

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









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

Cyberinfrastructure for Network Science Center

- Sci2 Overview scientometric analysis tool
- Open Refine data parsing, transformation, and editing tool

Tutorial Overview

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 <u>http://cns.iu.edu/docs/data/Sci2-CDC-data.zip</u>
- 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 <u>http://sci2.wiki.cns.iu.edu</u>
- Sci2 Tool v1.0 Alpha (June 13, 2012) <u>http://sci2.cns.iu.edu</u>

Cyberinfrastructure for

- Additional Datasets <u>http://sci2.wiki.cns.iu.edu/2.5+Sample+Datasets</u>
- Additional Plugins <u>http://sci2.wiki.cns.iu.edu/3.2+Additional+Plugins</u>
- 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

Cyberinfrastructure for Network Science Center

- 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

CNS Cyberinfrastructure for Network Science Center

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

CNS Cyberinfrastructure for Network Science Center

Visualization Framework

LEVELS



Atlas of Knowledge

Anyone Can Map

Workflow Design and Implementation

Cyberinfrastructure for Network Science Center









Insight Need Types	Data Scale Types	Visualization Types	Graphic Symbol Types	Graphic Variable Types	Interaction Types page 26
page 26	page 28	page 30	page 32	page 34	
 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







10



Insight Need Types	Data Scale Types	Visualization Types	Graphic Symbol Types	Graphic Variable Types	Interaction Types page 26
page 26	page 28	page 30	page 32	page 34	
 categorize/cluster order/rank/sort distributions (also outliers, gaps) comparisons trends (process and time) geospatial compositions (also of text) correlations/relationship 	 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







12



Insight Need Types	Data Scale Types	Visualization Types	Graphic Symbol Types	Graphic Variable Types	Interaction Types
page 26	page 28	page 30	page 32	page 34	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





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

Cyberinfrastructure for Network Science Center

- 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



Cyberinfrastructure for Network Science Center Workflow Design: Graphic Variable Types

Graphic Variable Types Versus Graphic Symbol Types

					Geometric Symbols			Lincolativ Surplush	Referated Sections
	_		raint	Line	ATE	safas	volume	Test, Numerals, Punctuation Marks	Images, Icons, Statistical Glyphs
1	ļ	1 (1997) 199				, i , i		Tiext	, <u>,</u> ,
F	1	sia questa	NA (Nar. Applicable)		• • • •	See Brunder Hay.	the Report Role Map.	Inc. Proper transit Spreaded Map, page 54	See Height of the Principal Manufacturepage 67
	ľ	shape guillant	PA.		• • • • •			Test Red But Test Bat	
	5	acitation quantizati	PEA.	1///		>\		토 관 · · · · · · · · · · · · · · · · · ·	🛔 (aloo) 🗰 (alood)
	-	curature question	HA.	((((P D D O			Test Test Vev Vev	000000
Mar		Angle question	PLA,	VVVLL	P D D O		Some table cells are left blank to encourage Future exploration of combinations.	Text Rost Rost Rost Aber	\odot \odot \odot \odot \odot \odot
		ciours questas	HA HA	CCCCO	0604			S ST ST ST ST ST	000000
		wice wice	- · · · · · · · · · · · ·					Test Test Test Test Test	* * * * *
	ġ				10 No.			Test Test Test Test	1 (alwa) 1 (alwa)
L			••••••••					Text Text Text Text Text	internation (1990) 🥌 (1999 nation)
			roint	Line	ATE	sarbas	volume	Text, Namwah, Punctuation Marks	Images, Icons, Statetical Glyphs
	1	spacing quantitation			医医肠管管管管管管管管管管管管管管管管管管管管管管管管管管管管管管管管管管管管		医脱酸酸酶		
		control quantitativ					麗区区 4 ::		
		ration quinte	19 12 12 13 13 19 12 13 13 13				333 III 33 III III		232 図 田 田 🎬
	ľ	orientation quantitation	PLA		5550 CO 0000 HANG 7-27		THE REPORT OF THE PARTY OF THE		
	1				100		55 S 55 S 85 S		See Field Tectors of Bondom Pesitions page S1
	+	anders generation					<u> </u>		And Participants
1.0.1		andert garitati skr garitati managery garitati	= / / / / / /					Image: Section of the sectio	Image: A set of the location of the loc
Net of		anderi yariki ikir yariki tungareny yariki ulain yariki						Image: Non-State Image: Non-State<	
Nev-1	dite	nadint yaniki skr yaniki tangangy yaniki shalog yaniki timorapi ogli yaniki						Image Image <th< th=""><th>Statistical Postanci Scale Amage Image Image Image Image Image Image Image Image Image Image Image Image Image Image Image Image Image Image Image Image Image Image Image <t< th=""></t<></th></th<>	Statistical Postanci Scale Amage Image Image Image Image Image Image Image Image Image Image Image Image Image Image Image Image Image Image Image Image Image Image Image <t< th=""></t<>
interes.	dete	endert ender dar ender tanparen ender vlang besten vlancopic orgh ender speed ender	 Anticipal and the second second		233 233 233 111 111 111 111 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4	Sufferent - budgeoard		Text Text Text Text Text Text Text Text	Image: Section 2 model Section 2 model Image: Section 2 model
Port of the second s	0000	exadent examine alter examine anteres examine thating examine thating examine theoremails and participation examine theoremail				Sofier In Engrand		Image: Second	Statistical Postancia Image: Statistical Postanci Image: Statistic
Perform 1	0000	endert werden der werden tanparen werden vlading werden ubrocept oph upred werden vlader werden	Image: state	Image: Second	Image: State	Sofier In Sofier In Sofie		Image: Second	Image: Section 2 and the independent of the ind

Cyberinfrastructure for Network Science Center Workflow Design: Graphic Variable Types

Graphic Variable Types Versus Graphic Symbol Types



erinfrastructure for 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** (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** (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.

Quantitative









Workflow Design: Color Schemes



See Color Use Guidelines for Mapping and Visualization by Cynthia Brewer

Cyberinfrastructure for Network Science Center



Questions?



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

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

9:15 Tool Overview and Trouble Shooting

Cyberinfrastructure for Network Science Center

- Sci2 Overview scientometric analysis tool
- Open Refine data parsing, transformation, and editing tool

Tutorial Overview

Gephi – network visualization

9:45 Geospatial Analysis

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

11:00 Topical/Temporal Analysis: Burst Detection

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

12:30 Lunch

1:30 Network Analysis

- Introduction to Network Analysis
- Network Analysis: Bimodal networks with Morbidity Data
- Network Analysis: Co-authorship Network with CDC publications
 4:00 Wrap-up

4:30 Adjourn



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

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



Microscopes





Telescopes

Macroscopes



Macroscopes (cont.)

Plug-and-Play Macroscopes

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

Macroscopes make it easy to

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



8:00 Welcome and Overview of Tutorial and Attendees8:30 Visualization Framework and Workflow Design

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

9:15 Tool Overview and Trouble Shooting

Cyberinfrastructure for Network Science Center

- Sci2 Overview scientometric analysis tool
- Open Refine data parsing, transformation, and editing tool

Tutorial Overview

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

Shell (CIShell)

Sci2 Infrastructure

- CIShell (<u>http://cishell.org</u>) is an open source software specification for the integration and utilization of datasets, algorithms, and tools
- It extends the Open Services Gateway Initiative (OSGi) (<u>http://osgi.org</u>), 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 <u>http://sci2.wiki.cns.iu.edu/display/SCI2TUTORIAL/2.3+Data+Formats</u>.

29

Network Sci

Cyberinfrastructure for

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 <u>http://sci2.wiki.cns.iu.edu</u>

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.

