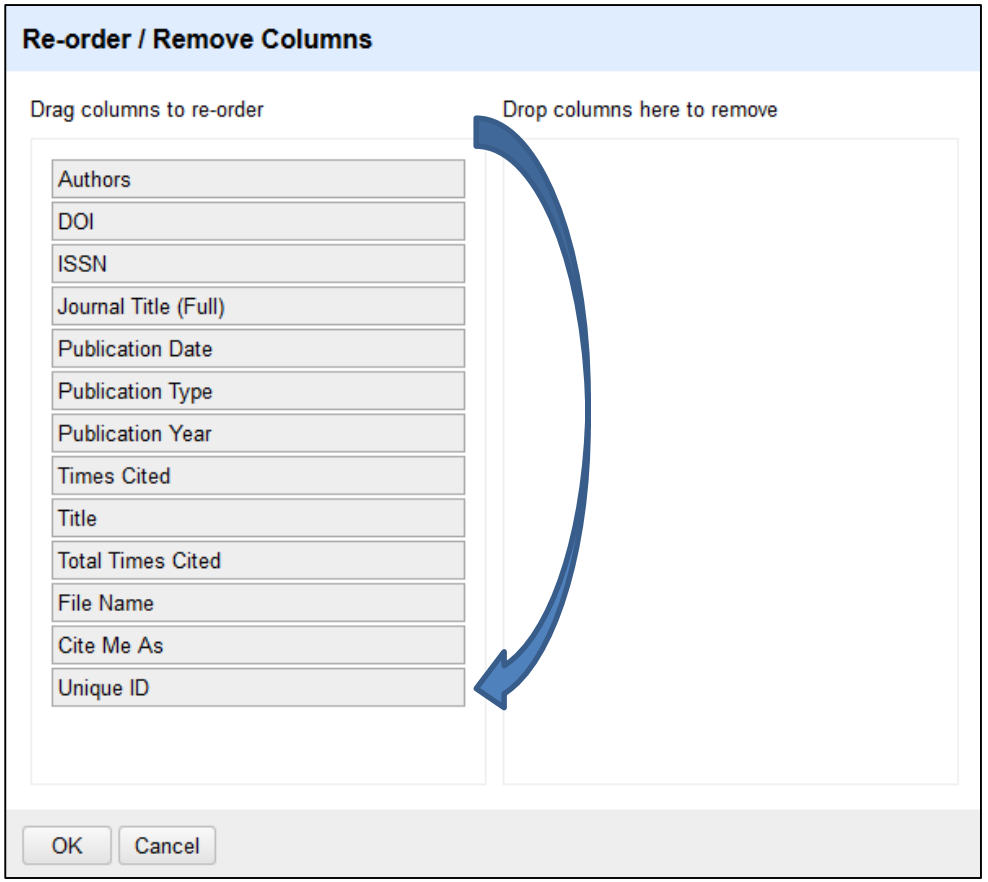




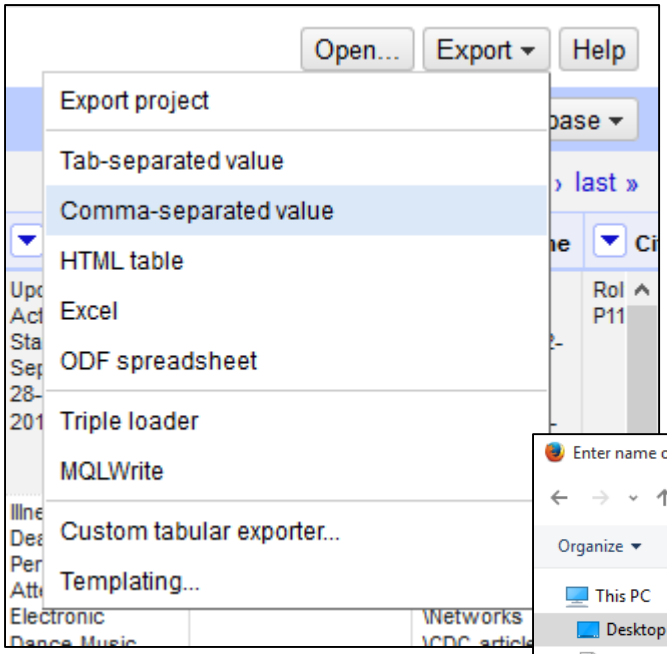
Open Refine: Post-Cluster & Editing data processing



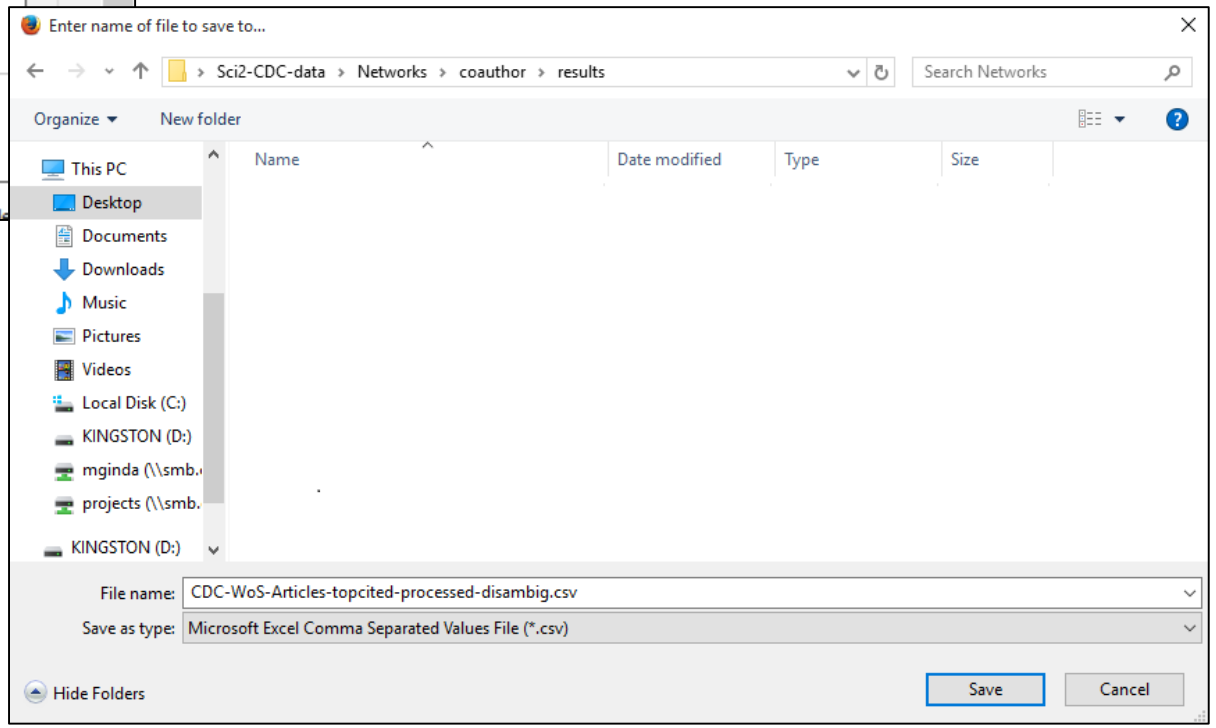
Next, reorder the remaining columns by moving **Unique ID** to the end of the list.



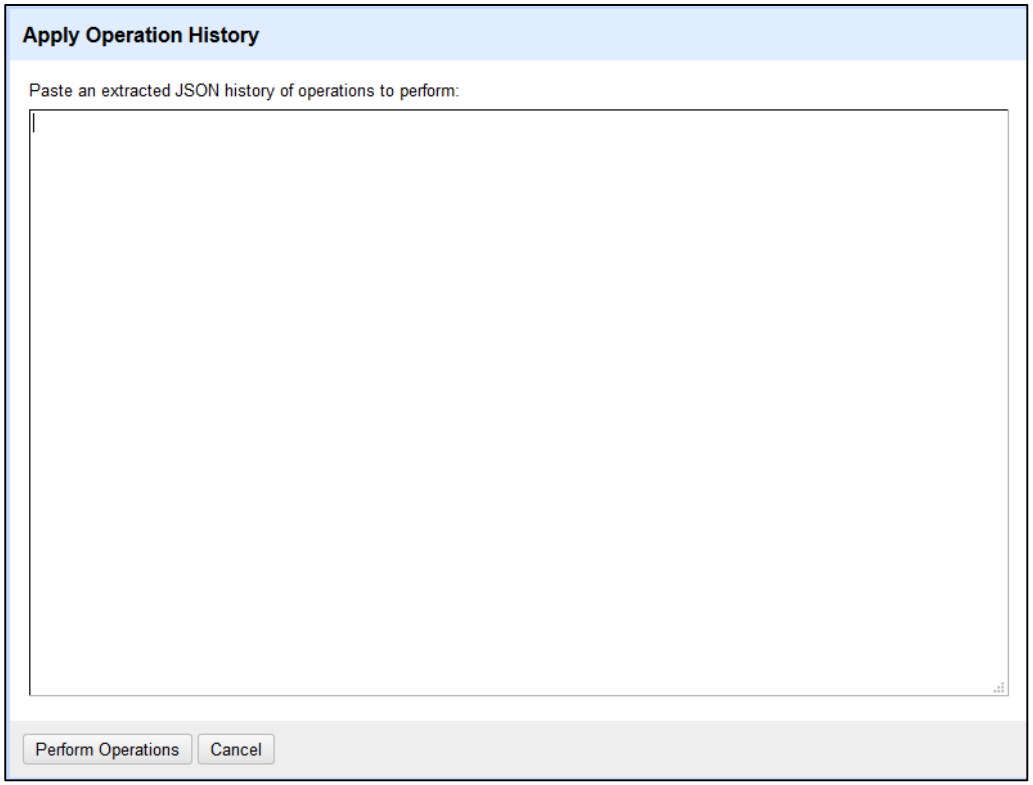
Open Refine: Exporting of CDC Web of Science Publications



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.



Open Refine: Cluster & Editing the Authors column in bulk

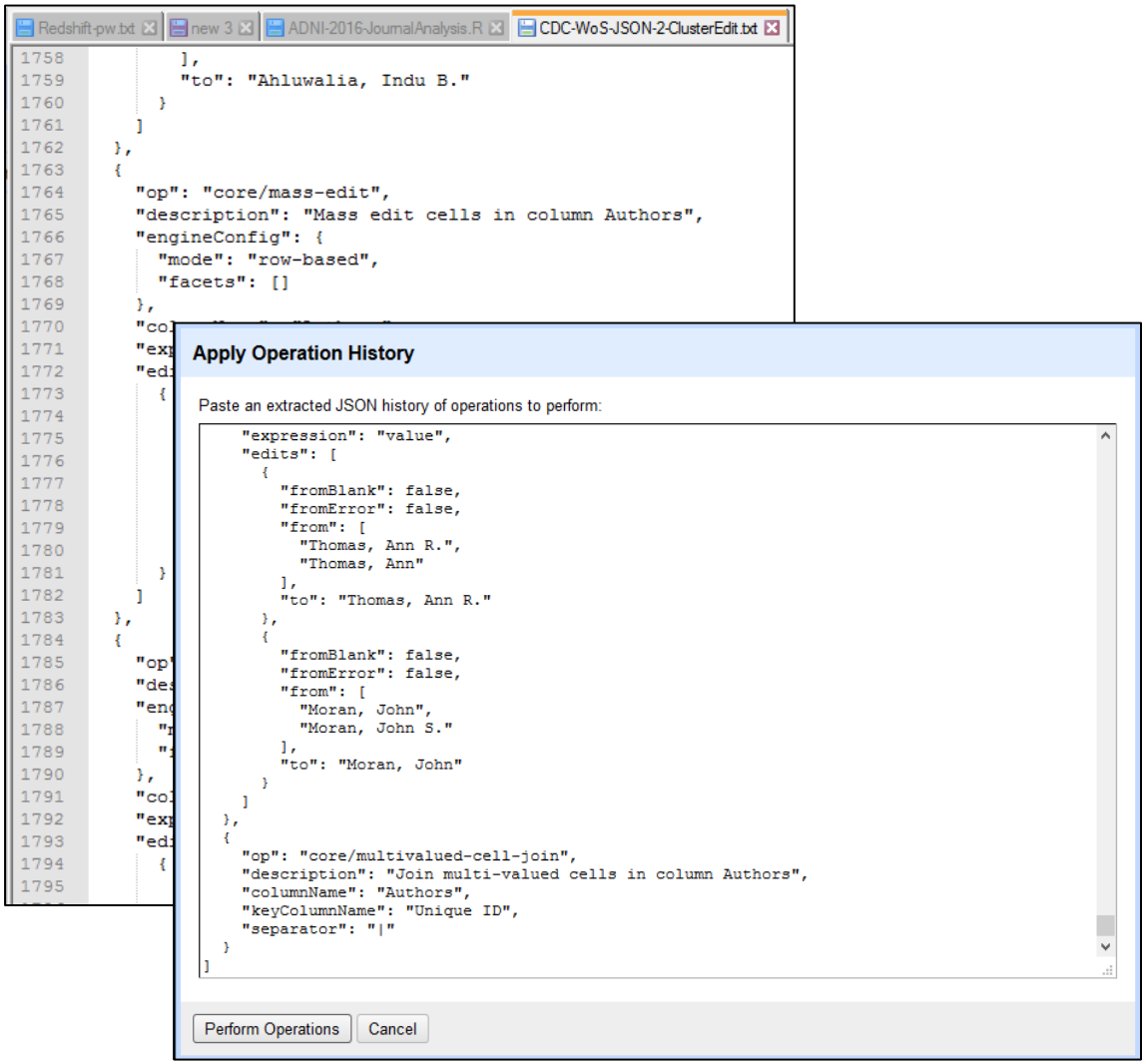


To speed up the process, in advance of the workshop, I have identified the all columns that need to be removed, and saved a JSON script that will allow you to duplicate this data preprocessing task.

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.