Troubleshooting

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
```

Organize 👻 🦳 Open 👻	Share with 🔻 🛛 Print 🛛 Burn	New folder		
🔆 Favorites	Name	Date modified	Туре	Size
🥅 Desktop	Configuration	1/3/2013 2:06 PM	File folder	
🗼 Downloads	鷆 features	12/21/2012 9:21 AM	File folder	
🔠 Recent Places	licenses	12/21/2012 9:21 AM	File folder	
	퉬 logs	1/3/2013 2:31 PM	File folder	
🥽 Libraries	plugins	1/3/2013 9:16 AM	File folder	
	鷆 sampledata	12/21/2012 9:21 AM	File folder	
👰 Computer	퉬 scripts	12/21/2012 9:21 AM	File folder	
🚢 Local Disk (C:)	workspace	12/21/2012 9:33 AM	File folder	
🖵 dapolley (Y:)	artifacts.xml	12/21/2012 9:21 AM	XML Document	41
🖵 projects (Z:)	eclipsec.exe	12/21/2012 9:21 AM	Application	24
	🜍 sci2.exe	12/21/2012 9:21 AM	Application	52
年 Network	😰 sci2.ini	12/21/2012 9:21 AM	Configuration sett	1
	Type: Config Size: 24 byte	guration settings s		

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.

Troubleshooting

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



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

			[- • ×					
O ↓ sci2 ↓		✓ Sear	ch sci2	Q					
Organize 🔻 🦳 Open 🔻	Share with 👻 Print Burn	New folder	:	• 🔟 🔞					
🔶 Favorites	Name	Date modified	Туре	Size					
🧮 Desktop	Configuration	1/3/2013 2:06 PM	File folder						
🗼 Downloads	鷆 features	12/21/2012 9:21 AM	File folder						
📃 Recent Places	퉬 licenses	12/21/2012 9:21 AM	File folder						
	퉬 logs	1/3/2013 2:31 PM	File folder						
🥽 Libraries	퉬 plugins	1/3/2013 9:16 AM	File folder						
	퉬 sampledata	12/21/2012 9:21 AM	File folder						
👰 Computer	퉬 scripts	12/21/2012 9:21 AM	File folder						
🚢 Local Disk (C:)	퉬 workspace	12/21/2012 9:33 AM	File folder						
🚽 dapolley (Y:)	📄 artifacts.xml	12/21/2012 9:21 AM	XML Document	41 KB					
🖵 projects (Z:)	eclipsec.exe	12/21/2012 9:21 AM	Application	24 KB					
	Sci2.exe	12/21/2012 9:21 AM	Application	52 KB					
📬 Network	📰 sci2.ini	12/21/2012 9:21 AM	Configuration sett	1 KB					
	Type: Confi Size: 24 byte	guration settings es							
sci2.ini Date modified: 12/21/2012 9:21 AM Configuration settings Size: 24 bytes									

Troubleshooting

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

Cyberinfrastructure for Network Science Center

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+Lo cale





8:00 Welcome and Overview of Tutorial and Attendees8:30 Visualization Framework and Workflow Design

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

9:15 Tool Overview and Trouble Shooting

Cyberinfrastructure for

- Sci2 Overview scientometric analysis tool
- Open Refine data parsing, transformation, and editing tool

Tutorial Overview

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

CNS Cyberinfrastructure for Network Science Center 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 <u>created online tutorials</u> and <u>documentation</u> to support new users.

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



8:00 Welcome and Overview of Tutorial and Attendees8:30 Visualization Framework and Workflow Design

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

9:15 Tool Overview and Trouble Shooting

Cyberinfrastructure for Network Science Center

- Sci2 Overview scientometric analysis tool
- Open Refine data parsing, transformation, and editing tool

Tutorial Overview

Gephi – network visualization

9:45 Geospatial Analysis

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

11:00 Topical/Temporal Analysis: Burst Detection

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

12:30 Lunch

1:30 Network Analysis

- Introduction to Network Analysis
- Network Analysis: Bimodal networks with Morbidity Data
- Network Analysis: Co-authorship Network with CDC publications
 4:00 Wrap-up

4:30 Adjourn




Gephi is an open source network analysis and visualization tool developed by the Gephi Consortium.

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

A <u>full list of features</u> is found on the Gephi.org website. <u>Documentation</u> & <u>tutorials</u> for the tool, <u>system</u> <u>requirements</u> and <u>GitHub repository</u> also are useful resouces.



Questions?



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

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

9:15 Tool Overview and Trouble Shooting

Cyberinfrastructure for Network Science Center

- Sci2 Overview scientometric analysis tool
- Open Refine data parsing, transformation, and editing tool

Tutorial Overview

Gephi – network visualization

9:45 Geospatial Analysis

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

11:00 Topical/Temporal Analysis: Burst Detection

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

12:30 Lunch

1:30 Network Analysis

- Introduction to Network Analysis
- Network Analysis: Bimodal networks with Morbidity Data
- Network Analysis: Co-authorship Network with CDC publications
 4:00 Wrap-up

4:30 Adjourn



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

Terminology

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

CNS Cyberinfrastructure for Network Science Center Introduction to Geospatial Analysis

Representation of Geospatial Data

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

• Latitude/Longitude



CNS Cyberinfrastructure for Network Science Center Introduction to Geospatial Analysis

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?



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

Cyberinfrastructure for

- Sci2 Overview scientometric analysis tool
- Open Refine data parsing, transformation, and editing tool

Tutorial Overview

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

CNS Cyberinfrastructure for Network Science Center 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.



CNS (cns.iu.edu)



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.

Proportional Symbol Map: CDC HIV/AIDS Diagnosis Data

CDC WON	I DER FAQ	Help	Contact Us	WONDER Search	
	Al	DS Publi	c Use (Vinta	ge 2002) Request	
Request Form	Results Map	Chart	About		
		Processin	g Request, pleas	e wait for results	
			Cancel Requ	est	
AIDS Public Inf	ormation Dataset Doc	umentation Ot	her Data Access Data L	Ise Restrictions How to Use WONDER	Reset
	Make all desired	selections an	d then click any Sen	d button one time to send your request.	
1. Organize	table layout:				Send <u>Help</u>
	<u> </u>				
Group Resul	ts By Location		~		
Ar	d By Year Diagnosed		~		
Ar	d By Race or Ethnicity		~		
Ar	None		~		
Ar	None		\sim		
	Title				
2. §	WONDER	FAQ I	Help Contact (Js WONDER Search	
Loc		AIDS	Public Use (Vintage 2002) Results	
	Desults		Chart About		
Val	quest Form Results	мар	Chart About		
Wa	AIDS Public Information	Dataset Docume	entation Other Data Acc	ess Help for Results Printing Tips Help with Expo	rts Export Reset
(Quick Options	More Options	J	Top Notes	Citation Query Criteria
	Messages:				
	Rows with zero Case	es are hidden	. Use Quick Options	above to show zero rows.	
	Location	ŧ	Year Diagnosed	Race or Ethnicity	Cases
	Akron, OH		1982	Black (and also not Hispanic)	1
	Akron, OH		1984	White (and also not Hispanic)	3
	Akron, OH		1985	White (and also not Hispanic)	4
	Akron, OH		1986	Black (and also not Hispanic)	3
	Akron, OH		1986	White (and also not Hispanic)	9
				······	

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 <u>NCHHSTP Atlas</u>."

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



Cyberinfrastructure for Network Science Center

DEPLOY

verinfrastructure for Open Refine: Review of CDC HIV/AIDS Diagnosis data

te Project	« Start O	ver Configure Parsing O	ptions		Proje	ct name CDC AIDS Diagn	osis CityLocation 819	6 Create Project
n Project	Notes	Location	Location Code	Year Diagnosed	Year Diagnosed Code	Race or Ethnicity	Race or Ethnici	ty Code Cases
art Project	1.	Akron, OH	80	1982	1982	Black (and also not Hispan	ic) 2054-5	1
in Project	2.	Akron, OH	80	1984	1984	White (and also not Hispan	ic) 2106-3	3
	3.	Akron, OH	80	1985	1985	White (and also not Hispan	ic) 2106-3	4
	4.	Akron, OH	80	1986	1986	Black (and also not Hispan	ic) 2054-5	3
	5.	Akron, OH	80	1986	1986	White (and also not Hispan	ic) 2106-3	9
	6.	Akron, OH	80	1987	1987	Black (and also not Hispan	ic) 2054-5	6
	7.	Akron, OH	80	1987	1987	White (and also not Hispan	ic) 2106-3	14
	8.	Akron, OH	80	1988	1988	Black (and also not Hispan	ic) 2054-5	11
	9.	Akron, OH	80	1988	1988	Hispanic	2135-2	1
	10.	Akron, OH	80	1988	1988	White (and also not Hispan	ic) 2106-3	15
	11.	Akron, OH	80	1989	1989	Black (and also not Hispan	ic) 2054-5	7
	12.	Akron, OH	80	1989	1989	White (and also not Hispan	ic) 2106-3	19
	13.	Akron, OH	80	1990	1990	Black (and also not Hispan	ic) 2054-5	6
	14.	Akron, OH	80	1990	1990	White (and also not Hispan	ic) 2106-3	23
	15.	Akron, OH	80	1991	1991	Black (and also not Hispan	ic) 2054-5	11
	16.	Akron, OH	80	1991	1991	White (and also not Hispan	ic) 2106-3	38
	17.	Akron, OH	80	1992	1992	Black (and also not Hispan	ic) 2054-5	17
	18.	Akron, OH	80	1992	1992	White (and also not Hispan	ic) 2106-3	41
	19.	Akron, OH	80	1993	1993	Black (and also not Hispan	ic) 2054-5	28
	20.	Akron, OH	80	1993	1993	White (and also not Hispan	ic) 2106-3	37
	21.	Akron, OH	80	1994	1994	Black (and also not Hispan	ic) 2054-5	15
	22.	Akron, OH	80	1994	1994	Hispanic	2135-2	1
	23.	Akron, OH	80	1994	1994	White (and also not Hispan	ic) 2106-3	31
	24.	Akron, OH	80	1995	1995	Black (and also not Hispan	ic) 2054-5	27
	25.	Akron, OH	80	1995	1995	Hispanic	2135-2	1
	26.	Akron, OH	80	1995	1995	White (and also not Hispan	ic) 2106-3	32
	27.	Akron, OH	80	1996	1996	Black (and also not Hispan	ic) 2054-5	17
	Parse	data as	Charact	er encoding				Update Preview
	CSV / T	SV / separator-based file	es Column	s are separated b	y	□ Ignore first	0 line(s) at begin	ining of file
	Line-bas	ed text files	O con	nmas (CSV)		Parse next	1 line(s) as colu	mn headers
	Fixed-wi	dth field text files	tab:	s (TSV)		Discard initial	0 row(s) of data	
	DO A 1	4 61	⊖ cus	tom \t		□ Load at most	n row(s) of data	
	PC-Axis	text illes	Escape	special characte	rs with \		o row(s) or data	
	JSON fil	es						
	RDF/N3	files				Parse cell text into	Store	blank rows
rsion 2.5 [r2407]	XML file	5				Quotation marks an	⊠ Store	biank cells as nulls
Help	Open De	ocument Format spreadshe	eets			to enclose cells co	ntaining (file n	file source ames, URLs)

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

Cyberinfrastructure for Network Science Center project

	,	Application Tools Refine	- C	
File Hom	e Share View	Manage		
$\leftarrow \rightarrow \cdot \cdot 1$	h 📙 > This PC > Lo	ocal Disk (C:) → Users → mgir	nda > Refine v 🖏 Search Refine	
📥 Ouick acc	C:\Users\mg	ginda\Refine\google-refine.exe		
	13:38:23.373 13:38:23.380	[refine_ser	<pre>vver] Starting Server bound to '127.0.0.1:3333' (Oms) vver] Initializing context: '/' from 'C:\Users\mginda\Refine\web</pre>	app' (7ms)
Deskton	13:38:24.175 ables (795ms	[refine_ser	over] Failed to use jdatapath to detect user data path: resortin	g to environment var
J Downloa	13:38:24.176	[refine_ser	ver] Failed to use jdatapath to detect user data path: resortin	g to environment var
Pictures	13:38:24.183	[ref	ine] Starting Google Refine 2.5 [r2407] (7ms)	
mginda				
ConeDrive				
This PC	(127 0 0 1·33	33	C Q OpenRefine voutube → ☆ 自 📄 ▾ Z 👃 🎓	
Desktop				
🔮 Docume	Google refine	A power tool for working with messy dat	a.	
🕹 Downlo	Create Project	Create a project by importing	g data. What kinds of data files can I import?	
👌 Music	Open Project	TSV, CSV, *SV, Excel (.xls and can be added with Google Refin	I .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Su ne extensions.	
Pictures	Import Project	Get data from	Locate one or more files on your computer to upload:	
Videos		This Computer	Browse No files selected.	
🏪 Local Di		Web Addresses (URLs)	Next	
		Clipboard		
🛫 mginda		Google Data		
rojects				
KINGSTO				
💣 Network				
]
	Version 2.5 [r2407]			
	Help About			

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

0 127.0.0.1:333	33 A power tool for working with messy data	C Q OpenRefin	ne youtube 🦻 🏠 🗎 📄 🖿	· Z + â	Once (want t	OpenRei	rine has a proje	startec ct. Firs	d. You w t. you n
Create Project Open Project	Create a project by importing TSV, CSV, *SV, Excel (.xls and can be added with Google Refin	data. What kinds of dat .xlsx), JSON, XML, RDF a e extensions.	a files can I import? as XML, and Google Data document	ts are all supported. Su	to sele	ect your	data sou	irce, w	which in t
mport Project	Get data from	Locate one or more file	es on your computer to upload:		case is	a samp	le data f	ile on t	this
	This Computer	Browse No files	selected.		compi	iter			
	Web Addresses (URLs)	Next »			compt				
	Clipboard				"CDC-	AIDS-Dia	agnosis-	CityLo	cation-
	Google Data				8196.t Once t	x t" :he data	file is up	oloade	d, hit ne
			📵 File Upload						×
			\leftarrow \rightarrow \checkmark \uparrow \square \Rightarrow This	s PC > Desktop > Sci2-1	DC-data → Geo		v Ö	Gearch Geo	Q
			Organize 👻 New folder					:== :== :===	- 🔳 🔞
			A Quick access	Name		Date modified	Туре	Size	
					-Cityl ocation-8196				
			🗄 Documents 🖈	CDC-AIDS-Diagnosis	citycocation orbonii	6/9/2016 9:18 AM	Text Document	403 KB	
			🔮 Documents 🖈 🔜 Desktop 🖈	STD-Morbid-Chlam	dia-20022013.csv	6/9/2016 9:18 AM 6/7/2016 4:25 PM	Text Document Microsoft Excel C	403 KB 31 KB	
Version 2.5 [r2407] Help About			 Documents * Desktop * Downloads * Pictures * mginda * OneDrive This PC Desktop 	STD-Morbid-Chlamy	dia-20022013.csv rrhea-20022013.csv	6/9/2016 9:18 AM 6/7/2016 4:25 PM 6/7/2016 4:25 PM	Text Document Microsoft Excel C Microsoft Excel C	403 KB 31 KB 31 KB	
Version 2.5 [r2407] Help About			 Documents Desktop Desktop Downloads Pictures mginda OneDrive This PC Desktop Documents 	STD-Morbid-Chlam	dia-20022013.csv rrhea-20022013.csv	6/9/2016 9:18 AM 6/7/2016 4:25 PM 6/7/2016 4:25 PM	Text Document Microsoft Excel C Microsoft Excel C	403 KB 31 KB 31 KB	
Version 2.5 [r2407] Help About			 Documents Desktop Downloads Pictures mginda OneDrive This PC Desktop Documents Documents Downloads 	STD-Morbid-Chlam	dia-20022013.csv rrhea-20022013.csv	6/9/2016 9:18 AM 6/7/2016 4:25 PM 6/7/2016 4:25 PM	Text Document Microsoft Excel C Microsoft Excel C	403 KB 31 KB 31 KB	
Version 2.5 [r2407] Help About			 Documents Desktop Downloads Pictures mginda OneDrive This PC Desktop Documents Downloads Music Pictures 	STD-Morbid-Chlam	dia-20022013.csv rrhea-20022013.csv	6/9/2016 9:18 AM 6/7/2016 4:25 PM 6/7/2016 4:25 PM	Text Document Microsoft Excel C Microsoft Excel C	403 KB 31 KB 31 KB	
Version 2.5 [r2407] Help About			 Documents Desktop Downloads Pictures mginda OneDrive This PC Desktop Documents Downloads Music Pictures File name 	me: CDC-AIDS-Diagnosis	CityLocation-8196.txt	6/9/2016 9:18 AM 6/7/2016 4:25 PM 6/7/2016 4:25 PM	Text Document Microsoft Excel C Microsoft Excel C	403 KB 31 KB 31 KB	~

Cyberinfrastructure for Network Science Center

Open Refine: Creating an CDC HIV/AIDS Diagnosis data

Help

> last »

Ethnicit

project

Cyberinfrastructure for Network Science Center

Project	« Sta	rt Over	Configure Parsing (Options		P	oject nam	e CDC AIDS Diagno	osis CityLocation 8	3196 C	Create Project	»
roiect	N	lotes L	ocation	Location Code	Year Diagnosed	Year Diagnosed Co	de Race	or Ethnicity	Race or Ethr	nicity Code	Cases	^
During a	1.	A	Akron, OH	80	1982	2 19	82 Black	(and also not Hispani	c) 2054-5		1	
Fioject	2.	ŀ	Akron, OH	80	1984	19	84 White	(and also not Hispan	ic) 2106-3		3	
	3.	F	Akron, OH	80	1985	i 19	85 White	(and also not Hispan	ic) 2106-3		4	
	4.	ł	Akron, OH	80	1986	i 19	86 Black	(and also not Hispani	c) 2054-5		3	
	5.	ŀ	Akron, OH	80	1986	5 19	86 White	(and also not Hispan	ic) 2106-3		9	
	6.	ŀ	Akron, OH	80	1987	/ 19	87 Black	(and also not Hispani	c) 2054-5		6	
	7.	F	Akron, OH	80	1987	7 19	87 White	(and also not Hispan	ic) 2106-3		14	
	8.	ŀ	Akron, OH	80	1988	19	88 Black	(and also not Hispani	c) 2054-5		11	
	9.	F	Akron, OH	80	1988	3 19	88 Hispa	nic	2135-2		1	
	10.	F	Akron, OH	80	1988	8 19	88 White	(and also not Hispan	ic) 2106-3		15	
	11.	ŀ	Akron, OH	80	1989) 19	89 Black	(and also not Hispani	c) 2054-5		7	
	12.	ŀ	Akron, OH	80	1989) 19	89 White	(and also not Hispan	ic) 2106-3		19	
	13.	F	Akron, OH	80	1990) 19	90 Black	(and also not Hispani	c) 2054-5		6	
	14.	ŀ	Akron, OH	80	1990) 19	90 White	(and also not Hispan	ic) 2106-3		23	
	15.	F	Akron, OH	80	1991	19	91 Black	(and also not Hispani	c) 2054-5		11	
	16.	ŀ	Akron, OH	80	1991	19	91 White	(and also not Hispan	ic) 2106-3		38	
	17.	ŀ	Akron, OH	80	1992	2 19	92 Black	(and also not Hispani	c) 2054-5		17	
	18.	ŀ	Akron, OH	80	1992	2 19	92 White	(and also not Hispan	ic) 2106-3		41	
	19.	ŀ	Akron, OH	80	1993	3 19	93 Black	(and also not Hispani	c) 2054-5		28	
	20.	ŀ	Akron, OH	80	1993	3 19	93 White	(and also not Hispan	ic) 2106-3		37	
	21.	F	Akron, OH	80	1994	19	94 Black	(and also not Hispani	c) 2054-5		15	
	22.	ŀ	Akron, OH	80	1994	19	94 Hispa	nic	2135-2		1	
	23.	ŀ	Akron, OH	80	1994	19	94 White	(and also not Hispan	ic) 2106-3		31	
	24.	ŀ	Akron, OH	80	1995	5 19	95 Black	(and also not Hispani	c) 2054-5		27	
	25.	F	Akron, OH	80	1995	5 19	95 Hispa	nic	2135-2		1	
	26.	ŀ	Akron, OH	80	1995	5 19	95 White	(and also not Hispan	ic) 2106-3		32	
Г	27	4	Akron OH	80	1996	19	96 Black	(and also not Hispani	c) 2054-5		17	~
	Google	refin	e CDC AIDS Dia	gnosis CityLo	ation 8196 tx	t Permalink					Open	Exp
		or	Inda / Dada a	5244							Extens	ions.
	Facel / File			JJTT	0W5	Ohana 5 40 05	50 ·····				End and investigation	4 40
				Show as	s. Tows records	Snow. 5 10 25	50 TOWS			«	linst (previous	1 - 10
	Using fa	acets a	and filters 🛛 🔰		Notes	Location Loca	tion Code	Year Diagnosed	Year Diagnosed	Race o	r Ethnicity	Ra
				x 5	1.	Akron, OH	80	1982	1982	Black (and a	also not Hispanic)	2054-5
	Use facet	ts and fi	ilters to select subsets	1 값 되	2.	Akron, OH	80	1984	1984	White (and a	also not Hispanic)	2106-3
	filter met	hods fro	m the menus at the to	\$ \$ 5	3.	Akron, OH	80	1985	1985	White (and a	also not Hispanic)	2106-3
	of each d	lata colu	umn.	다 다	4.	Akron, OH	80	1986	1986	Black (and a	also not Hispanic)	2054-5
on 2.5 (r240	Not	how to	ant started?	<u> </u>	5.	Akron, OH	80	1986	1986	White (and a	also not Hispanic)	2106-3
	Watch th	NOW TO	ger started?	\$ 5	6.	Akron, OH	80	1987	1987	Black (and a	also not Hispanic)	2054-5
Help	watch u		i o o noti ata	\$ 5	7.	Akron, OH	80	1987	1987	White (and a	also not Hispanic)	2106-3
About				<u>\$</u>	8.	Akron, OH	80	1988	1988	Black (and a	also not Hispanic)	2054-5
				😒 🗐	9.	Akron, OH	80	1988	1988	Hispanic		2135-2
				☆ 57	10.	Akron, OH	80	1988	1988	White (and a	also not Hispanic)	2106-3

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.

Cyberinfrastructure for 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.

Cyberinfrastructure for 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. Cyberinfrastructure for Network Science Center Open Refine: Cleaning the CDC HIV/AIDS Diagnosis data

Go	ogl	le r	əfin	e CDC AIDS Diagr	iosis C	tyLocatio	n 8196 t	xt Permalink					
Face	et / F	ilter	U	Indo / Redo o	5	253 ma	ching I	rows (5311	total)				
Refre	sh			Reset All Remove	AII S	now as: ro	ws record	s Show:	5 10 25	50 rows			
× Not	es			change invert re	set 🔽	All	Notes	Location		Locat	ion C	Code 💌 Year Diagnosed 💌) 1
53 cho	oices	Sort	by: n	ame count Clust	er 🖄	· 🗐 1.	Facet	•				80 1982	
Sugge	sted	Cita	tion:	US	A 23	· · · 2.	Text filter					80 1984	
Depar	tmen HHS	t of H	lealth	h and Human Services			Edit cells	s 🕨 -				80 1985	
Prever	ntion), CC (CD	C), 1	Tor Disease Control and			Edit colu	mn 🕨	Split into s	everal colu	mns	1986	
Title:	52	53	ma	tching rows (53)	11 total								
Vital	01				- 5 40	05 50							
in a r	Sho	ow a	s: ro	ows records Show	1:510	25 50 rov	vs						
metro		All		Location	💌 Lo	cation Cod	e 💌 Yea	ar Diagnosed	Year	Diagnose	1	Race or Ethnicity	Race o
(blar			1.	Akron, OH		8	0	1982	Facet		•	ick (and also not Hispanic)	2054-5
Facet	☆	Ч	2.	Akron, OH		8	0	1984	Text filt	er		nite (and also not Hispanic)	2106-3
			3.	Akron, OH		8	0	1985	Editor			nite (and also not Hispanic)	2106-3
	ಬ	Ы	4.	Akron, OH		8	0	1986	Edit ce	lis		ick (and also not Hispanic)	2054-5
			5.	Akron, OH		8	0	1986	Edit col	lumn		Split into several columns	S
	ಬ	Ы	6.	Akron, OH		8	0	1987	Transp	ose		Add column based on thi	s column
			7.	Akron, OH		8	0	1987	Sort			Add column by fetching U	RIS
	ជ	9	8.	Akron, OH		8	0	1988	View			Add columns from Freeb	
			9.	Akron, OH		8	0	1988			·	Add columns from Freeb	d5e
	ជ	5	10.	Akron, OH		8	0	1988	Recon	cile	•	Rename this column	
	ਨਿ		11.	Akron, OH		8	0	1989		1989	Bl	Remove this column	
	ដ	5	12.	Akron, OH		8	0	1989		1989	W	Move column to beginnin	0
			13.	Akron, OH		8	0	1990		1990	Bk	move column to beginnin	y
	ដ	5	14.	Akron, OH		8	0	1990		1990	W	Move column to end	
			15.	Akron, OH		8	0	1991		1991	Bk	Move column left	
	ដ	5	16.	Akron, OH		8	0	1991		1991	W	Move column right	
	শ্ব		17.	Akron, OH		8	0	1992		1992	Bla	ack (and also not mispanic)	2004-0
	5	F.1	18	Akron OH			0	1992		1992	10/0	hite (and also not Hispanic)	2106-3

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.

> 14 11

53	11	row	/s						
Sho	ow a	s: ro	ws records Show	: 5 10 25 50 rows					
•	AII		Location	Location Code	Year Diagnosed	•	Year Diagnosed	Race or E	thnicity
	Ч	1.	Akron, OH	80	Facet I	Þ	Text facet		not Hispanic)
없	Ч	2.	Akron, OH	80	Text filter		Numeric facet		not Hispanic)
		3.	Akron, OH	80			Timeline facet		not Hispanic)
슔	Ц	4.	Akron, OH	80	Edit cells	•			not Hispanic)
	Ы	5.	Akron, OH	80	Edit column		Scatterplot face		not Hispanic)
ជ	Ы	6.	Akron, OH	80	Transpose I	۲Ì	Custom text fac	et	not Hispanic)
	5	7.	Akron, OH	80	Sort		Custom numeri	c facet	not Hispanic)
슔	Ч	8.	Akron, OH	80			Customized fac	oto 🕨	not Hispanic)
	Ы	9.	Akron, OH	80	View	•	Custoniizeu iac]
없	Ы	10.	Akron, OH	80	Reconcile I	•	1988	White (and als	o not Hispanic)
		11.	Akron, OH	80	1989		1989	Black (and als	o not Hispanic)
ជ	Ч	12.	Akron, OH	80	1989		1989	White (and als	o not Hispanic)





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	/ m	atch	i ng rows (5311	l total)						
Sh	ow a	as: rov	vs records Sh	ow: 5 10 25 50 ro	ws					
	AII		Location	Location Code	Year Diagnosed	Year Diagnosed	Race or Ethnicity	Race or Ethnicit	• 💌	Cases
	5	193.	Atlanta, GA	520	Before 1982	Data type: numbe	r v			3
었	5	194.	Atlanta, GA	520	Before 1982	1981				6
	5	300.	Bakersfield, CA	680	Before 1982	1301				2
☆	5	349.	Baltimore, MD	720	Before 1982				.::	1
슔	5	350.	Baltimore, MD	720	Before 1982	Apply Apply to	All Identical Cells Cancel			1
☆	5	443.	Bergen-Passaic, NJ	875	Before 1982	Enter	Ctrl-Enter Esc			1
প্ন	5	444.	Bergen-Passaic, NJ	875	Before 1982	1981	Black (and also not Hispanic)	2054-5		1
53	G	445.	Bergen-Passaic, NJ	875	Before 1982	1981	White (and also not Hispanic)	2106-3		4

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

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

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



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

Cuberinfrastructure for

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



the Geocoding API may only be used in conjunction with a Google map; geocoding result

It is your responsibility to observe the terms of service, such as, in this case, plotting your data on

Clustering

Clustering In Depth Column Editing Data Sources

Cyberinfrastructure for 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

Add column by	fetching URL	s based on column Location	
New column name	Geocode	Throttle delay 200 milliseconds	5
On error	Set to blank	○ store error	
Formulate the URI	s to fetch:		
Expression	La	inguage $\ $ Google Refine Expression Language (GREL) $\ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \ \$	
"http://maps.g address=" + es Preview H	oogle.com/map cape(value, " story Starred	No syntax error.	
row value	"ht ad	ttp://maps.google.com/maps/api/geocode/json?sensor=false& Aldress=" + escape(value, "url")	
1. Akron, Ol	H http add	p://maps.google.com/maps/api/geocode/json?sensor=false& dress=Akron%2C+OH	
29. Albany-Se NY	chenectady, http add	p://maps.google.com/maps/api/geocode/json?sensor=false& dress=Albany-Schenectady%2C+NY	
75. Albuquero	que, N.M. http add	p://maps.google.com/maps/api/geocode/json?sensor=false& dress=Albuquerque%2C+N.M.	
121. Allentown	, PA http	p://maps.google.com/maps/api/geocode/json?sensor=false& v	
OK Crea on c grel: /json	te column Geo olumn Location "http://maps.go ?sensor=false& 4% complete	ocode at index 1 by fetching URLs based n using expression oogle.com/maps/api/geocode &address=" + escape(value, "url")	

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")

Cyberinfrastructure for Open Refine: Geocoding the CDC HIV/AIDS Diagnosis data

Add column based on column Geocode	
New column name lation	
On error	copy value from original column
Expression Language Google F	Refine Expression Language (GREL) \vee
<pre>with(value.parseJson().results[0].geomet pair.lat +", " + pair.lng) Proving History Stared Help</pre>	No syntax error.
Preview History Starred Help	^
row value wit pai	h(value.parseJson().results[0].geometry.location, ir, pair.lat +", " + pair.lng)
 { "results" : [{ "address_components" : 41. [{ "long_name" : "Akron", "short_name" : "Akron", "types" : ["locality", "political"] }, { "long_name" : "Summit County", "short_name" : "Summit County", "types" : ["administrative_area_level_2", "political"] }, { "long_name" : "Ohio", "short_name" : "OH", "types" : [0814447, -81.51900529999999
OK Cancel	

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 "lation", and the GREL expression for parsing is: with(value.parseJson().results[0].geometry .location, pair, pair.lat +", " + pair.lng)

refor Open Refine: Geocoding the CDC HIV/AIDS Diagnosis data

w: 5 1	0 25 50 rows				Refresh	Reset All	Remove All
	Iation	_	Year Diagnosed		X lation		
	Facet 🕨	Text fac	et	lac			change
NY	Text filter	Numer	ic facet	isp	104 choices Sort by: nai	me count	Cluster
	Edit cells	Timelin Scatter	e facet plot facet	/hi isp /hi	18.4655394, -66.105735 21.3069444, -157.85833 25.7616798, -80.191790	55 1 333 1 02 1	^
	Transpose 🕨	Custon	n text facet	lac	26.1224386, -80.13731740000001 1		
	Sort View ▶	Custon Custon	n numeric facet nized facets	lac ∕hi	26.3524992, -98.124288 26.7153424, -80.053374	34 1 46 1	
	Reconcile		1983	Blac	27.3364347, -82.53065269999999 1		
	40.8567662, -74.128476	4	1981	Am	27.8498701, -82.482767	75 1	
	33.5206608, -86.802489	999999999	1982	Whi	28.5383355, -81.379236	65-1	
	42.3600825, -71.058880	1	1981	Blac	-00-0400447 04-000002	<u>14</u>	~

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 **"lation"** 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 <u>Clustering and Editing algorithms</u> for text data. 65

104 n	natcł	ning rows (5311 total))				
Show a	s: row	records Show: 5 1	0 25 50 rows				
		Location	Iation 1	•	lation 2	Year Diagnosed	Race or
☆ 🗐	1.	Akron, OH	Facet	F	.51900529999999	1982	Black (and als
☆ 57	29.	Albany-Schenectady, NY	Text filter		-73.7562317	1983	Hispanic
☆ 57	75.	Albuquerque, N.M.			-106.6055534	1983	White (and al
3 9	121.	Allentown, PA	Edit cells	۲.	-75.4901833	1982	Hispanic
	161.	Ann Arbor, MI	Edit column	١.	Split into several	l columns	nd al
3 57	193.	Atlanta, GA	Transpose	F	Add column bas	ed on this column	nd als
3 5	251.	Austin, TX	Sort		Add column by f	atching LIPL e	nd als
3 57	300.	Bakersfield, CA	View				nd al
	349.	Baltimore, MD	VIEW	•	Add columns fro	m Freebase	nd als
3 57	408.	Baton Rouge, LA	Reconcile	F	Rename this co	lumn	nd als
3 5	443.	Bergen-Passaic, NJ	40.8567662		Remove this col	umn	n Indi
3 57	505.	Birmingham, AL	33.5206608	-86			nd al
3 57	536.	Boston, MA	42.3600825		Move column to	beginning	nd als
3 57	608.	Buffalo, NY	42.88644679999999		Move column to	end	
	663.	Charleston, SC	32.7764749	-79	Move column lef	t	nd al
3 57	699.	Charlotte, NC	35.2270869		Move column rig	ht	nd al
3 57	747.	Chicago, IL	41.8781136		-01.0231302	1901	DIACK (and als
3000	***	hd also not His	panic) 2054-5	5	-84.5120196	1983	Black (and als
			-5		for no limit)	1	
		Enter new column n	ame _s	2			
		ation 1					
-	10						
9			-5	5			
			-2	2		.::	
		OK (Cancel	3	15		
14	5		· /•	83			
	130	not His	panic) 2054-5	1			

After determining if there are any duplicate locations, we need to split the **lation** 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 Cyberinfrastructure for Network Science Center

104 choices Sort by: name count Cluster Tutsa, OK, I Image: Count Cluster Software: Count Cluster Image: C	× Location		cha	nge					After tran	sformin	g the
Tucson, AZ Image: AZ	104 choices	Sort by: name count	Clust	er					latitude a	nd longi	itude dat
Tulka, OK 1 Tulka, OK 1 Search / Filter Undo / Redo 11	Tucson, AZ 1			^						ovu stor	filtoring
Vallejo-Fairing Ventura, CA Washington, No column named lation Reask all Ramove all Reasks Reask all Ramove all Reasks Show as: rooms record Reasks Nore of the rooms rec	Tulsa, OK 1	Eagot / Eiltor		5253	matching rows				we may n	ow stop	intering
Ventura, CA Vester Number Vester Num	Vallejo-Fairfie		A.II.	Show as: rows records Show 5 10 25 50 rows					the Location column, and		
Note of the second of	Ventura CA	Refresh Reset All Ren	nove All		Jesetien	₩. 5 10 25 5		Veer Diagnes			,