Open Refine: Cluster & Editing the Authors column in bulk



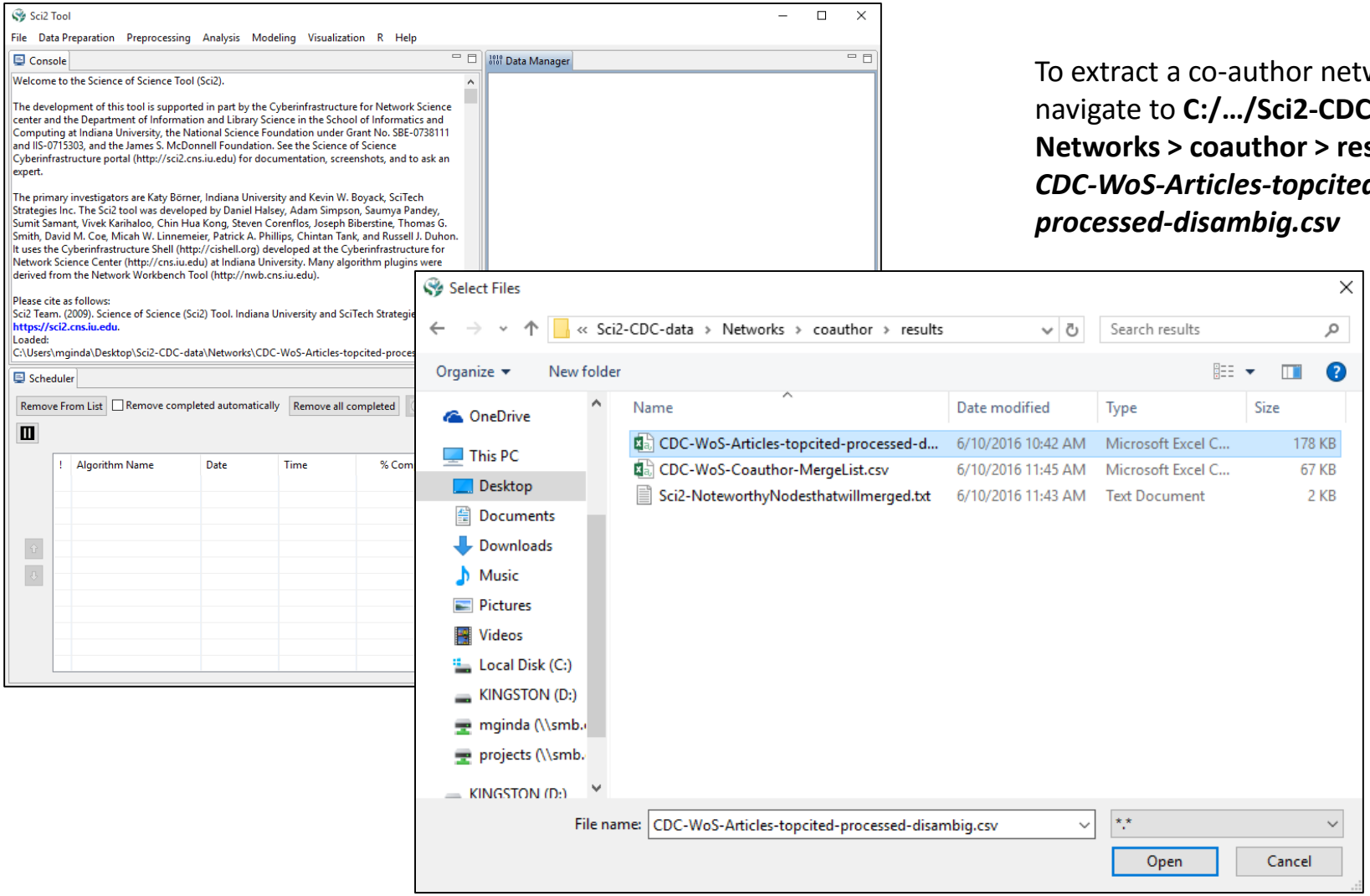
To apply the re-order tasks, navigate to **C:/…/Sci2-CDC-data > Networks > coauthor > refine > CDC-WoS-JSON-2-ClusterEdit.txt**, and open the file in Notepad.

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

Then select the **Perform Operations** button.

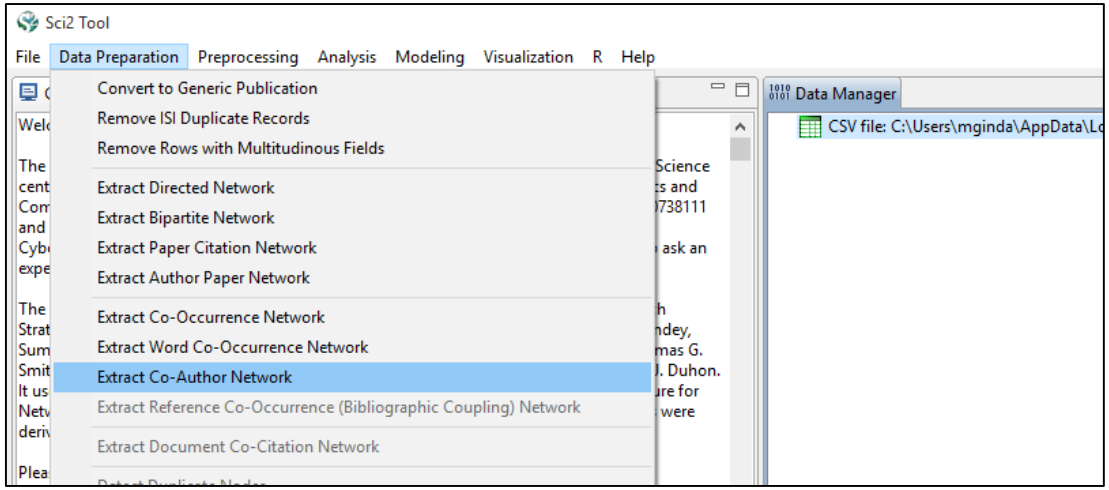
Questions?

Sci2: Co-Author Network Analysis



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

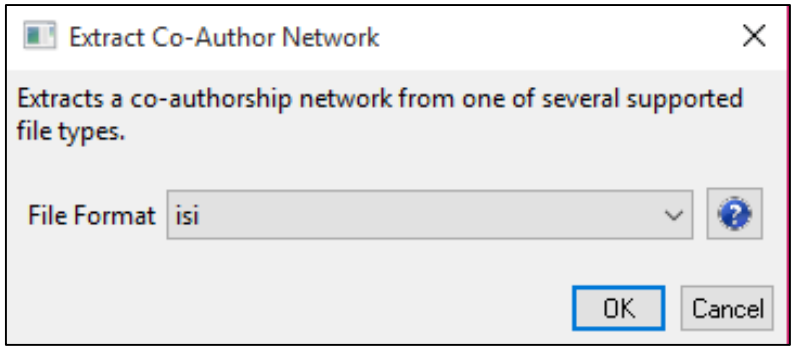
Sci2: Extracting a co-author network

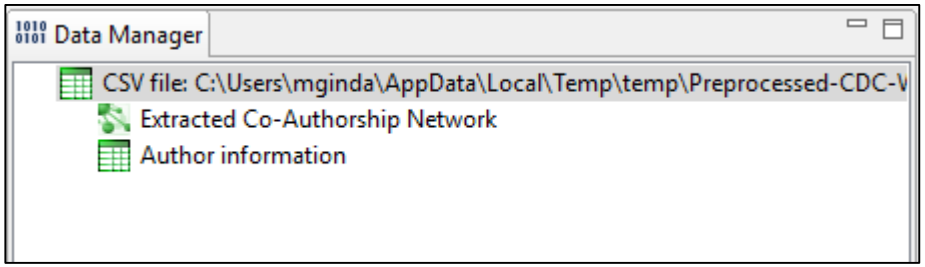


Select the data CDC articles in the Data manager and the then navigate in the menu ***Data Preparation > Extract Co-Author Network.***

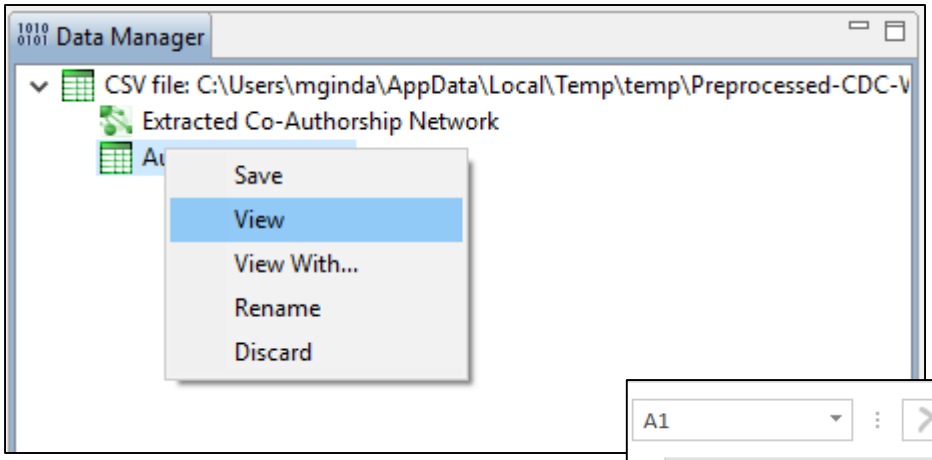
A pop-up window will appear; select the format ISI.

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





Sci2: Viewing the Authors list

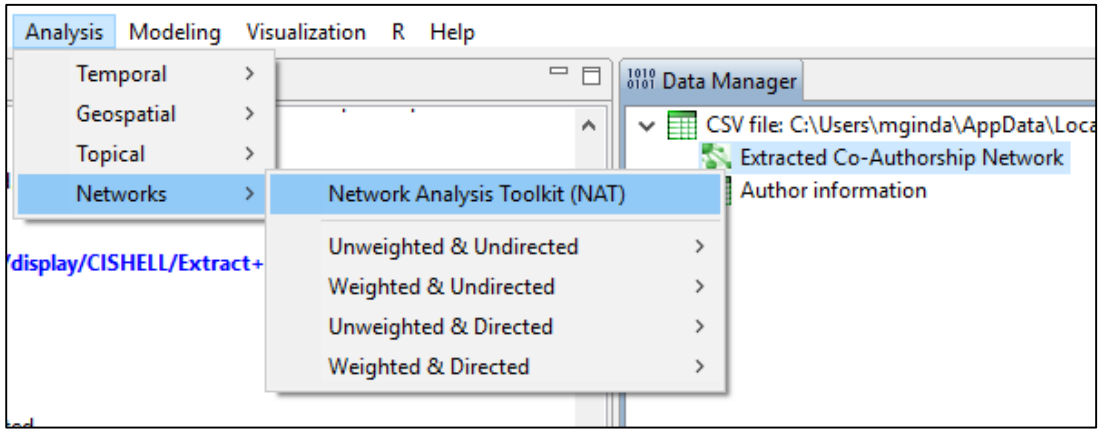


To see the resulting node variables calculated for each author in the data set, right click on the Author list data table in the Data Manager, and select **View** or **View with…**

A1 | label

	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	*

Sci2: Applying the Network Analysis Toolkit

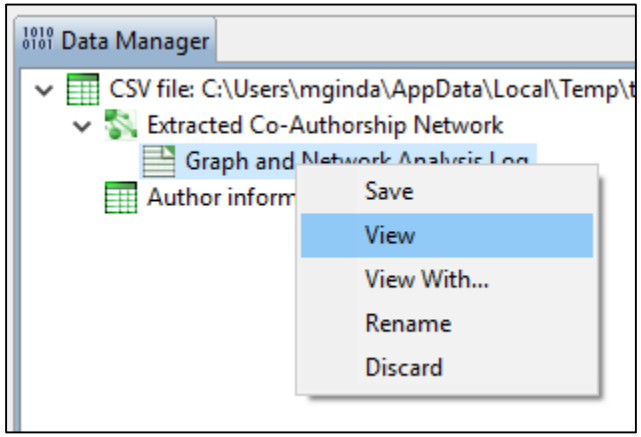


After extracting a network, it is good practice to get the initial statistics for the network, before further processing and analysis.

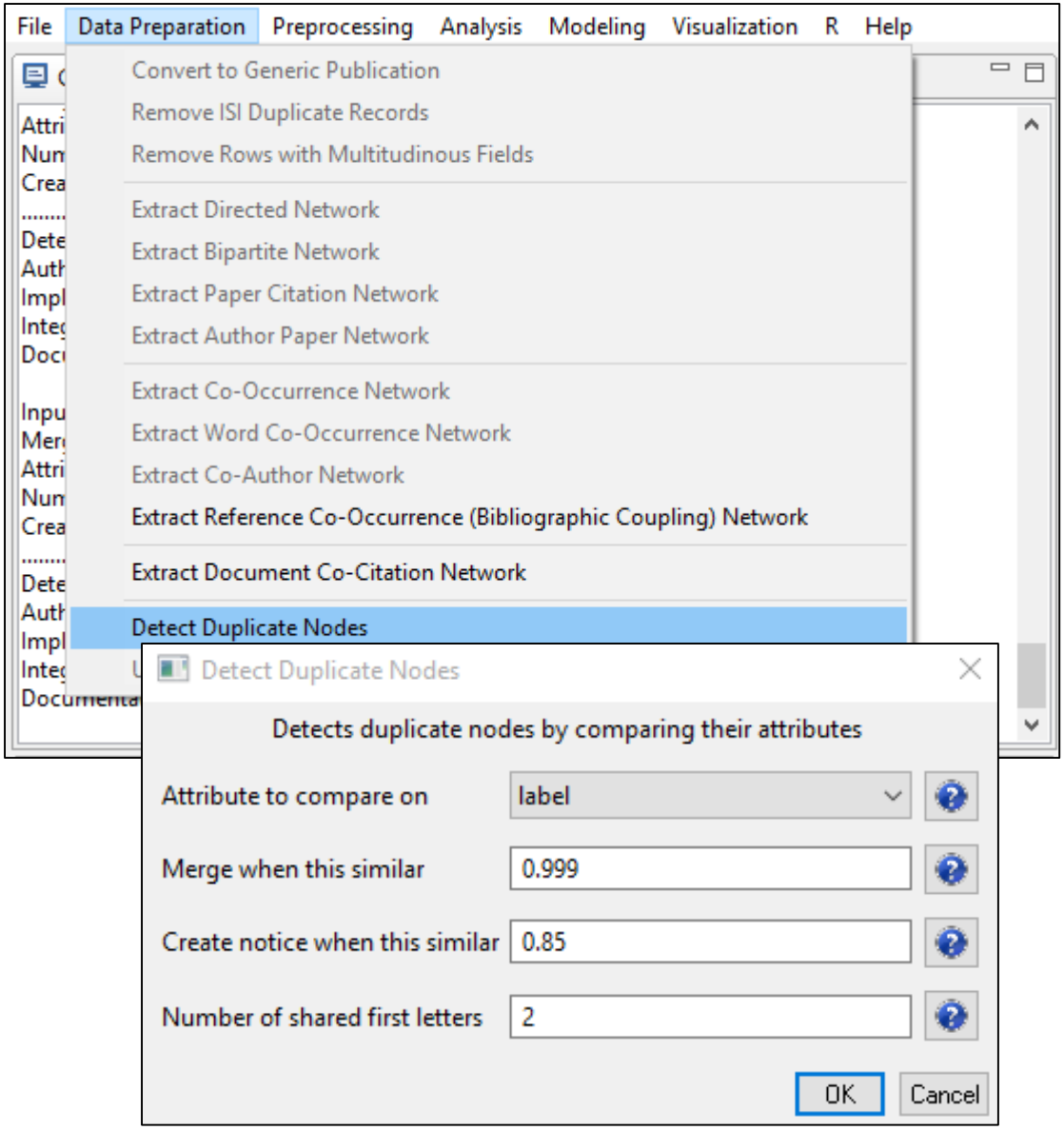
Navigate to **Analysis > Networks > Network Analysis Toolkit (NAT).**