Wash Pall Fig We column named lation Image of the pale of	Washington.	× lation	change	· All	1. Akron. OH	41.0814447	-81,51900529999999	198	show blar	ik value	S.
West Palm E View (and also not Huganic) 200-3 4 Winnington, Youngstown, S - - 1965 Bick (and also not Huganic) 200-3 9 Youngstown, Youngstown, S - - - 1965 Bick (and also not Huganic) 200-3 9 Youngstown, S - - - - 1967 Bick (and also not Huganic) 200-3 16 Yaungstown, S - - - - 1967 Bick (and also not Huganic) 200-3 16 Yaungstown, - - - - 1968 Hear (and also not Huganic) 200-3 16 Yaungstown, - - - - 1968 Hear (and also not Huganic) 200-3 16 Yaung Saungstown, 1981.00 - 19 - - 198 Hear (and also not Huganic) 200-3 16 Yaung Saung Sa	vvasnington,	No column named lation			2.			198	White (and also not Hispanic)	2106-3	3
Wichita, KS Wilmington, Youngstown,	West Palm E				3.			198	5 White (and also not Hispanic)	2106-3	4
Villinington, Youngstown, (blank) 5149 Vice Diagnosed S M	Wichita KS				4.			198	6 Black (and also not Hispanic)	2054-5	3
Wilmington, Youngstown, (blank) 5149 C <thc< th=""> C <thc< th=""></thc<></thc<>					5.			1980	6 White (and also not Hispanic)	2106-3	9
Youngstown (blank) 5149 Yme Make (and also not Hispan) 210-3 14 Gel (blank) 5149 Image (blank) 5149 <td>Wilmington,</td> <td></td> <td></td> <td>☆ 되</td> <td>6.</td> <td></td> <td></td> <td>198</td> <td>7 Black (and also not Hispanic)</td> <td>2054-5</td> <td>6</td>	Wilmington,			☆ 되	6.			198	7 Black (and also not Hispanic)	2054-5	6
Nome Nome <th< td=""><td>Youngstown</td><td></td><td></td><td></td><td>7.</td><td></td><td></td><td>198</td><td>7 White (and also not Hispanic)</td><td>2106-3</td><td>14</td></th<>	Youngstown				7.			198	7 White (and also not Hispanic)	2106-3	14
(blank) 5149 S <t< td=""><td>roungotonn,</td><td></td><td></td><td></td><td>8.</td><td></td><td></td><td>198</td><td>Black (and also not Hispanic)</td><td>2054-5</td><td>11</td></t<>	roungotonn,				8.			198	Black (and also not Hispanic)	2054-5	11
Sect by choice Image: Sect by choice <td>(blank) 5149</td> <td></td> <td></td> <td></td> <td>9.</td> <td></td> <td></td> <td>198</td> <td>8 Hispanic</td> <td>2135-2</td> <td>1</td>	(blank) 5149				9.			198	8 Hispanic	2135-2	1
Albert by Childe Image: Section of the sectin of the section of the section of the section of the section of t	Fact by choice				10.			198	8 White (and also not Hispanic)	2106-3	15
X Year Diagnosed change reat If I I <t< td=""><td>acer by choice</td><td></td><td></td><td></td><td>11.</td><td></td><td></td><td>198</td><td>9 Black (and also not Hispanic)</td><td>2054-5</td><td>7</td></t<>	acer by choice				11.			198	9 Black (and also not Hispanic)	2054-5	7
Image: State of the state		X Year Diagnosed char	nge reset	23 51	12.			198	9 White (and also not Hispanic)	2106-3	19
$ \begin{vmatrix} \mathbf{k} & \mathbf{k}$					13.			1990	Black (and also not Hispanic)	2054-5	6
1 9 4					14.			1990	Vvnite (and also not Hispanic)	2106-3	23
1,981.00 -1,997.00 17 Image: 18 Image: 18 <t< td=""><td></td><td></td><td>Ψ</td><td></td><td>15.</td><td></td><td></td><td>199</td><td>1 Black (and also not Hispanic)</td><td>2054-5</td><td>11</td></t<>			Ψ		15.			199	1 Black (and also not Hispanic)	2054-5	11
Wumeric Using <		1,981.00 — 1,997.00			10.			199	Plack (and also not Hispanic)	2100-3	30
Landing in Bank Link In			Error		17.			195	2 White (and also not Hispanic)	2004-0	41
X Locationchange a		5253 0 58	0		10.			199	Black (and also not Hispanic)	2054-5	28
A Location Change No No <td></td> <td></td> <td></td> <td></td> <td>20</td> <td></td> <td></td> <td>199</td> <td>3 White (and also not Hispanic)</td> <td>2106-3</td> <td>37</td>					20			199	3 White (and also not Hispanic)	2106-3	37
104 choices Sort by: name count Cluster Tucson, AZ 1 2 Tulsa, OK 1 2 Vallejo-Fairfield-Napa, CA 1 2 Ventura, CA 1 2 Vest Palm Beach, FL 1 2 Vininington, DE 1 2 Youngstown, OH 1 2 (blank) 5149 2 Facet by choice counts 2 Vest 3 Vallejo-State 3 Vest 3 Vest 4 Volta (ad also not Hispanic) 206-3 Vest 2 Volta (ad also not Hispanic) 206-3 Vest 2 Volta (ad also not Hispanic) 206-3 Volta (ad also not Hispanic) 206-3			change		21			199	4 Black (and also not Hispanic)	2054-5	15
Tucson, AZ 1 Image: Constraint of the		104 choices Sort by: name count	Cluster		22			199	4 Hispanic	2135-2	1
Tulsa, OK 1 Vallejo-Fairfield-Napa, CA 1 Vallejo-		Tucson, AZ 1	^		23.			1994	4 White (and also not Hispanic)	2106-3	31
Vallejo-Partheld-Mapa, CA 1		Tulsa, OK 1			24.			199	5 Black (and also not Hispanic)	2054-5	27
Ventura, CA 1 Vashington, DC 1		Vallejo-Fairfield-Napa, CA 1			25.			199	5 Hispanic	2135-2	1
West Palm Beach, FL 1 Image: Constraint of the second		Washington DC 1			26.			199	5 White (and also not Hispanic)	2106-3	32
Wichita, KS 1 Image: Constraint of the		West Palm Beach, El. 1			27.			1990	6 Black (and also not Hispanic)	2054-5	17
Wilmington, DE 1		Wichita, KS 1			28.		edit	199	6 White (and also not Hispanic)	2106-3	28
Youngstown, OH 1 Image: Constraint of the system Image: Constraint of the syst		Wilmington, DE 1			29. Albany-Schenectady, N	Y 42.6525793	-73.7562317	198	3 Hispanic	2135-2	2
(blank) 5149 (x) (x) <t< td=""><td></td><td>Youngstown, OH 1</td><td></td><td>☆되</td><td>30.</td><td></td><td></td><td>198</td><td>3 White (and also not Hispanic)</td><td>2106-3</td><td>4</td></t<>		Youngstown, OH 1		☆되	30.			198	3 White (and also not Hispanic)	2106-3	4
Facet by choice counts Image: Column 2 Ima		(blank) 5149			31.			1984	4 Black (and also not Hispanic)	2054-5	2
· · · · · · · · · · · · · · · · · · ·		Facet by choice counts	~		32.			1984	4 Hispanic	2135-2	3
		=			33.			1984	4 White (and also not Hispanic)	2106-3	14

Cyberinfrastructure for Network Science Center 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.

Google refine CDC AIDS Diagnosis CityLocation 8196 txt Permalink											
Facet / Filter Undo	5253 matching rows (5311 total)										
Refresh	Show as: rows records Show: 5 10 25 50 rows										
× lation	change		Location		Iatitude	Iongitude	Year Diagnosed	Race or Ethnicity	Race or Ethnicit	Cases	
		☆ 🏹 1	. Facet	•	41.0814447	-81.51900529999999	1982	Black (and also not Hispanic)	2054-5	1	
No column n	amed lation	☆ 듸 2	Text filter				1984	White (and also not Hispanic)	2106-3	3	
	☆ 🍕 3	. Edit calla		1		1985	White (and also not Hispanic)	2106-3	4		
	☆ 듸 4	☆ · · · Edit cells	•	Transform		1986	Black (and also not Hispanic)	2054-5	3		
	☆ 🎝 5	Edit column	•	Common tra	nsforms 🕨	1986	White (and also not Hispanic)	2106-3	9		
	☆ 듸 6	Transpose	anspose 🕨 🕨			1987	Black (and also not Hispanic)	2054-5	6		
	☆ 🎝 7	Sort		Blank down		1987	White (and also not Hispanic)	2106-3	14		
		☆ 되 8	-7 8.			1988	Black (and also not Hispanic)	2054-5	11		
		☆ 🎝 9	. View		Split multi-valued cells Join multi-valued cells		1988	Hispanic	2135-2	1	
		☆ 듸 1	0. Reconcile	►			1988	White (and also not Hispanic)	2106-3	15	
		☆ 🎝 1	1.		Olympics		1989	Black (and also not Hispanic)	2054-5	7	
		☆ 더 1	2.		Cluster and	eail	1989	White (and also not Hispanic)	2106-3	19	
Year Diagnosed	change reset	☆ 🎝 1	3.				1990	Black (and also not Hispanic)	2054-5	6	
		☆ 듸 1	4.				1990	White (and also not Hispanic)	2106-3	23	
Û	☆ 🖓 1	5.				1991	Black (and also not Hispanic)	2054-5	11		
	☆ 대 1	6.				1991	White (and also not Hispanic)	2106-3	38		

Race or Ethnicit	Cases	Race or Ethnicity cha	inge
Facet 🕨	Text facet	6 choices Sort by: name count Clust	ter
Text filter	Numeric facet	American Indian /Alaskan Native 422	
Edit cells	Timeline facet	Asian / Pacific Islander 551	
	Scottorplat facat	Black (and also not Hispanic) 1392	
Edit column	Scallerpiol lacet	Hispanic 1201	
Transpose 🕨 🕨	Custom text facet	Unknown 151	
Sort	Custom numeric facet	White (and also not Hispanic) 1536	
View •	Customized facets	Facet by choice counts	
Reconcile •	15		
2054-5	7		
2106-3	19		

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.





Cyberinfrastructure for Network Science Center Open Refine: Exporting the CDC HIV/AIDS Diagnosis data

Open Export -	Help
Export project	pase 👻
Tab-separated value	v last »
Comma-separated value	7 1050 //
HTML table	
Excel	<u>^</u>
ODF spreadsheet	
Triple loader	
MQLWrite	
Custom tabular exporter	
Templating	
	-

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 **Commaseparated value**.



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

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

9:15 Tool Overview and Trouble Shooting

Cyberinfrastructure for Network Science Center

- Sci2 Overview scientometric analysis tool
- Open Refine data parsing, transformation, and editing tool

Tutorial Overview

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

Cyberinfrastructure for Sci2: Loading the CDC HIV/AIDS Diagnosis data

S :	Sci2 Tool	Research A			1 A					
File	Data Preparation Preprocess	sing Analysis Modeling Vi	isualization R Help							
	Load	Ctrl+Alt+O		배 Data Manager						
	Read Directory Hierarchy Facebook Google Scholar Elictr Reader	, Strea	ims. Proc. 8th ACM SIGKDD							
	Twitter Reader	😵 Select Files					×			
	Save	← → ~ ↑ 🔒 > Tł	his PC → Desktop → Sci2-CDC	-data > Geo >	ٽ ~	Search Geo	م			
	View View with	Organize 🔻 New fold	ler			===	- 🔳 😢			
	Merge Node and Edge Files	a OneDrive	Name		Date modified	Type	Size			
	Split Graph to Node and Edge	💻 This PC	CDC-AIDS-Diagnosis-Cit	bd acation_9106	6/9/2010 3:38 PIVI	Text Document	402 KB			
	Preferences	📃 Desktop	CDC-AIDS-Diagnosis-Cit	vl ocation-8196	6/9/2016 2:31 PM	Microsoft Excel C	403 KB			
	Converter Graph	Documents	STD-Morbid-Chlamydia	-20022013.csv	6/7/2016 4:25 PM	Microsoft Excel C	31 KB			
	Exit	🔶 Downloads	STD-Morbidity-Gonorrh	ea-20022013.csv	6/7/2016 4:25 PM	Microsoft Excel C	31 KB			
Π	1	b Music								
		Pictures								
	i Algorithm Name L	😽 Videos								
		🏪 Local Disk (C:)		Г						
l r	ф	KINGSTON (D:)			Load					×
	\$	🛖 mginda (\\smb.) 🛖 projects (\\smb.			The file 'C:\User Please select the	s\mginda\Desktop\Sci2-(format you would like to	CDC-data\Burst\CD o try.	C-Grants-20102015-Envir	mentalHealt	h.csv' can be
		- KINGSTON (D:)			Standard csv format					
		File n	name: CDC-AIDS-Diagnosis-City	/Location-8196-ge	NSF csv format Scopus csv format					
								Select	Cancel	Details >>

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

Cyberinfrastructure for Network Science Center



Sci2: Time Slicing and Aggregating CDC HIV/AIDSCyberinfrastructure for
Network Science CenterDiagnosis data

 > Workflow Manager > ➡ Workflow 1 > ⊕ Load > ⊕ Slice Table by Time > ⊕ Parameters 	Next, crea by right cl Workflow aggregatio	Next, create a new workflow in the Workflow manage by right clicking the menu area and selecting "New Workflow". This allows us to duplicate the next aggregation step.					
New Workflow Load	Then navi Data.	gate to Preproce	essing > Genera	al > Aggregate			
Sci2 Tool	1						
File Data Preparation Preprocessing Analysis Modeling Visual	lization R Help						
General > Extract Top N	% Records	🔳 Aggregate Data		×			
Alian With Calendar fal	Records	Aggregate data i	Aggregate data in the table based on a column.				
Date/Time Format: yyyy Geospatial > Aggregate Da Cumulative?: true Topical >	ita	Aggregate on column	Location	~ 😧			
Slice Table by Time was Networks >		latitude	Max	~			
Implementer(s): Russell Duhon Integrator(s): Russell Duhon Documentation: http://wiki.cns.iu.edu/display/CISHELL/Slice+Table+	longitude	Max	~ 😧				
Input Parameters:		Year Diagnosed	None	~ 📀			
Set the following parameters for the		Cases	None	~ 0			
Aggregate Data algorithm, and select	Delimiter for Location	None Sum Difference	•				

INS

0

Cancel

UN.

Difference Average

Min Max

Delimiter for Race or Ethnicity

Sci2: Time Slicing and Aggregating CDC HIV/AIDS

Diagnosis data

	А	В	С	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

Cyberinfrastructure for Network Science Center

> 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 Cyberinfrastructure for Network Science Center

×

ng	Visualization R	Help			
	General	>	🗖 🗖 🗤 Data Manager		
_	Temporal	>	A SV file: C\Users\mainda\		
	Geospatial	>	Proportional Symbol Map 1g of		
	Topical	>	Choropleth Map :rfor		
	Networks	>	Geospatial Network Layout with Base Map 19 of		
	Proportional Syn	mbol Maj	p		×
		aps geospatial coordinates as circles that can be size- and lor-coded in proportion to associated numeric data.			
	Subtitle	HI	V /AIDS Diagnosis Counts from before 1982		0
	Map	Uni	ited States	~	0
	Latitude	lati	tude	~	0
	Longitude	Igitude			0
	Size Circles By	Cas	ses	~	0
	Size Scaling	Lin	ear	~	0
	Color Circle Exterior	rs By Cas	ses	~	0
	Exterior Color Scalir	ng Lin	ear	~	0
	Exterior Color Range	e Wh	ite to Green	~	0
	Color Circle Interior	rs By No	ne (no coloring)	~	0
	Interior Color Scalin	ling Lin	ear	~	0
	Interior Color Range	e Wh	ite to Green	~	0
				OK	Cancel

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



Cancel

×

Q

403 KB

109 KB

31 KB

31 KB

-

Size

Sci2: Converting the Output PostScript Maps to PDF

Command Prompt Desktop app

Cyberinfrastructure for Network Science Center

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

0:5___

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.

Command Prompt				-	×	
crosoft Windows [Version 10.0	9.10240]			^	
) 2015 Microsoft	Corporation	. All rights res	served.			
\Users\mginda≻						
Command	Prompt				- [) X
Microsoft W	indows [Vers	ion 10.0.10240				
(C) 2015 M1	crosoft Corp	poration. All r	ignts reserved.			_
C:\Users\mg Volume in Volume Ser	inda≻DIR drive C has ial Number i	no label. is 8222-07E9				
Directory	of C:\Users	mginda				
05/18/2016	11:59 AM	<dir></dir>				
05/18/2016	11:59 AM	<dir></dir>				
01/04/2016	03:00 PM	<dir></dir>	.continuum			
05/24/2016	01:23 PM	<dir></dir>	.oracle_jre_usage			
03/10/2016	01:15 PM	<dir></dir>	.sqlworkbench			
01/04/2016	02:51 PM	<dir></dir>	3D Objects			
04/13/2016	03:51 PM	<dir></dir>	Contacts			
06/11/2016	02:06 PM	<dir></dir>	Desktop			
06/09/2016	06:45 PM	<dir></dir>	Documents			
03/30/2016	01:48 PM	<dir></dir>	DocumentsADNI			
06/06/2016	04:42 PM	<dir></dir>	Downloads			
04/13/2016	03:51 PM	<dir></dir>	Favorites			
04/13/2016	03:51 PM	<dir></dir>	Links			
01/14/2016	01:02 PM	<dir></dir>	logs			
01/04/2016	02:49 PM	<dir></dir>	Mallet			
04/13/2016	03:51 PM	<dir></dir>	Music			
05/18/2016	11:59 AM	<dir></dir>	OneDrive			
01/26/2016	05:11 PM	<dir></dir>	Pajek			
04/13/2016	03:51 PM	<dir></dir>	Pictures			
06/10/2016	09:37 AM	<dir></dir>	Refine			
04/13/2016	03:51 PM	<dir></dir>	Saved Games			

Cyberinfrastructure for 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
```

ov. Select Com	nmand Prompt			-	\times
	27 Dir(s) 134,791,598,08	30 bytes free		
:\Users\mg	inda≻cd C:\	Users\mginda\Docu	iments		
·\Users\mg	inda\Docume	nts>DTR			
Volume in	drive C has	no label.			
Volume Ser	ial Number	is 8222-07E9			
Directory	of C:\Users	\mginda\Documents			
6/09/2016	06:45 PM	<dtr></dtr>			
6/09/2016	06:45 PM	<dir></dir>			
6/06/2016	04:16 PM	<dir></dir>	Admin		
5/17/2016	09:38 AM	<dir></dir>	ADNI		
6/06/2016	10:07 AM	13,409	ADNI-JournalImpactFactors.xlsx		
2/10/2016	05:40 PM	23,568	ADNI-JournalList-countsstats.xlsx		
2/10/2016	06:22 PM	81,738	ADNI-JournalsStats Query.xlsx		
1/08/2016	03:16 PM	<dir></dir>	Adobe		
1/04/2016	02:32 PM	<dir></dir>	BBSRC		
6/09/2016	06:45 PM	13,926	Burst Detection.docx		
4/28/2016	12:16 PM	<dir></dir>	Conference		
2/07/2016	05:26 PM	3,318,937	Copy of icd102010enMeta-csv.txt		
4/13/2016	11:39 AM	<dir></dir>	Custom Office Templates		
1/04/2016	02:36 PM	<dir></dir>	DocumentationPrep		
5/16/2016	10:29 AM	<dir></dir>	Endnote		
5/23/2016	12:14 PM	<dir></dir>	IAI		
5/25/2016	11:44 AM	<dir></dir>	IVMOOC		
6/07/2016	10:06 AM	<dir></dir>	LA		
3/21/2016	09:27 AM	95,721	LIGHTR-ESRICONFERENCEnotes.rtf		
3/08/2016	12:24 PM	360	modelLCSH.txt		