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:
                                min     max     mean
                number_...      1       10      1.10811
                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
```



Sci2: Detecting Duplicate Nodes in Sci2



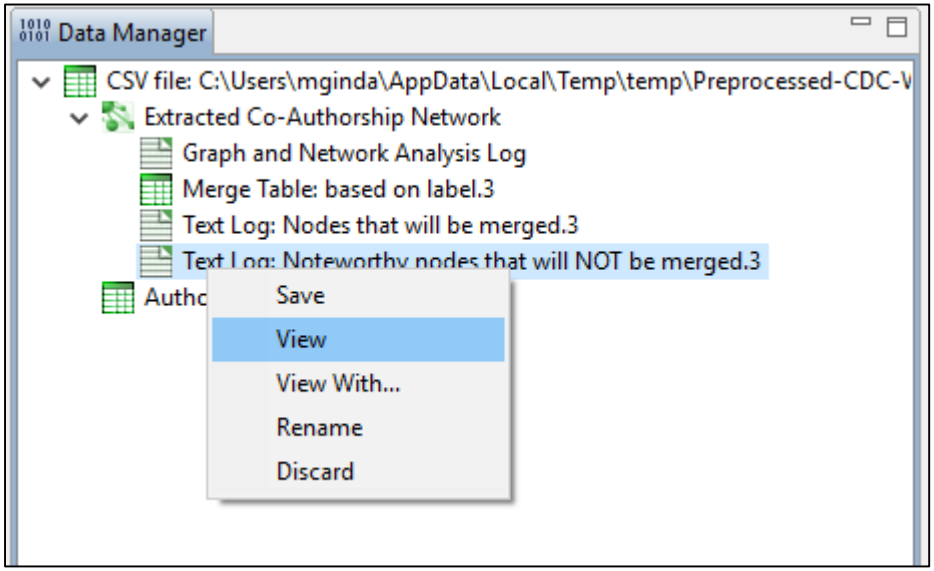
OpenRefine found a majority of the duplicate nodes in the network quickly. However, Sci2 also permits a simple duplicate node detection algorithm.

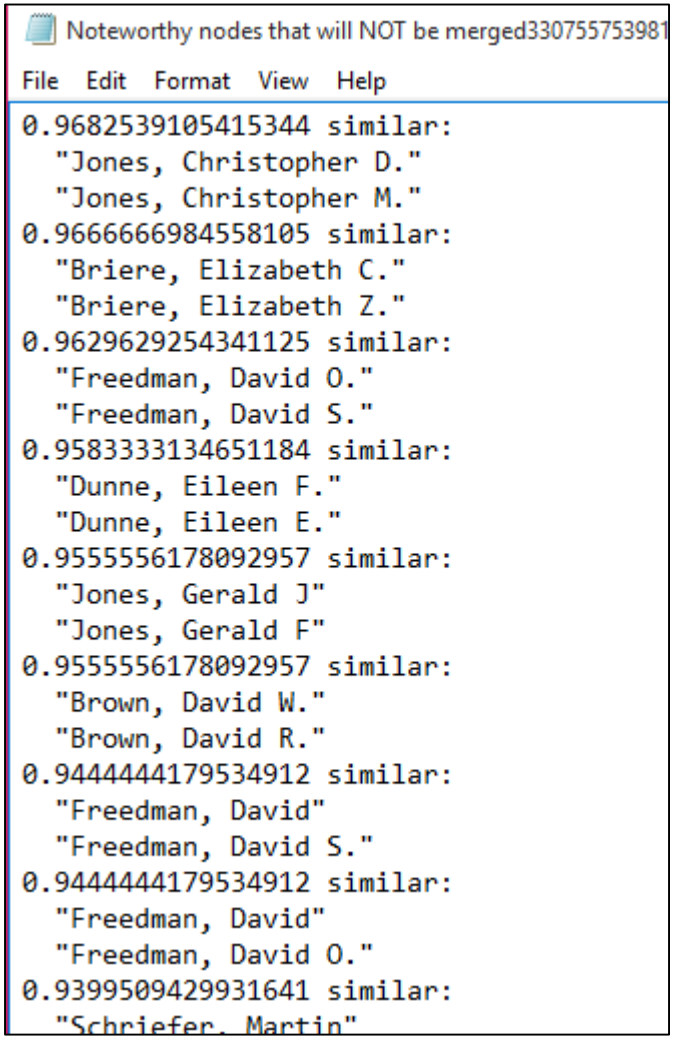
Select the network file in the data manager, and then navigate to **Data Preparation > Detect Duplicate Nodes** algorithm.

A pop-up box will let you set initial parameters. These are set to create three file: 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.

Sci2: Detecting Duplicate Nodes in Sci2





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**

Sci2: Updating a merge list

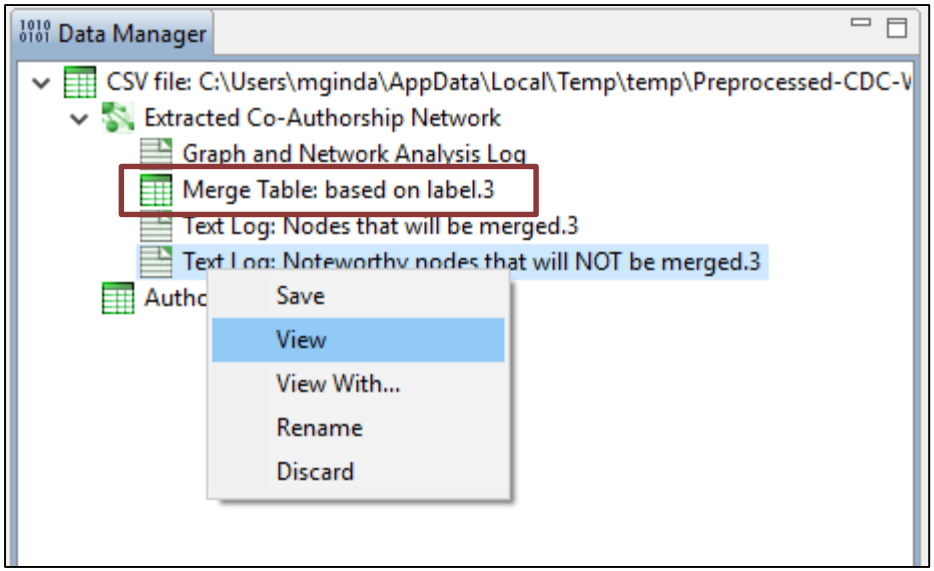
Noteworthy nodes that will NOT be merged330755753981

File Edit Format View Help

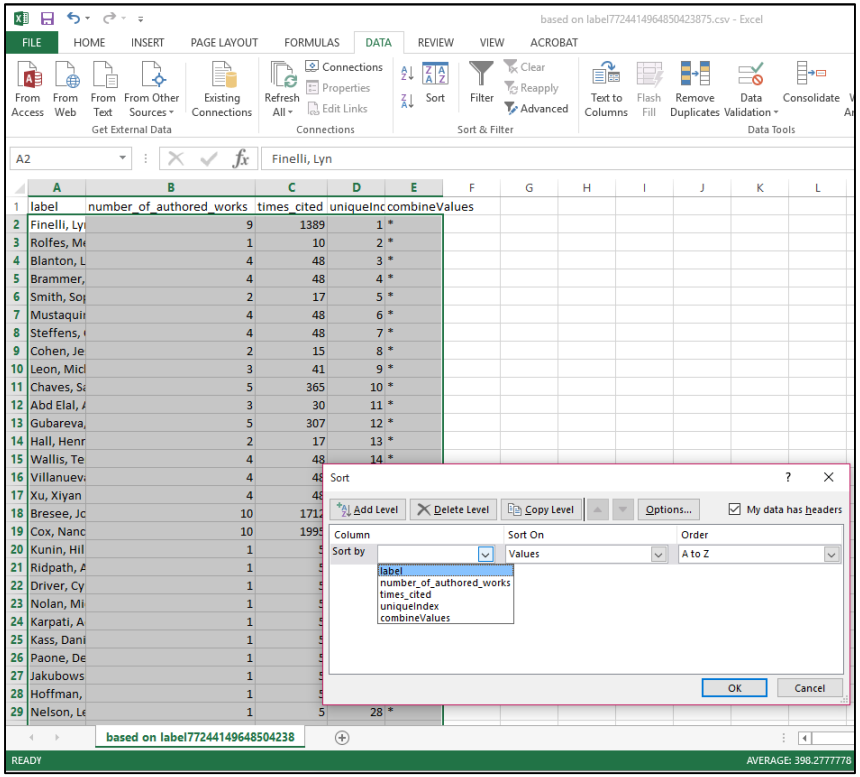
```
0.9682539105415344 similar:
  "Jones, Christopher D."
  "Jones, Christopher M."
0.9666666984558105 similar:
  "Briere, Elizabeth C."
  "Briere, Elizabeth Z."
0.9629629254341125 similar:
  "Freedman, David O."
  "Freedman, David S."
0.9583333134651184 similar:
  "Dunne, Eileen F."
  "Dunne, Eileen 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.



Sci2: Updating a merge list

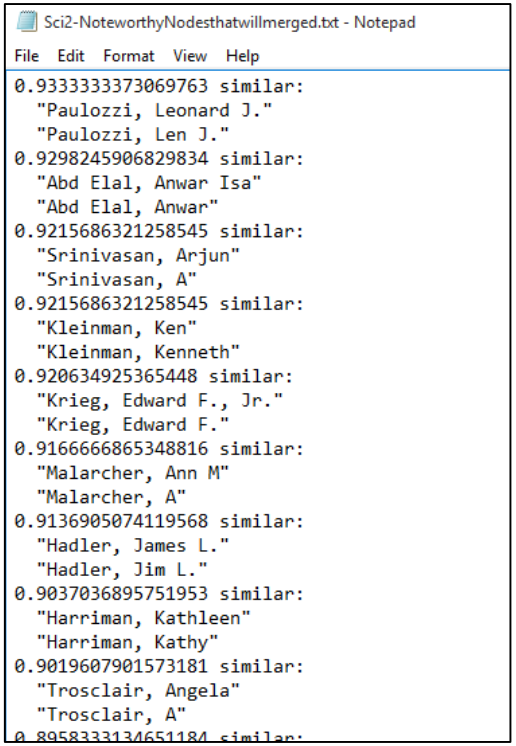


Open the merge list saved from Sci2 in a tabular data editor, such as Excel.

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.**

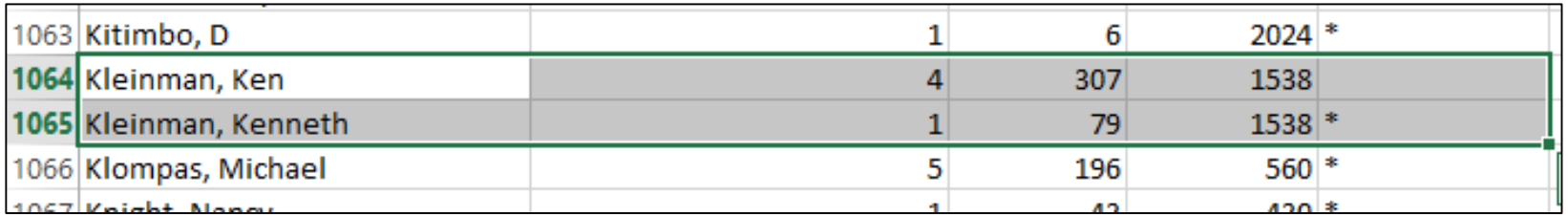


Sci2: Updating a merge list

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

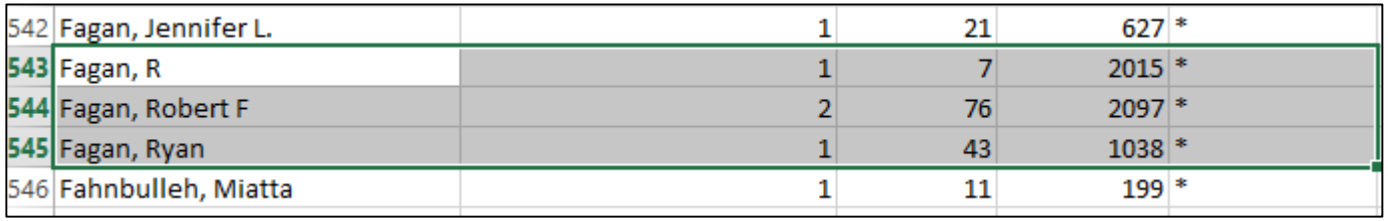
	label	number_of_authored_works	times_cited	uniqueInc	combineValue
1					
2	Finelli, Lu	9	1389	1	*

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	*

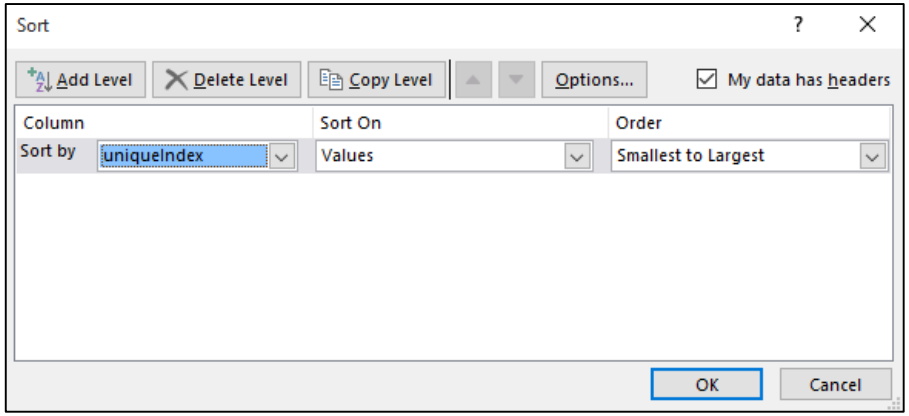
After updating a pair of **uniqueIndex** & **combinedValue** columns, search for the next author listed by using the Find tool (Ctrl-F), and copy (Ctrl-C) the name from the list and select (Ctrl-C) **OK,** and repeat the process listed above in 1). Remember, not every pair is necessarily good, sometimes, 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).



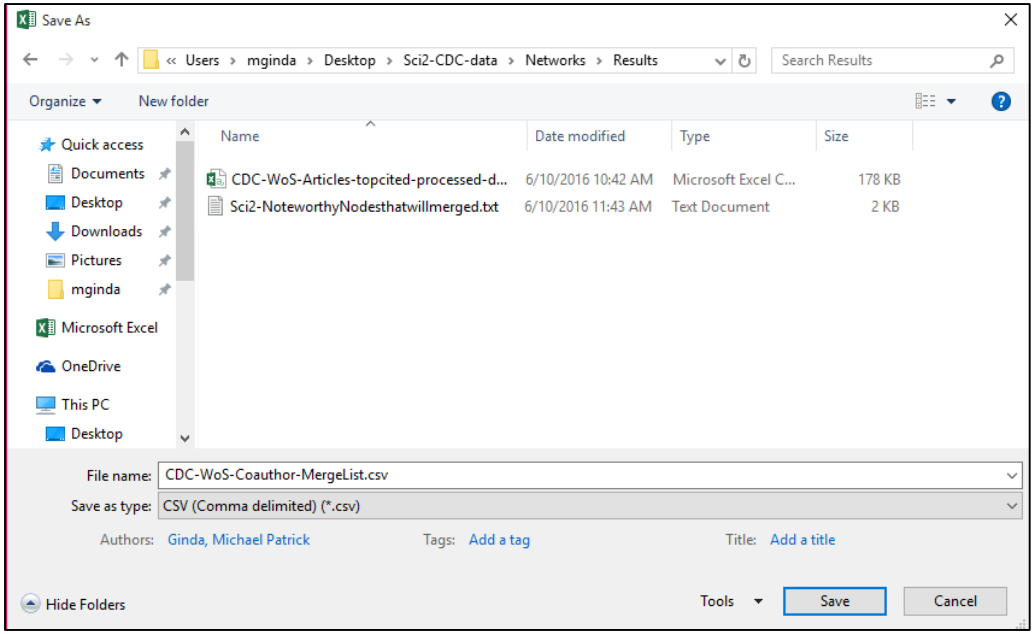
542	Fagan, Jennifer L.	1	21	627	*
543	Fagan, R	1	7	2015	*
544	Fagan, Robert F	2	76	2097	*
545	Fagan, Ryan	1	43	1038	*
546	Fahnbulleh, Miatta	1	11	199	*

Sci2: Updating a merge list

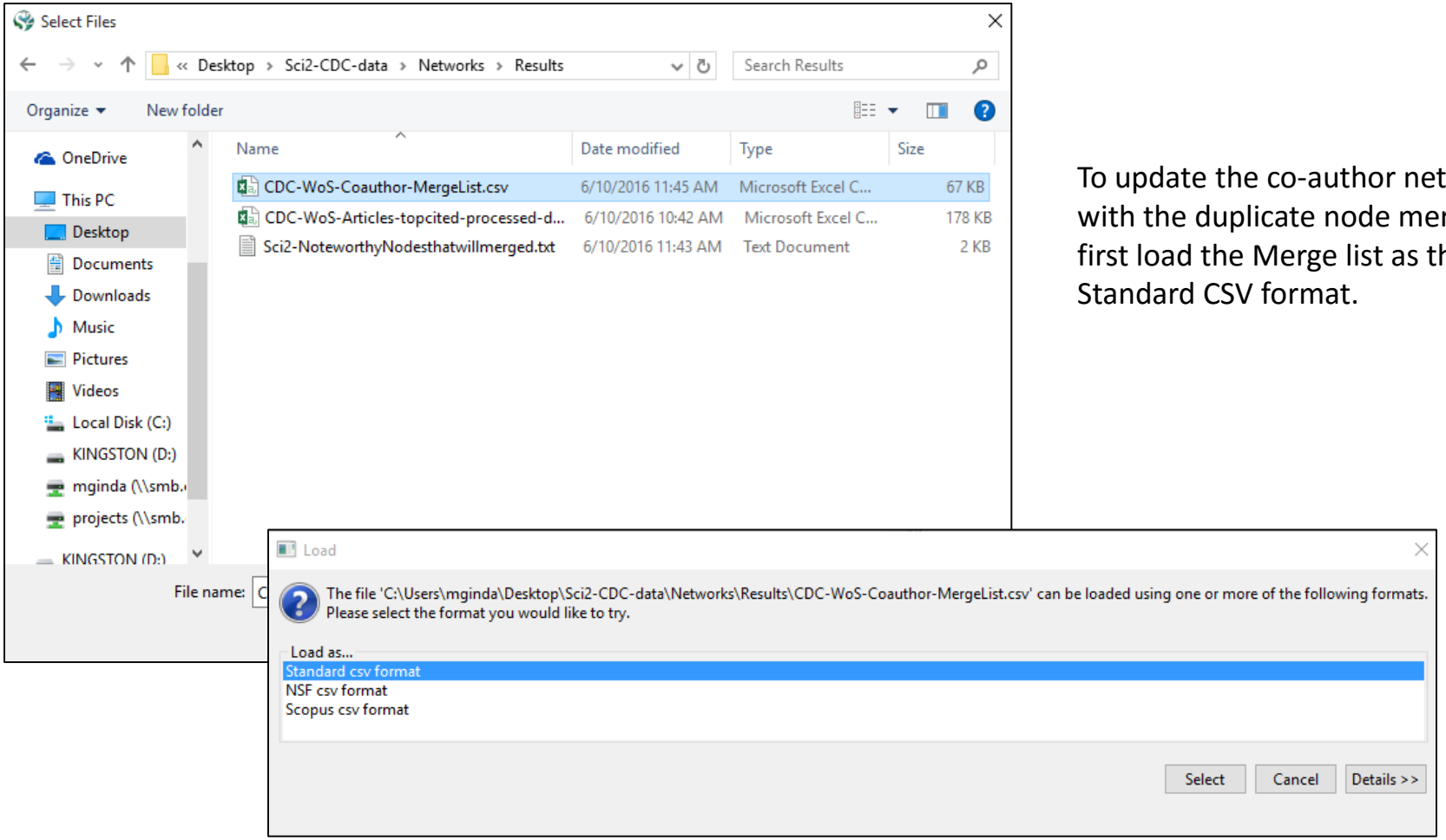
Last, selecting all of the data, resort the Merge list by the **uniqueIndex** values from smallest to largest.



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

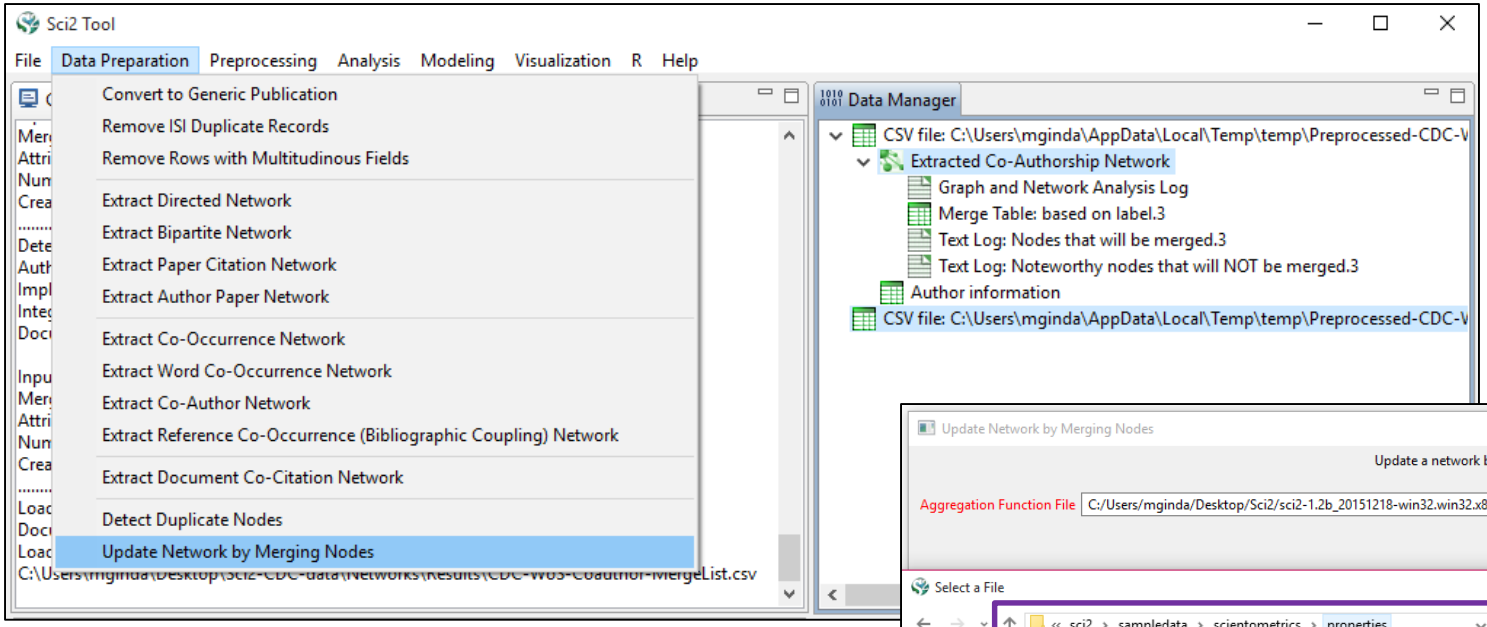


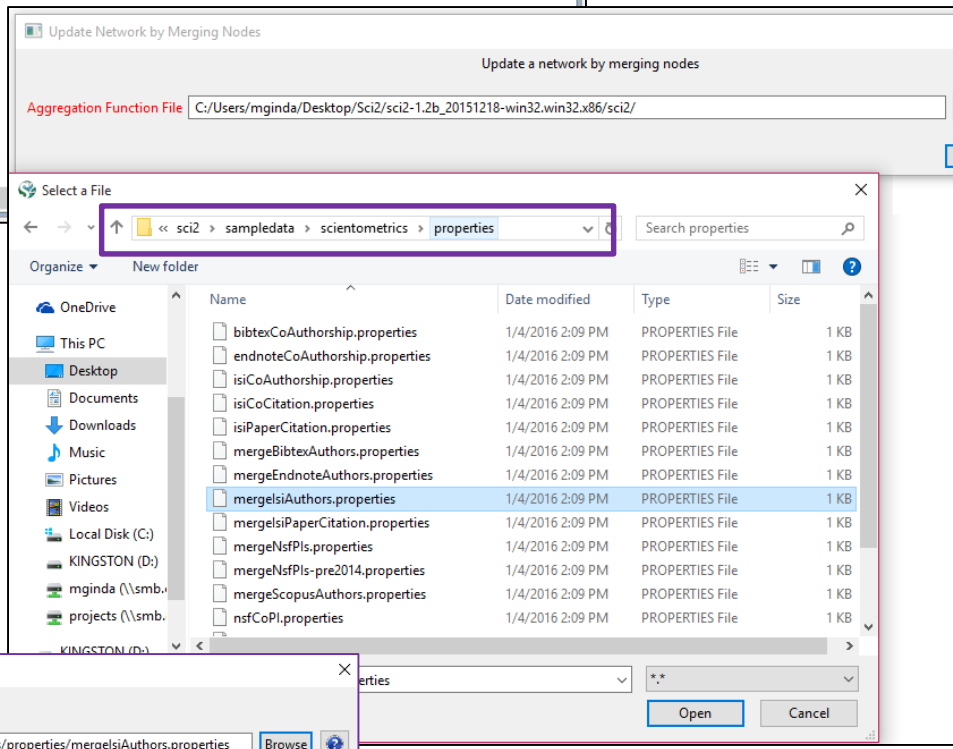
Sci2: Updating the co-author network



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

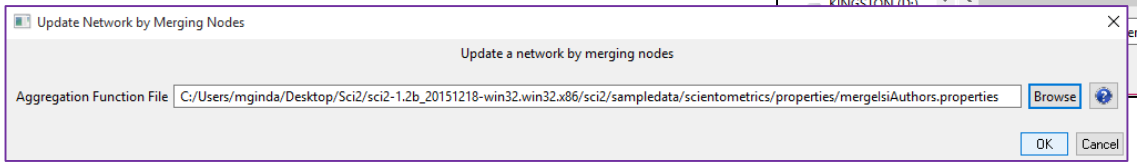
Sci2: Updating the co-author network





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 **“mergeIsiAuthors.properties”** file, and then select **OK.**



Sci2: Sci2: Applying the Network Analysis Toolkit

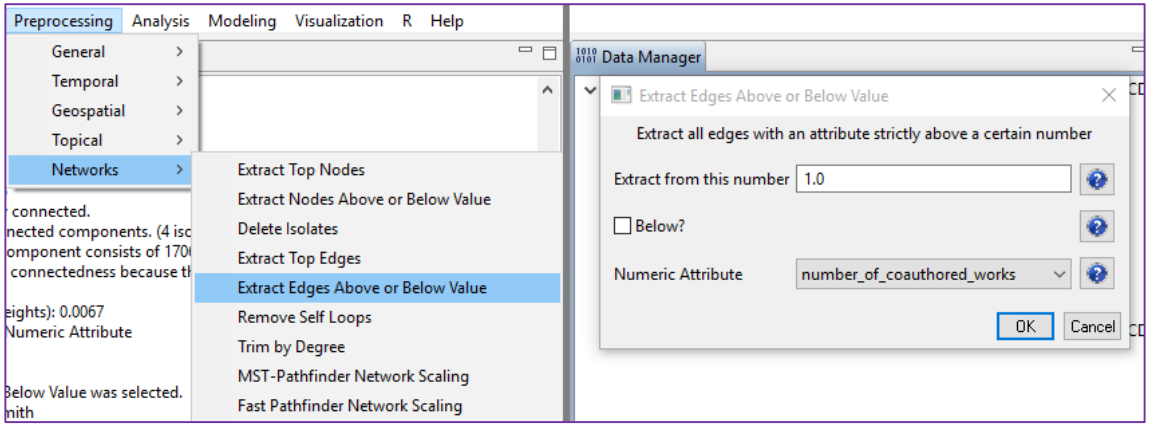
```
This graph claims to be undirected.

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
```

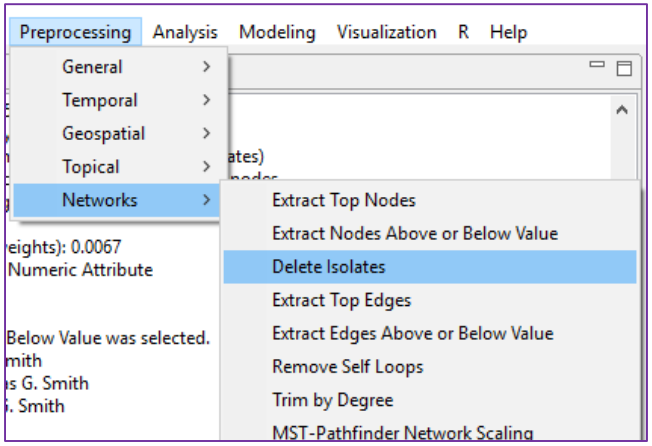
After updating the network by merging nodes, we can review the effects of our update on the network by using the NAT algorithm.