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

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

Cyberinfrastructure for Network Science Center

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.



Sci2: Converting the Output PostScript Maps to PDF

You can view the PDF file with Adobe Reader. More information on this script tool is available in the <u>PS2PDF documentation site</u>.

Geospatial Visualization (Proportional Symbol Map)

Generated from Aggregation performed using unique values in 'Location' column. Jun 09, 2016 | 05:19:34 PM EDT





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?



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

Cyberinfrastructure for Network Science Center

- Sci2 Overview scientometric analysis tool
- Open Refine data parsing, transformation, and editing tool

Tutorial Overview

Gephi – network visualization

9:45 Geospatial Analysis

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

11:00 Topical/Temporal Analysis: Burst Detection

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

12:30 Lunch

1:30 Network Analysis

- Introduction to Network Analysis
- Network Analysis: Bimodal networks with Morbidity Data
- Network Analysis: Co-authorship Network with CDC publications
 4:00 Wrap-up

4:30 Adjourn

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

Cyberinfrastructure for Network Science Center 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.



Burst Detection	Burst Detection						
Perform Bu	rst Detection on time-series textual data.						
Gamma	1.0	0	Â				
Density Scaling	2.0	0					
Bursting States	1	0					
Date Column	Abstract	0					
Date Format	уууу	0					
Burst Length Unit	Years	0	Ξ				
Burst Length	1	0					
Text Column	Abstract	0					
Text Separator		0					
Document Column	No document column.	0					
Ignore Input Wit	h Empty Text	0	-				
	ОК	Can	cel				

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



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

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

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

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

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

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

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

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

Text Separator – delimits the tokens in the text column.



Questions?



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

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

9:15 Tool Overview and Trouble Shooting

Cyberinfrastructure for Network Science Center

- Sci2 Overview scientometric analysis tool
- Open Refine data parsing, transformation, and editing tool

Tutorial Overview

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

\$	ci2 Tool								
File	Data Preparation Preprocess	sing Analysis Modelin	g Visualization R H	elp					
	Load	Ctrl+Alt+O		- 8	🕮 Data Manager				
	Read Directory Hierarchy Facebook Google Scholar Flickr Reader Twitter Reader	} }	Streams. Proc. 8th ACM Iurst+Detection	SIGKDD					
	Save	Ctrl+Alt+S							
	View View with	Ctrl+Alt+V Ctrl+Alt+W	'emporal + Bar + Graph						
	Merge Node and Edge Files Split Graph to Node and Edge Preferences	Files	'emporal + Bar + Graph		Select Files	DC ▶ sampledata ▶ scientometrics ▶ isi		✓ 4→ Search isi	x م
	Evit				Organize 🔻 New folder			· · · · · · · · · · · · · · · · · · ·	
			Remove all completed		👢 Conference 🔺	Name	Date modified	Туре	Size
					L DocumentationPr	FourNetSciResearchers.isi	2/26/2015 4:50 PM	ISI Common Expor	875 KB
[I Algorithm Name I	Date Time	% Complete		 AmazonCloudEr IVMoocForumEc SCI2 Twitter WorkflowManac Packages Research educatic sci2-IVMOOC SCI2NWB Vis Tools Data Cle 	Test5Papers.isi	2/26/2015 4:50 PM	ISI Common Expor	3 KB
I	oad <i>FourNe</i>	tSciReseau	rchers isi		File name	: FourNetSciResearchers.isi		▼ [*] .* Open	Cancel

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

Suberinfrastructure for Network Science Center Temporal Analysis: Burst Detection

Sci2 Tool		
File Data Preparation Preprocessing Analysis Modeling Visualization	R Help	
General	🗖 🗇 🔤 Data Manager	
Load was selected.	ISI Data: D:\Users\mginda\Desktop\sci2-IVMOOC\sampled	
Documentation: http:// Geospatial CISHELL/Data + Formats	361 Unique ISI Records.2	
Found old-style ISI/Wet Topical Reconciled Journal Na	mes	
Found old-style ISI/Wet Networks Lowercase, Tokenize, S	Stem, and Stopword Text	
Found old-style ISI/Web Of Knowledge file.	Lowercase, Tokenize, Stem, and Stopword Text	_
The original 361 records have been processed to remove duplicate unique I	SI IDs leavin	
361 records.	Research Field	
Wrote log to		
D:\Users\mginda\AppData\Local\Temp\isiduplicateremoverlog7981651116	5405064838.	•
Loaded: D:\Llsers\mainda\Desktop\sci2-IVMOOC\sampledata\scientometrics\isi\Fou	IrNetSciRese	
hers.isi		S
	Subject Category	•
Bomovo From Liet Romovo completed automatically. Romovo all con	substal Supplement	•
Remove From List		
	v inte	*
! Algorithm Name Date Time % Complete	🔲 Unique ID	•
✓ Load 04/03/2015 11:36:49	Version Number	•
	Volume	@
	File Name	
	Cite Me As	😧 🗸
		OK Cancel

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

Then select **Title** from the input parameters

Temporal Analysis: Burst Detection

Sei2 Tool				
File Data Preparation Preprocessing Analysis Modeling Visualization R He	elp			
📮 Console	Data Manager			
Journal Name (Abbreviated): false Publication Type: false E-mail Addresses: false Conference Host: false Publisher: false ISI Document Delivery Number: false Special Issue: false File Type: false Version Number: false Reprint Address: false Unique ID: false City of Publisher: false Authors (Full Names): false	IISI Data: D:\Users\mginda\Desktop\sci2-IVMOOC ▲ III 361 Unique ISI Records.2 IIII with normalized Abstract, Title	\sampledata\scientometri		x
New Separator:	III	Burst Detection	6	
Scheduler C	Workflow Manager	Perform B	urst Detection on time-series textual data.	
Remove From List Remove completed automatically Remove all comple	Workflow 1	Gamma	1.0	0
Image: Algorithm Name Date Time % Complete	 Dovercase, Tokenize, Stem, and Stopword Parameters 	Density Scaling	2.0	0
✓ Lowercase, Tokeniz 04/03/2015 11:38:53 ✓ Load 04/03/2015 11:36:49		Bursting States	1	0
값		Date Column	Publication Year 👻	0
		Date Format	уууу	0
		Burst Length Unit	Years 🔻	0
		Burst Length	1	0
Highlight the table 'with normali	zed Title'	Text Column	Title 🔻	0
Select Analysis > Temporal > Burs	st	Text Separator	1	0
Detection		Document Column	No document column.	
Then set the parameters to what to the right	: is shown	I gnore Input With	n Empty Text	

CNS Cyberinfrastructure for Network Science Center Temporal Analysis: Burst Detection

Sci2 Tool		
File Data Preparation Preprocessing Analysis Mod	leling Visualization R	Help
E Console		🔐 Data Manager
aggreg starts: 1990 ends: 1995 sandpil starts: 1998 ends: 2000 growth starts: 1991 ends: 1997 network starts: 2002 ends: renorm starts: 1994 ends: 1999 index starts: 1955 ends: 1980 inform starts: 1957 ends: 1988 citat starts: 1955 ends: 1990 model starts: 1991 ends: 1996 	•	ISI Data: D:\Users\mginda\Desktop\sci2-IVMOOC\sampledata\scientometri 361 Unique ISI Records.2 with normalized Abstract, Title Bu Save View View View View With Rename Discard
Input Parameters: View file as: csv	III T	
Remove From List Remove completed automatica	ally Remove all comple	Workflow Manager Workflow 1 Load Lowercase Tokenize
Application Viewer Type	urst detection analy	Right-click on Burst Burst Detection Parameters View with Parameters View With Parameters Parameters View With Parameters Par
Please choose an application viewer to re w file as Firefox HTML Document Firefox HTML Document Microsoft Word 97 - 2003 Document Text Document Microsoft Excel Comma Separated Val	ad this file.	maximum burst level 1 in t data manager and view the file in the spreadsheet program of your choice

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

Temporal Analysis: Burst Detection

Pas Clipt	te 🛷 E	B I ∐ → 	A A A	E E = E ₽ ₩ Alianment	₩ × \$ × × 500 ÷	% , S	tyles		Pa	ste Joboard 5
	F6	•	f,	2						G2
	1									-
	Α	В	С	D	E	F	G	_		A
1	Word	Level	Weight	Length	Start	End			1	Word
2	transform	1	4.447253	8	1988	1995	i		2	transf
3	analysi	1	5.170601	22	1972	1993	•		3	analys
4	scienc	1	6.959986	31	1955	1985	i		4	scienc
5	critic	1	5.934392	6	1993	1998			5	critic
6	complex	1	8.823672	8	2000		<u> </u>		6	comp
7	chemic	1	4.026779	22	1957	1978	s		7	chemi
8	self	1	5.584457	10	1990	1999			8	self
9	fractal	1	4.654846	8	1990	1997	·		9	fracta
10	transit	1	4.855824	4	1997	2000)		10	transi
11	protein	1	4.017034	5	2003				11	protei
12	aggreg	1	4.791958	6	1990	1995	i		12	aggreg
13	sandpil	1	4.789993	3	1998	2000)		13	sandp
14	growth	1	5.953631	7	1991	1997	1		14	growt
15	network	1	21.68094	6	2002				15	netwo
16	renorm	1	4.301231	6	1994	1999)		16	renor
17	index	1	9.816979	26	1955	1980)		. 17	index
18	inform	1	4.867829	32	1957	1988	•		18	inform
19	citat	1	6.541599	36	1955	1990)		19	citat
20	model	1	9.214359	6	1991	1996	i		20	mode
21									21	

Cyberinfrastructure for Network Science Center

Pas	💼 1 ste 	B <u>I U</u> →	A A A	≣≣⊒ ≣∉i⊗≫	· * *.0 :	% ,	yles	Dele Dele	ete ≁ mat *	
Clip	board 🗔	Font	G.	Alignment	G Num	ber 🖫		Cell	s	
	G20	• (fa	-						
	А	В	С	D	E	F	G		H	
1	Word	Level	Weight	Length	Start	End				
2	transform	1	4.447253	8	1988	1995				
3	analysi	1	5.170601	22	1972	1993				
4	scienc	1	6.959986	31	1955	1985				
5	critic	1	5.934392	6	1993	1998				
6	complex	1	8.823672	8	2000	2014				
7	chemic	1	4.026779	22	1957	1978				
8	self	1	5.584457	10	1990	1999				
9	fractal	1	4.654846	8	1990	1997				
10	transit	1	4.855824	4	1997	2000				
11	protein	1	4.017034	5	2003	2014				
12	aggreg	1	4.791958	6	1990	1995				
13	sandpil	1	4.789993	3	1998	2000				
14	growth	1	5.953631	7	1991	1997				
15	network	1	21.68094	6	2002	2014				
16	renorm	1	4.301231	6	1994	1999				
17	index	1	9.816979	26	1955	1980				
18	inform	1	4.867829	32	1957	1988				
19	citat	1	6.541599	36	1955	1990				
20	model	1	9.214359	6	1991	1996				
21										

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

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

/isualization R	Help	_
General	•	ata Manager
Temporal	•	Temporal Bar Graph op\sci2-IVMOOC\sample
Geospatial	•	361 Unique ISI Records
Topical	•	S Extracted Co-Authorship Network
Networks	•	Author information

Temporal E	Bar Graph	
	Takes tabular data and generates PostScript for a temporal bar graph.	
Subtitle	Generated from 361 Unique ISI Records	
Label	Abstract	-
Start Date	Abstract	•
End Date	Abstract	- 0
Size By	Cited Reference Count	•
Date Format	Month-Day-Year Date Format (U.S., e.g. 10/15/2010)	•
Category	No Category Coloring	- 0
Scale Out	put?	0
Simplified	I Layout?	۲
	ОК	Cancel

Cyberinfrastructure for Network Science Center Temporal Analysis: Burst Detection

😵 Sci2 T	ool			
File Da	ta Preparation Preprocessing Analysis Modeling Visualizat	ion R Help		
📮 Con:	Sci2 Tool	_		
networ	File Data Preparation Preprocessing Analysis Modeling	Visualization R He	P	
renorm	E Console	General	la Manager 🔤 🗖	
inform	renorm starts: 1994 ends: 1999	Temporal	Temporal Bar Graph p\sci2-IVMOOC\sampledata\scientometri	
citat sta	index starts: 1955 ends: 1980	Geospatial	1 361 Unique ISI Records.2	
model	inform starts: 1957 ends: 1988	Topical	🔹 🖬 with normalized Abstract, Title	
View w	model starts: 1991 ends: 1996	Networks	Burst detection analysis (Publication Year, Title): maximum burst le	
			CSV file: D:\Users\mginda\AppData\Local\Temp\temp\terprocessed-Burst	
Input P View fil	View with was selected.			
	Input Parameters:	T T T T T T T T T T T T T T T T T T T	In the Cost	X
Docum	View file as: csv	1empo	ai bar Graph	-
	Load was selected.		Takes tabular data and generates PostScript for a	
📮 Sche	Documentation: http://wiki.cns.iu.edu/display/CISHELL/Da	ata + Forma	temporal bar graph.	
Remo	Loaded: D:\Users\mginda\Desktop\BurstDetectionResults.csv	Subtitle	Generated from CSV file: D:\Users\mginda\AppData\Local\Temp\temp\Preprocessed-BurstDetectionResults-2120606791465131481.csv	۲
	Scheduler	Label	Word	- 📀
	Remove From List Remove completed automatically	Remove all Start Date	Start	- 0
		End Date	End	▼ 🅙
Û	! Algorithm Name Date Time	% Comp Size By	Weight	- 📀
Ŧ	✓ Load 04/03/2015 11:59:11			
	✓ View With 04/05/2015 11:49:55 ✓ Rurst Detection 04/03/2015 11:47:56	Date For	Month-Day-Year Date Format (U.S., e.g. 10/15/2010)	▼ 🅙
	V Lowercase Tokeniz 04/03/2015 11:38:53	Category	No Category Coloring	- 📀
	✓ Load Ø4/03/2015 11:36:49			
		Scale	Jutput?	*
		Simpli	fied Layout?	۲
			OK	Cancel

- 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

Cyberinfrastructure for 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

CSV file: D:\Us	ers\mginda\AppData\Local\Temp\temp\Preprocesse Save View View With Rename Discard	d- PostScript Raster Image	Details label: file:text/ps -> file-ext:ps out_data: file:ext:ps in_data: file:text/ps conversion: lossless
r		S	elect Cancel Details >

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 <u>*PS2PDF.com*</u>.