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…**

Sci2: Updating the co-author network – Preprocessing to remove weak edges & isolate nodes





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
```

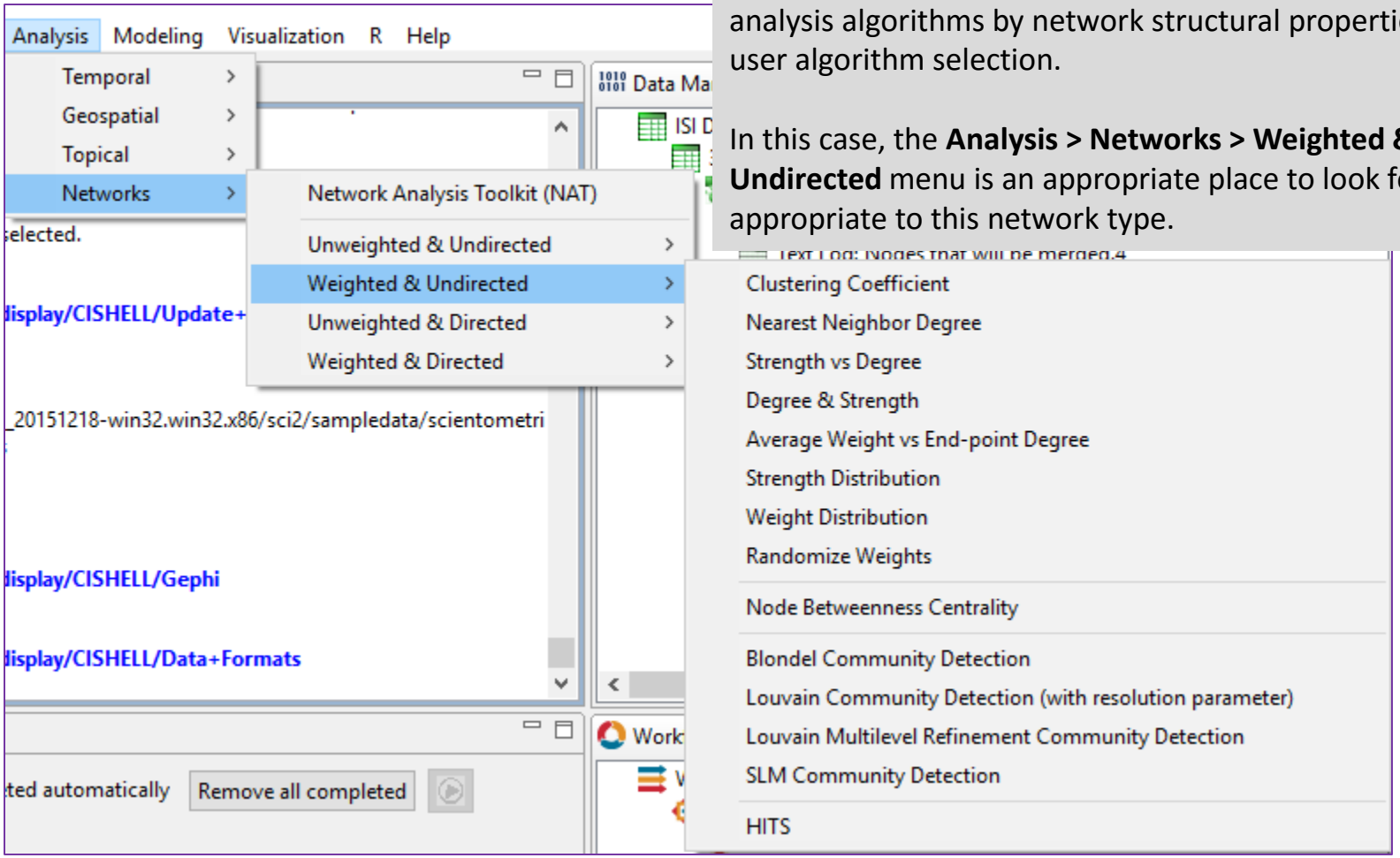
```
This graph claims to be undirected.

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

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

Sci2: Network Analysis appropriate for co-author networks

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.

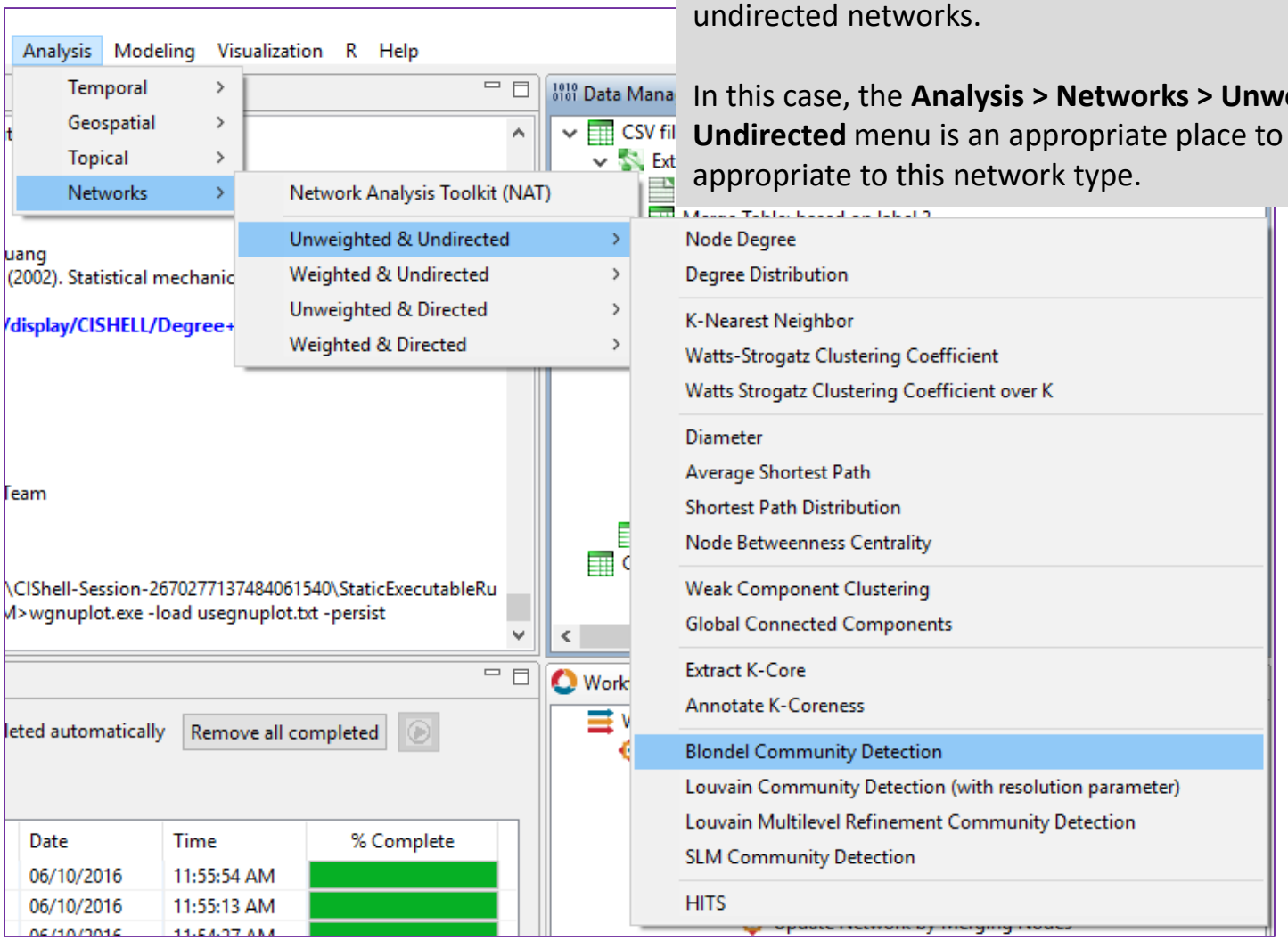
In this case, the **Analysis > Networks > Weighted & Undirected** menu is an appropriate place to look for analysis appropriate to this network type.



Sci2: Network Analysis appropriate for co-author networks

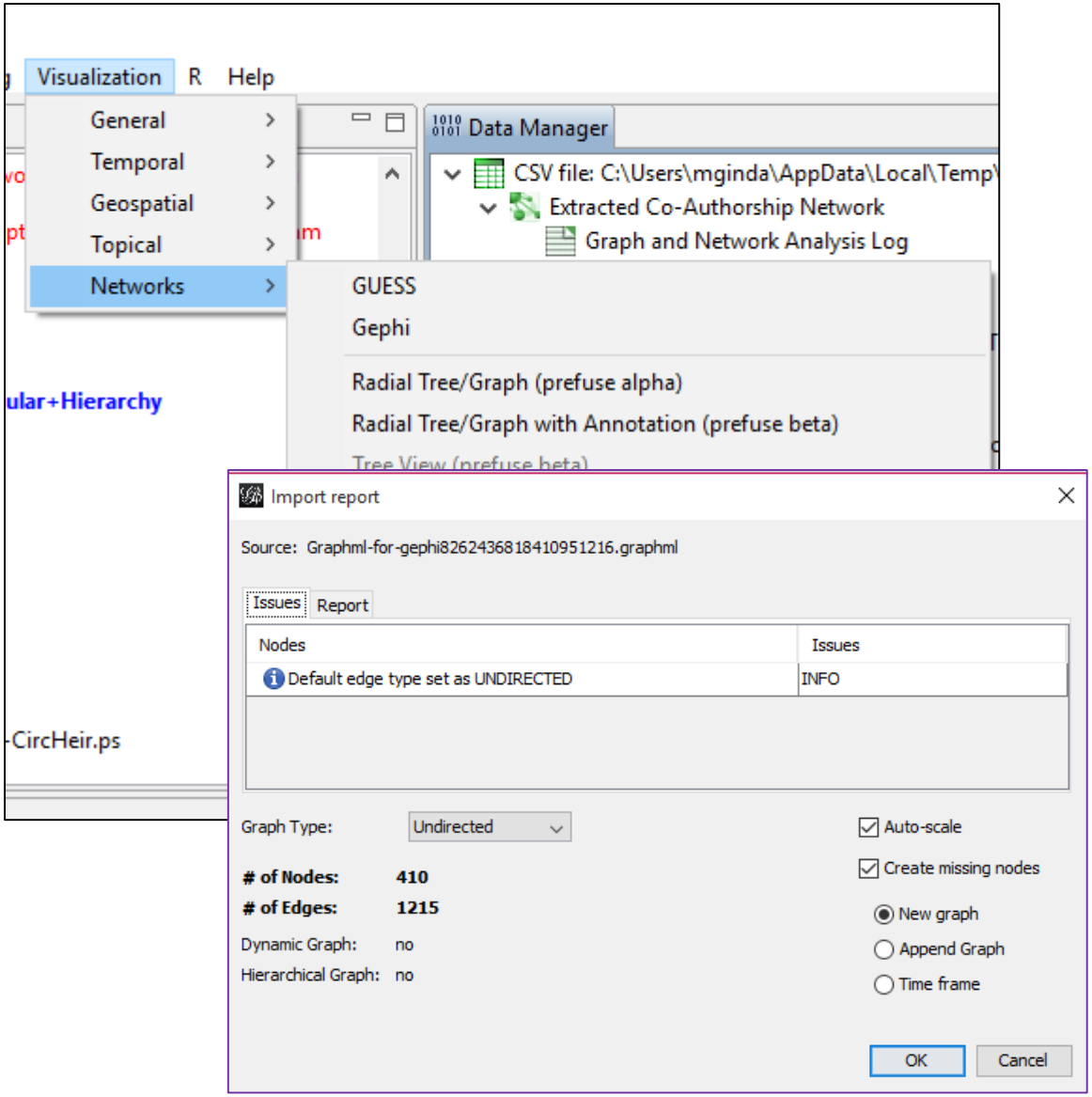
Additionally, co-authorship network may be treated as an unweighted network (node and edge weights). These algorithms may still provide insights into weighted & undirected networks.

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



Questions?

Sci2: Visualizing Co-Author Network in Gephi

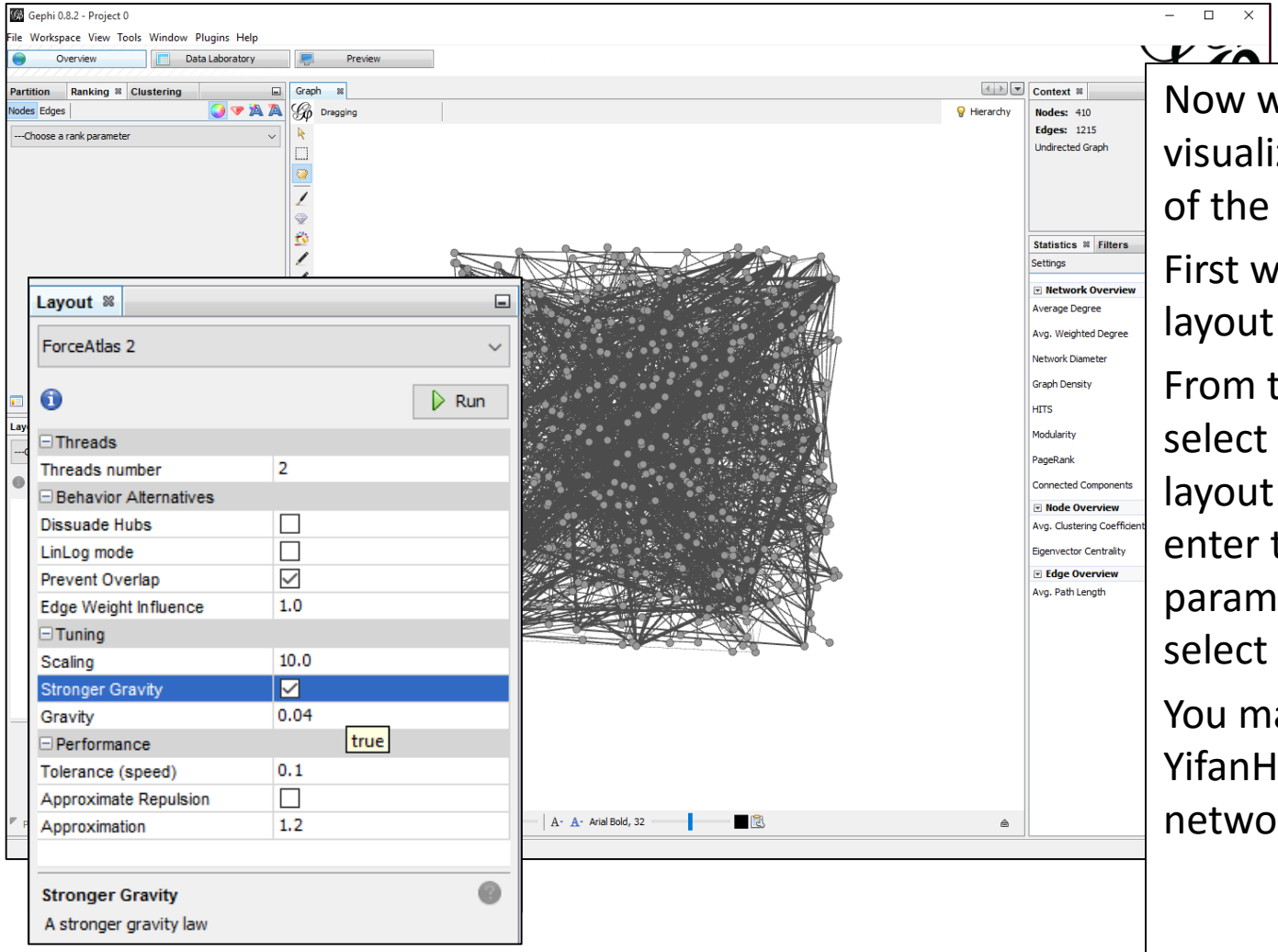


Next we will visualize the network in Gephi.

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.

Gephi: Initial Layout of Network



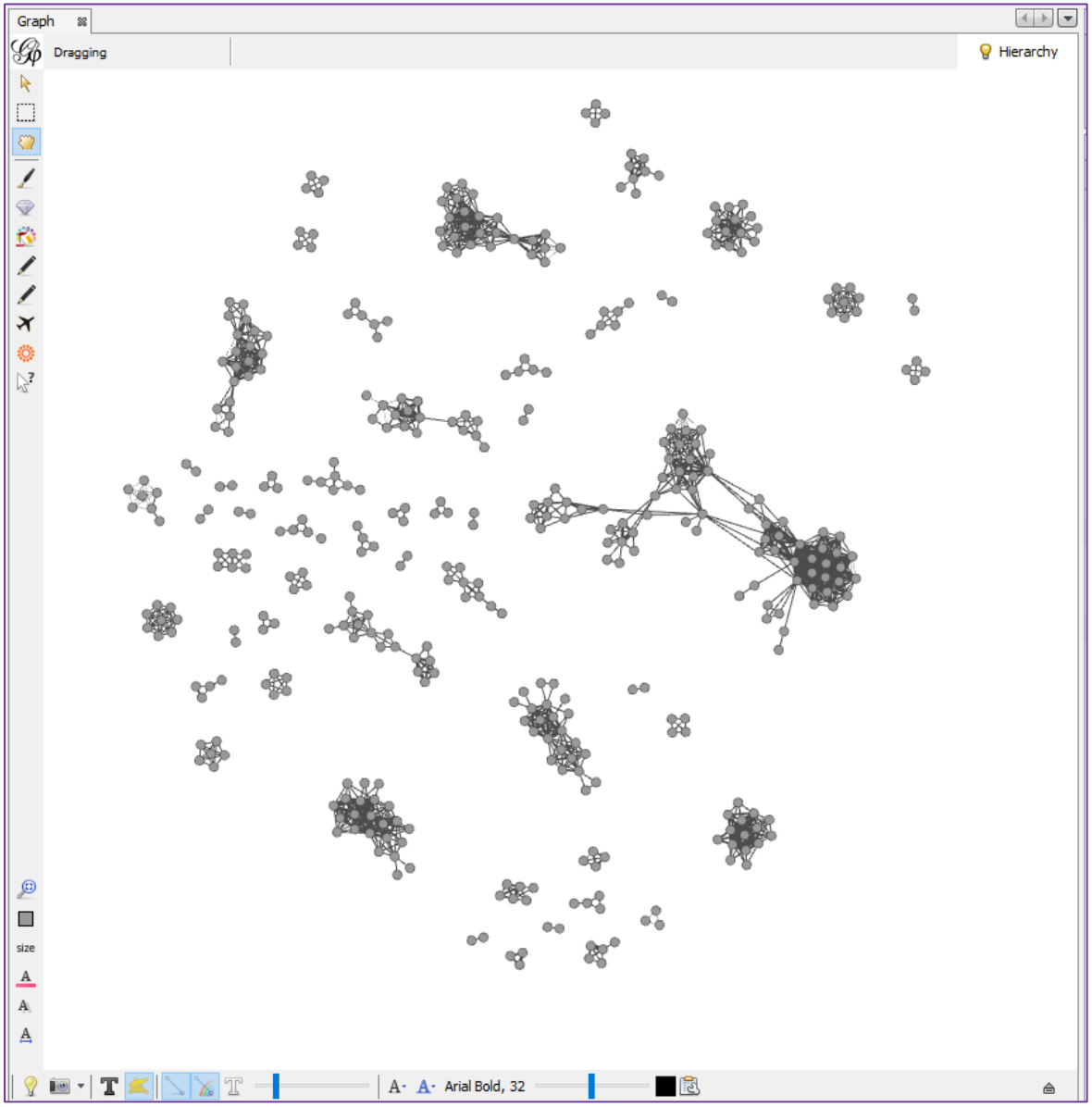
Now we can start an visualization and analysis of the network.

First we will adjust the layout of the network.

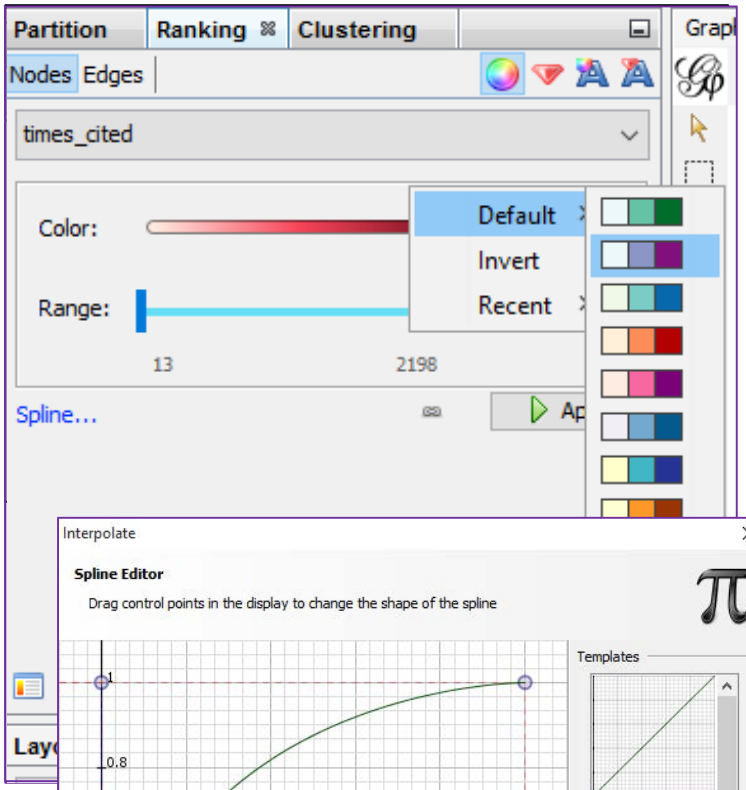
From the layout pane, select the “ForceAtlas2” layout algorithm and enter the following parameters, and then select “Run”.

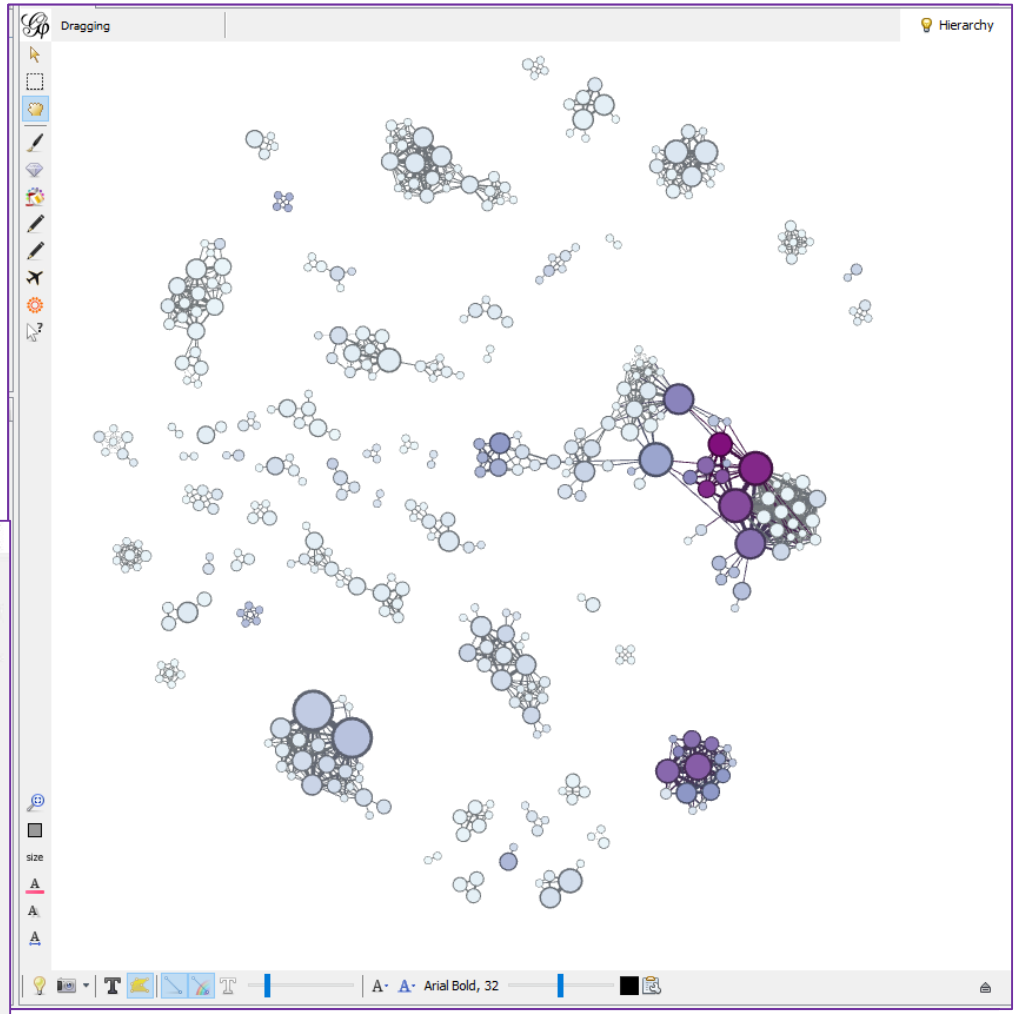
You may also select YifanHu’s Multilevel force network layout.

Gephi: Initial Layout of Network

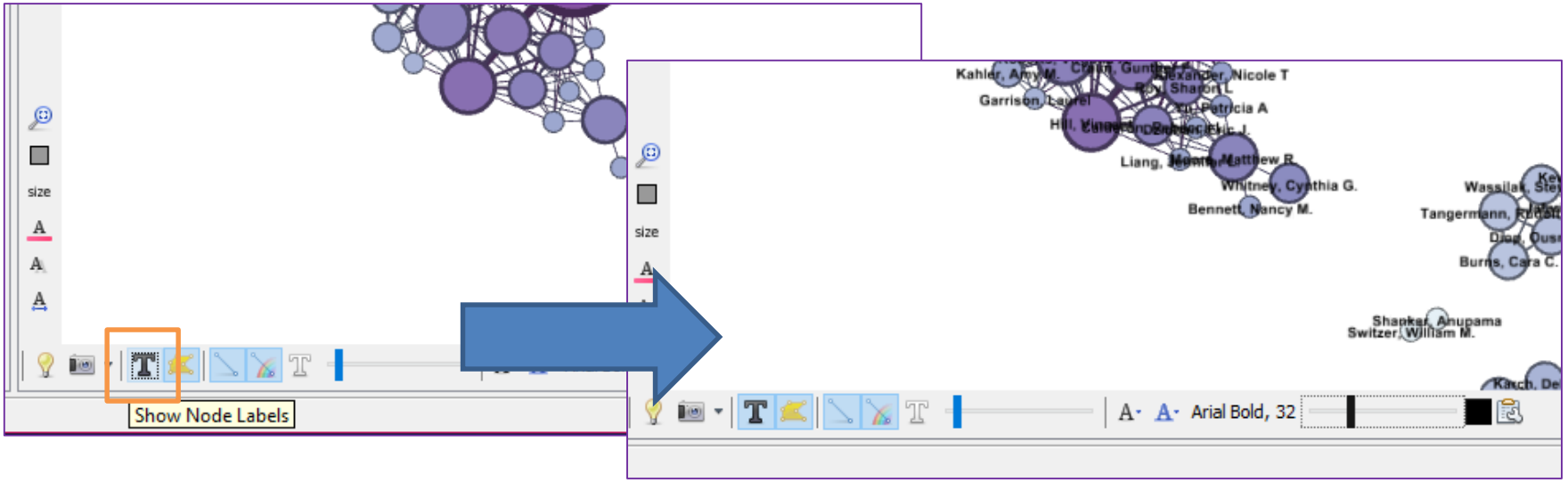


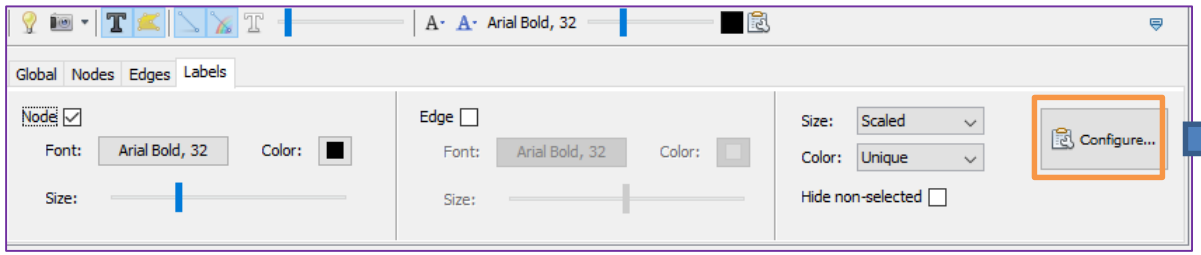
Gephi: Node Color Ranking & Scaling – Degree & Times Cited

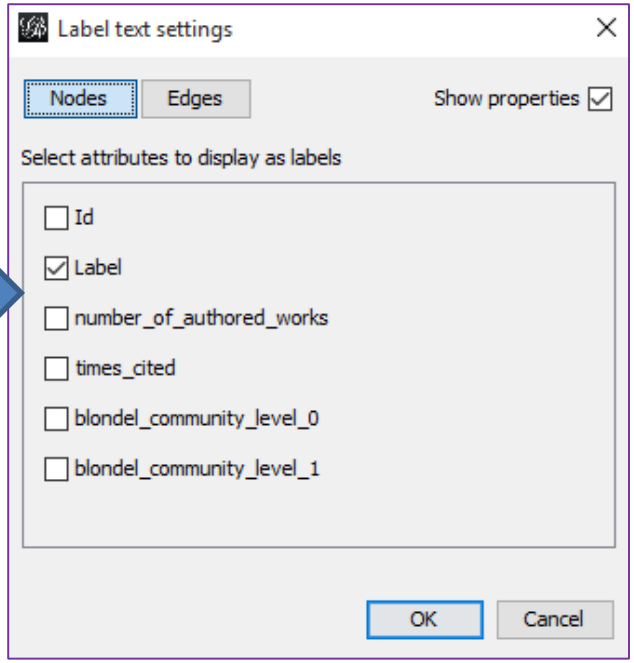




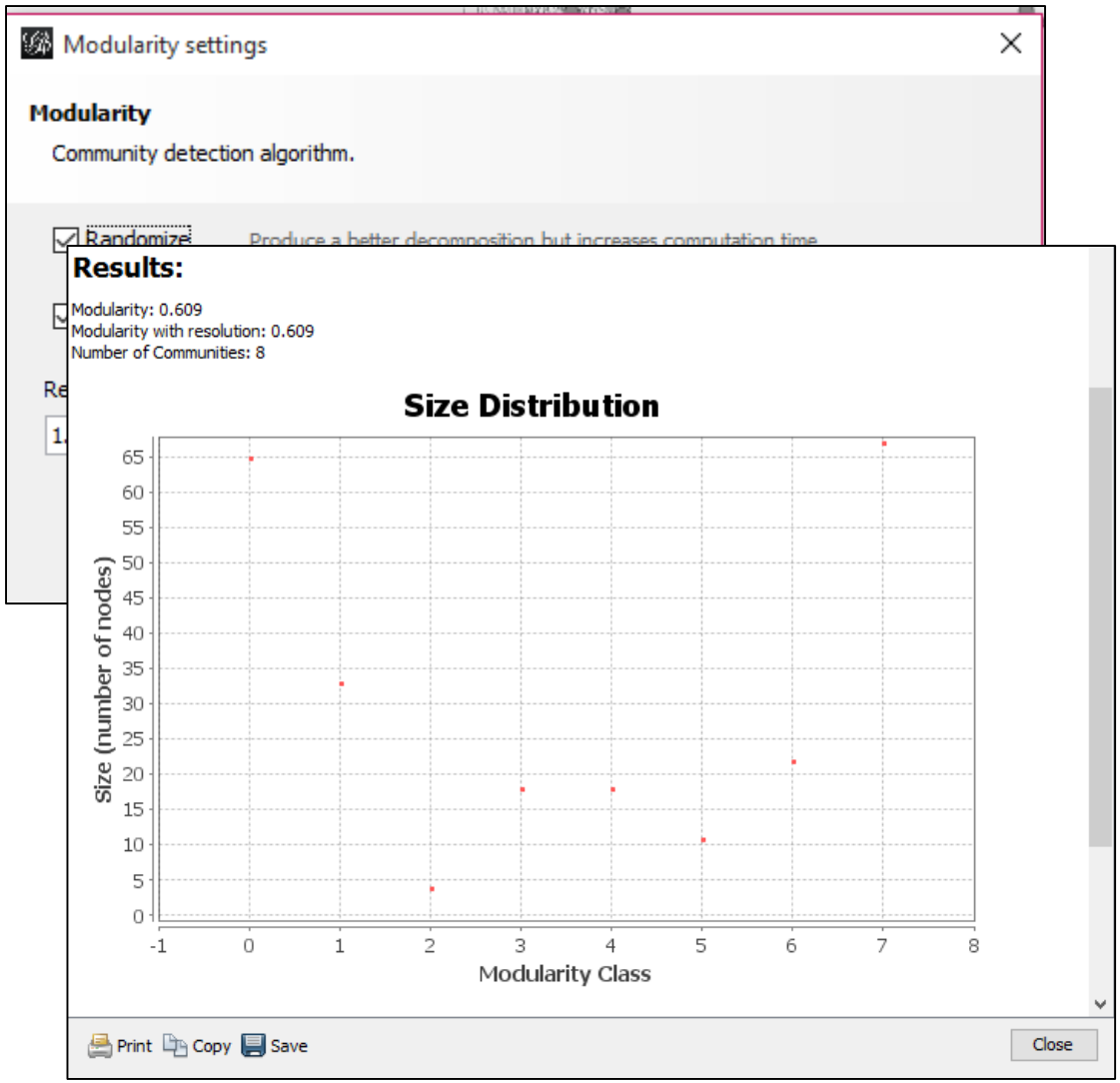
Gephi: Showing and Updating Node Labels





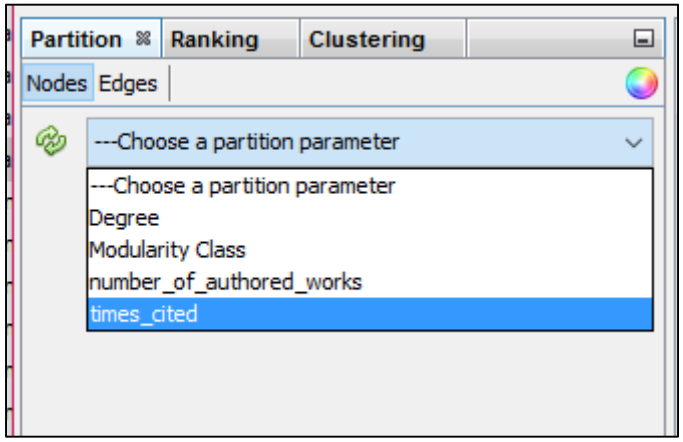


Gephi: Network statistics/Community Detection

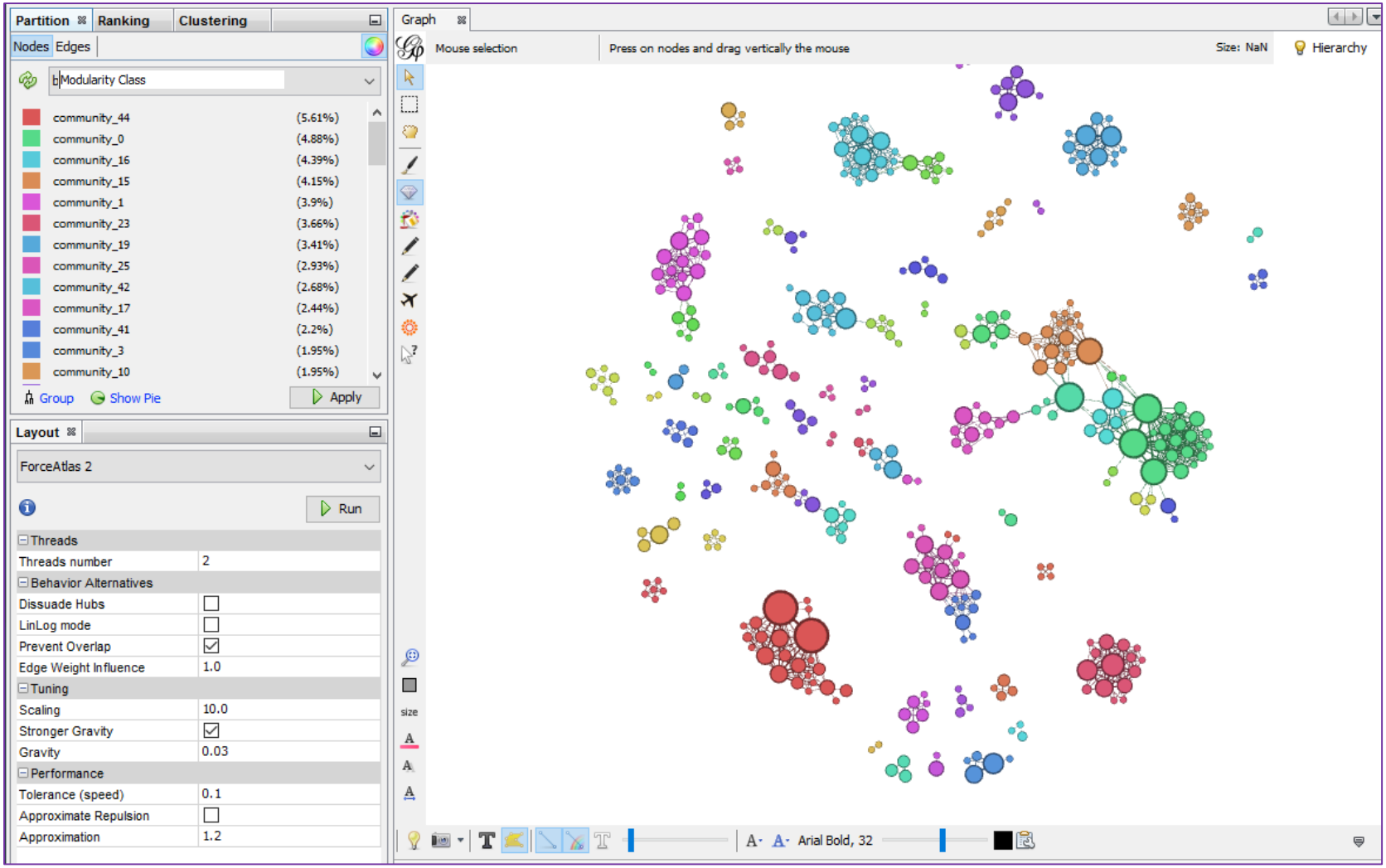


The modularity statistical algorithm calculates how the connectedness of a network, and the Blondel Communities that exist in the network. The communities are added as a partition to the nodes.

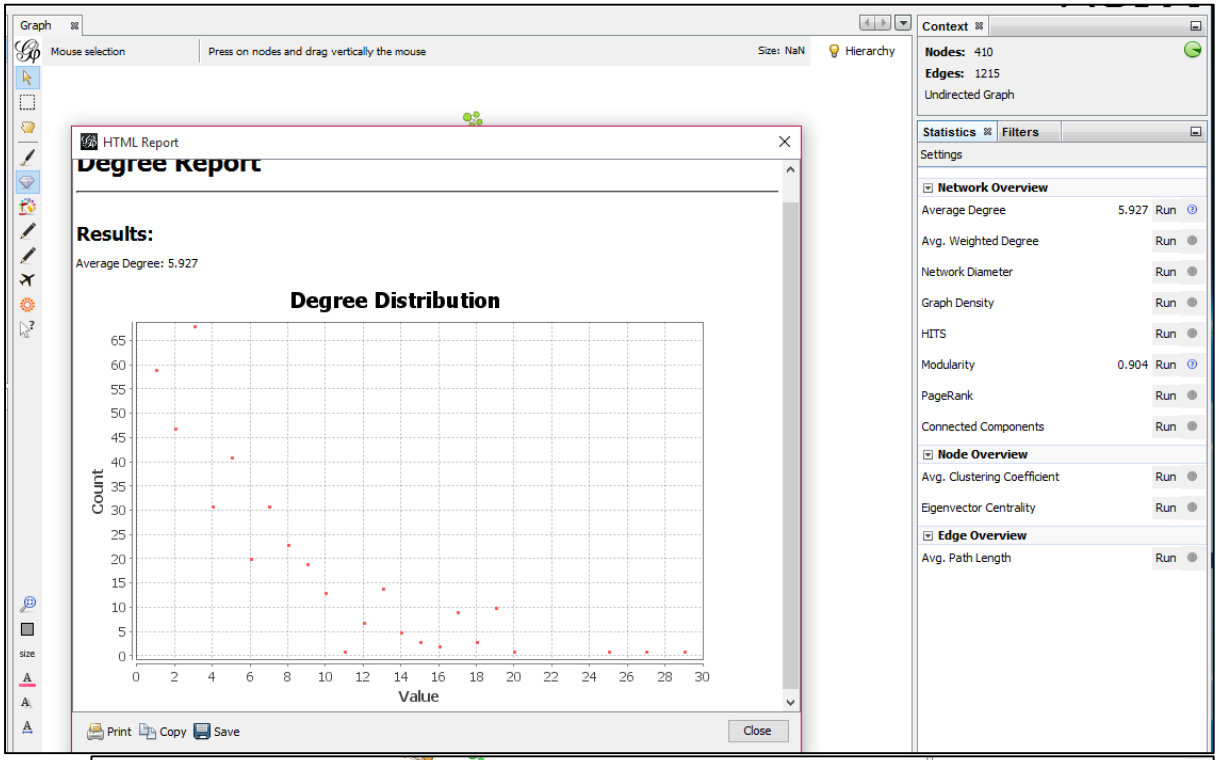
The modularity categories may be applied to the network from the Partitions window.



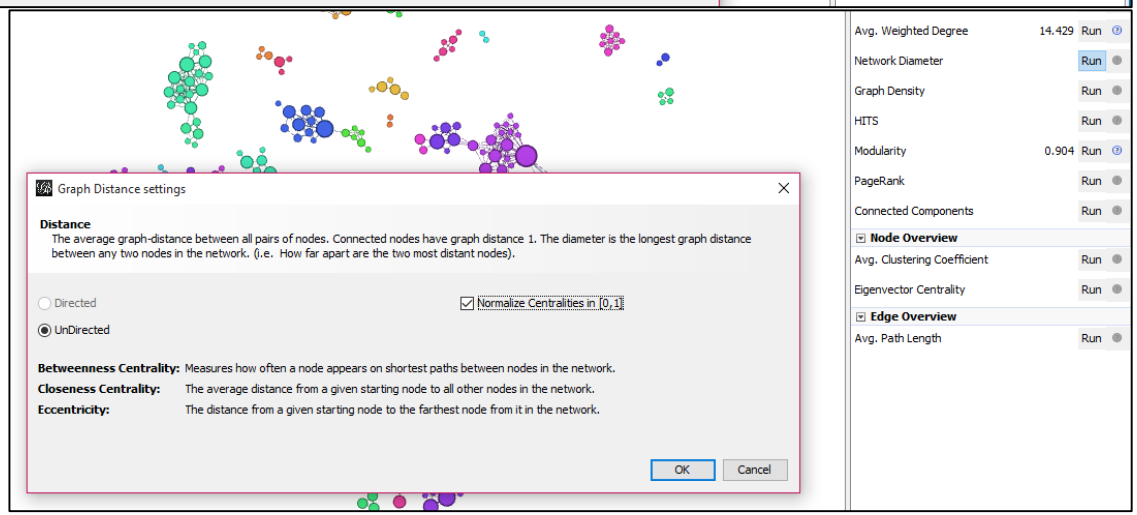
Gephi: Network statistics/Community Detection



Gephi: Network statistics

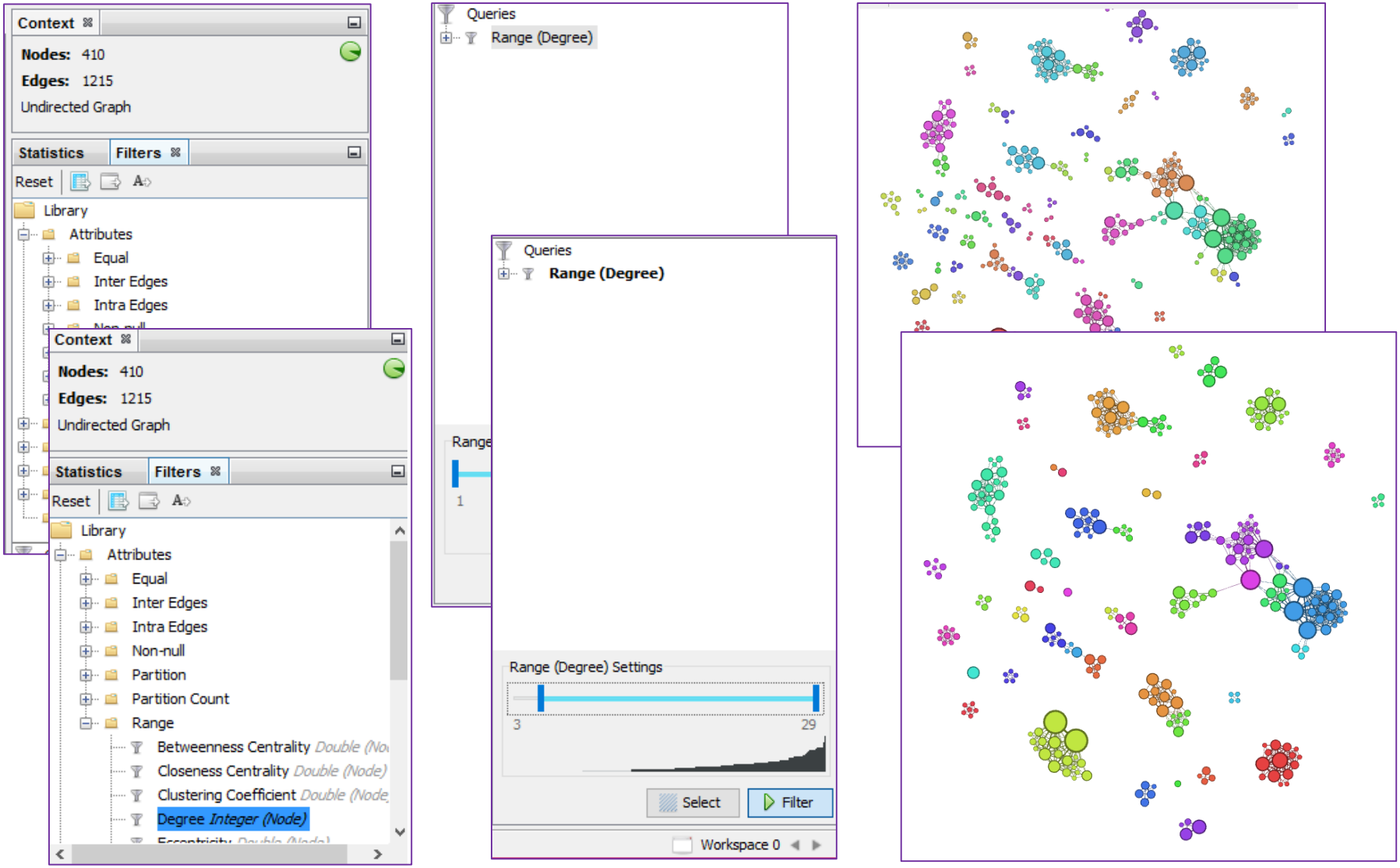


Gephi provides a variety of node and edge statistics to help understand the relationships, clustering, paths, centrality, and communities within a network.

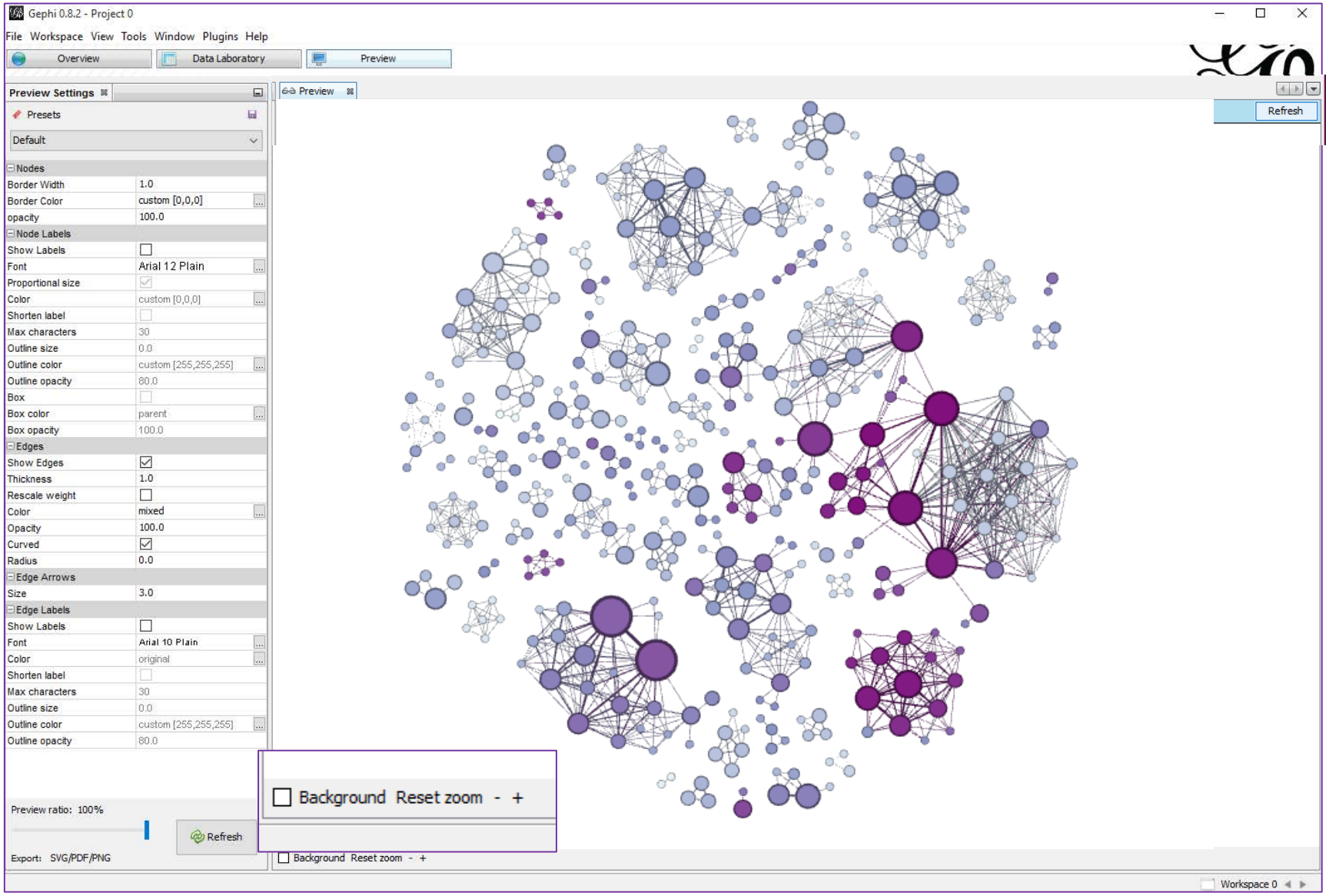


Try implementing the Average Degree, and Network Diameter statistics, which we will next visualize.

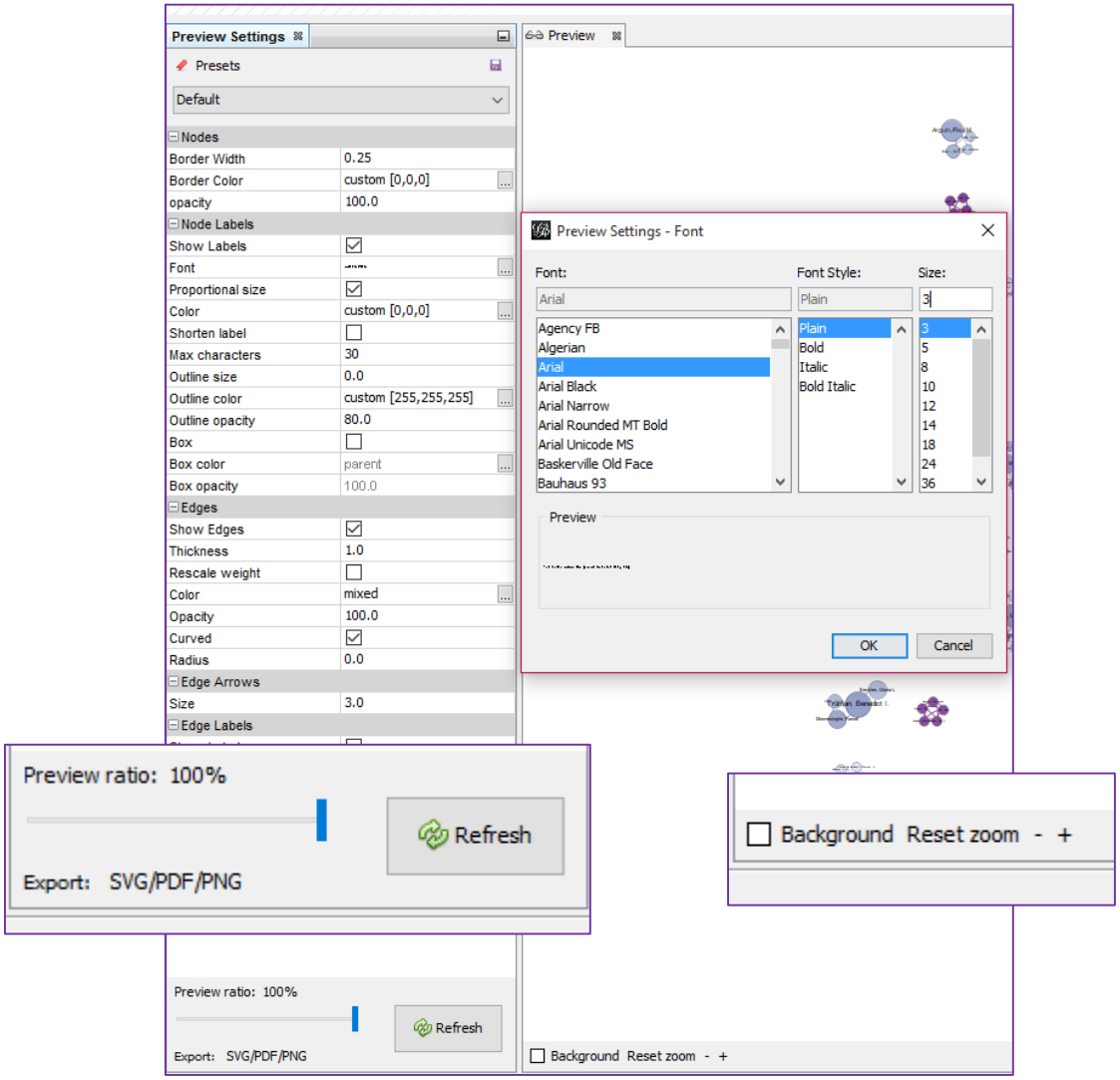
Gephi: Filtering Networks



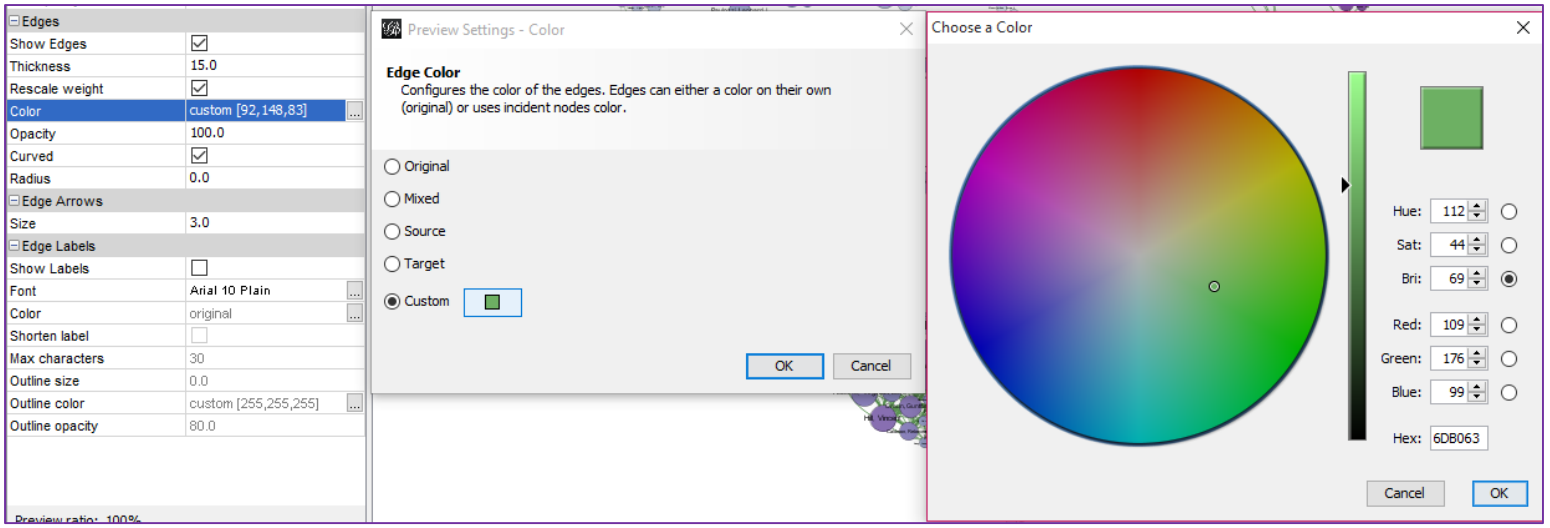
Gephi: Final visualization

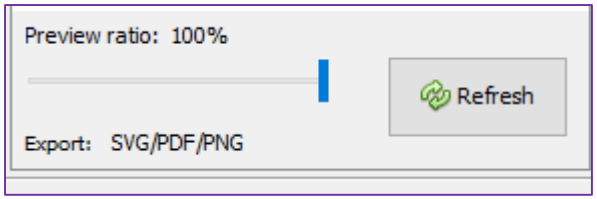


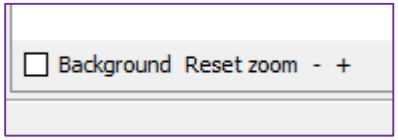
Gephi: Final visualization – Node Parameters



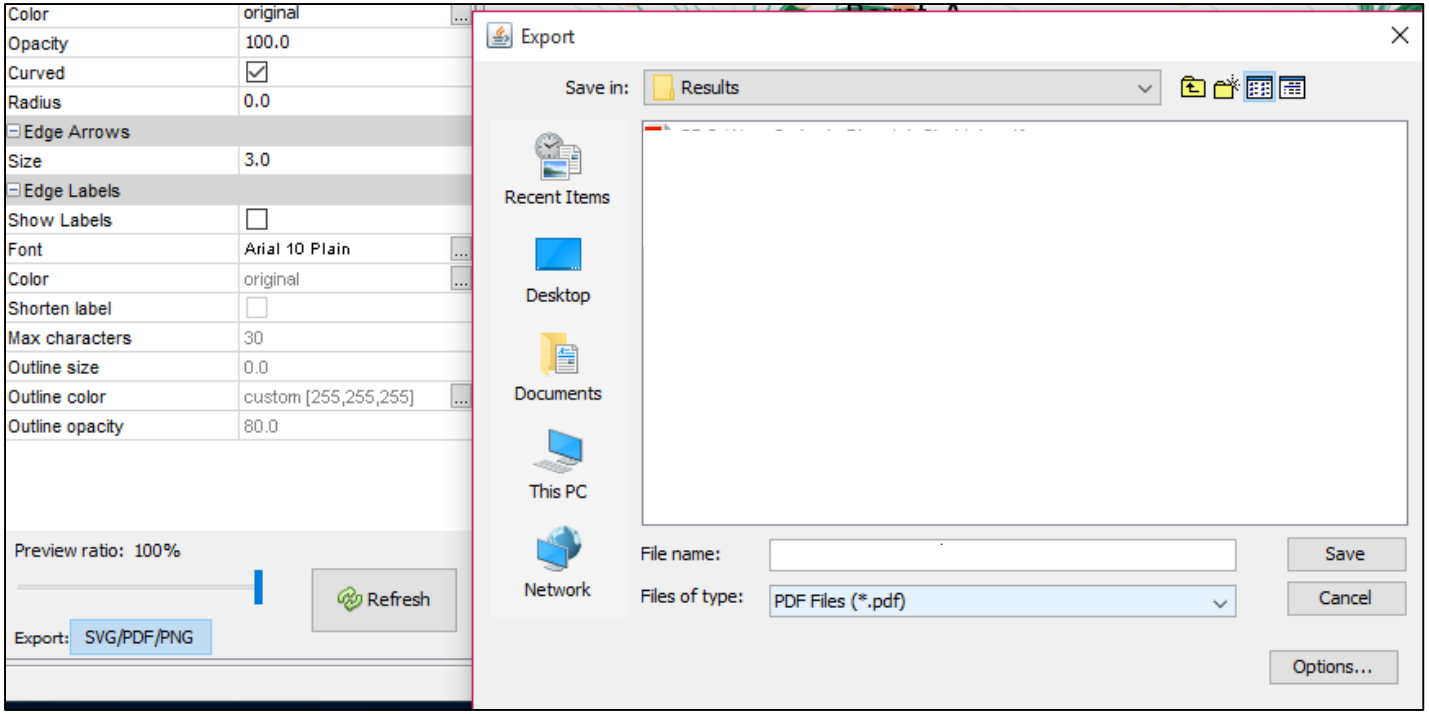
Gephi: Final visualization – Edge Color Parameters







Gephi: Exporting the visualization and networks



Questions?