— X

Sci2: Converting the Output PostScript Bursts to PDF



```
Command Prompt
```

Cyberinfrastructure for Network Science Center

0:5.

Desktop app

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.

Comm	and Prompt				-	×	
osof [.] 2015	t Windows [Ver Microsoft Cor	rsion 10.0.1 rporation. A	10240] All rights re	eserved.		^	
sers	\mginda>						
	(ing 2110 day						
	Command	Promot					 ×
	Command	a Prompt				_	
	Microsoft W	lindows [Ver	rsion 10.0.1	0240]			^
	(C) 2015 Mi	icrosoft Con	rporation. A	II rights reserved.			
	C:\Users\me	vinda≻DTR					
	Volume in	drive C has	s no label.				
	Volume Ser	rial Number	is 8222-07E	9			
	Directory	of C:\Users	s∖mginda				
	05/18/2016	11:59 AM	<dir></dir>				
	05/18/2016	11:59 AM	<dir></dir>				
	01/04/2016	03:00 PM	<dir></dir>	.continuum			
	05/24/2016	01:23 PM	<dir></dir>	.oracle_jre_usage			
	03/10/2016	01:15 PM	<dir></dir>	.sqlworkbench			
	01/04/2016	02:51 PM	<dir></dir>	3D Objects			
	04/13/2016	03:51 PM	<dir></dir>	Contacts			
	06/11/2016	02:06 PM	<dir></dir>	Desktop			
	06/09/2016	06:45 PM	<dir></dir>	Documents			
	03/30/2016	01:48 PM	<dir></dir>	DocumentsADNI			
	06/06/2016	04:42 PM	<dir></dir>	Downloads			
	04/13/2016	03:51 PM	<dir></dir>	Favorites			
	04/13/2016	03:51 PM	<dir></dir>	Links			
	01/14/2016	01:02 PM	<dir></dir>	logs			
	01/04/2016	02:49 PM	<dir></dir>	Mallet			
	04/13/2016	03:51 PM	<dir></dir>	Music			
	05/18/2016	11:59 AM	<dir></dir>	OneDrive			
	01/26/2016	05:11 PM	<dir></dir>	Pajek			
	04/13/2016	03:51 PM	<dir></dir>	Pictures			
	06/10/2016	09:37 AM	<dir></dir>	Refine			
	01/13/2016	03.51 DM	ZDTRN	Saved Games			

Cyberinfrastructure for 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
```

Select Com	nmand Prompt			-	\times
	27 Dir(s) 134,791,598,08	30 bytes free		
:\Users\mg	inda≻cd C:\	Users\mginda\Docu	uments		
·\Users\mg	inda\Docume	nts>DTR			
Volume in	drive C has	no label.			
Volume Ser	ial Number	is 8222-07E9			
- •	5 - 1				
Directory	ot C:\Users	\mginda\Documents			
6/09/2016	06:45 PM	<pre><pre>DTR></pre></pre>			
6/09/2016	06:45 PM	<dir></dir>			
6/06/2016	04:16 PM	<dir></dir>	Admin		
5/17/2016	09:38 AM	<dir></dir>	ADNI		
6/06/2016	10:07 AM	13,409	ADNI-JournalImpactFactors.xlsx		
2/10/2016	05:40 PM	23,568	ADNI-JournalList-countsstats.xlsx		
2/10/2016	06:22 PM	81,738	ADNI-JournalsStats Query.xlsx		
1/08/2016	03:16 PM	<dir></dir>	Adobe		
1/04/2016	02:32 PM	<dir></dir>	BBSRC		
6/09/2016	06:45 PM	13,926	Burst Detection.docx		
4/28/2016	12:16 PM	<dir></dir>	Conference		
2/07/2016	05:26 PM	3,318,937	Copy of icd102010enMeta-csv.txt		
4/13/2016	11:39 AM	<dir></dir>	Custom Office Templates		
1/04/2016	02:36 PM	<dir></dir>	DocumentationPrep		
5/16/2016	10:29 AM	<dir></dir>	Endnote		
5/23/2016	12:14 PM	<dir></dir>	IAI		
5/25/2016	11:44 AM	<dir></dir>	IVMOOC		
6/07/2016	10:06 AM	<dir></dir>	LA		
3/21/2016	09:27 AM	95,721	LIGHTR-ESRICONFERENCEnotes.rtf		
3/08/2016	12:24 PM	360	modelLCSH.txt		

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

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

Cyberinfrastructure for Network Science Center

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.



Sci2: Converting the Output PostScript Bursts to PDF

You can view the PDF file with Adobe Reader. More information on this script tool is available in the <u>PS2PDF documentation site</u>.

Temporal Visualization

(Generated from CSV file: D:\Users\mginda\AppData\Loca\\Temp\temp\Preprocessed-BurstDetectionResults-4411812064438418826.csv) April 03, 2015 | 12:05 PM EDT



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

Legend



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()()

CNS (cns.iu.edu)



Cuberinfrastructure for

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.

INS Cyberinfrastructure for Network Science Center Burst Detection: CDC Grants Dataset

Sý So	ci2 To	ol												
File	Dat	a Preparation Preproc	essing A	nalysis Modelin	g Visualization R Help			🌍 Select Files						Х
	Load			Ctrl+Alt+O	- 8	1010	Data I	← → • ↑ 📘	> Thi	is PC > Desktop > Sci2-CDC-data > Burst	ٽ ~	Search Burst	م ر	
	Reac Face	Directory Hierarchy		•	Streams. Proc. 8th ACM SIGKDD			Organize 🔻 New	/ folde	r		:== :==	- 💷 🔮	
	Goog	gle Scholar		۶.	urst+Detection			🖈 Quick access	^	Name	Date modified	Type	Size	
	Twitt	er Reader						Desktop 🦻		CDC-Burst-FR-Envirmeaith-Grant Lities.csv	6/8/2016 3:21 PM	Microsoft Excel C	13 722 KB	
	Cours			Ctrl - Alt - C				Downloads		DCDC-Grants-20102015-EnviromentalHeal	6/8/2016 1:19 PM	Microsoft Excel C	502 KB	Ľ
-	Save			Ctri+Ait+S				mainda		CDC-Grants-20102015-Infectious.csv	6/8/2016 1:26 PM	Microsoft Excel C	3,455 KB	
	View			Ctrl+Alt+V	emporal+Bar+Graph			inginua 🤇		DDC-Grants-20102015-InjuryPrevention.c	6/8/2016 1:22 PM	Microsoft Excel C	655 KB	
-	view	with		Ctri+Ait+W				i OneDrive						
	Merg	e Node and Edge Files			-			💻 This PC						
-	Split	Graph to Node and Ed	ge Files		emporal+Bar+Graph			📃 Desktop						
	Prefe	rences						Documents						
	Conv	erter Graph			- 8	0	Workf	🖊 Downloads						
	Exit				Remove all completed		≣w	b Music						
	1							Pictures						
	J							📕 Videos	~					-
	1	Algorithm Name	Date	Time	% Complete				File na	me: CDC-Grants-20102015-EnviromentalHealth.cs	sv ~	*.*	~	
												Open	Cancel	
														.:
Ú									4			×		
Q	7											~		
									l he file Please si	'C:\Users\mginda\Desktop\Sci2-CDC-data\Burst\CDC-Gra elect the format you would like to try.	ants-20102015-Enviromen	talHealth.csv' can be		
								- Load a	as					
								Standa NSF cs	rd csv f v forma	ormat it				
								Scopus	s csv for	mat				
											Select Ca	ncel Details >>		

Load CDC-Grants-20102015-EnviromentalHealth.csv

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

Cyberinfrastructure for Burst Detection: CDC Grants Dataset

Sci2 T								-		
File Data	a Preparation	Preprocessing A	nalvsis N	Aodeling Visualizati	ion R Help					
Conso	ole	General	>	,, ,			1010 Data M			
		Temporal	>			^	CS	SV file: C:\Users\mginda\AppData\Local\Temp\temp\Preproces:		
The devel	opment of thi	Geospatial	> /t	he Cyberinfrastructur	e for Network Science c	enter and		· · · · · · · · · · · · · · · · · · ·		
the Depar University	tment of Infor , the National	Topical	>	Reconciled Journa	il Names					
McDonne	Il Foundation.	Networks	>	Lowercase, Tokeni	ize, Stem, and Stopword	Text				
The prima Sci2 tool v Chin Hua Patrick A. developed Many algo Please cito Sci2 Team https://sc	ary investigators was developed I Kong, Steven C Phillips, Chinta d at the Cyberin orithm plugins e as follows: n. (2009). Scienc ci2.cns.iu.edu.	s are Katy Börner, I by Daniel Halsey, A corenflos, Joseph B ın Tank, and Russe frastructure for Ne were derived from e of Science (Sci2)	ndiana Uni Adam Simp liberstine, T II J. Duhon etwork Scie the Netwo Tool. India	iversity and Kevin W. J oson, Saumya Pandey, Thomas G. Smith, Dav. I tuses the Cyberinfr noce Center (http://cn ork Workbench Tool (J ana University and Sci	Boyack, SciTech Strategi , Sumit Samant, Vivek K. vid M. Coe, Micah W. Lir astructure Shell (http:// is.iu.edu) at Indiana Unin http://nwb.cns.iu.edu). Tech Strategies,	es Inc. The rrihaloo, nemeier, ishell.org) ersity.		Lowercase, Tokenize, Stem, and Stopword Text	>	<
Load wa	as selected.							Stopword List C:/Users/mginda/Desktop/Sci2/sci2-1.2b_20151218-win32.win32.x86/sci2/configuration/stopwords.txt Browse	0	
Documen	tation: http://w	viki.cns.iu.edu/dis	play/CISH	ELL/Data+Formats	2015-EnviromentalHeal	h cou		New Connector 1		
Loaded. C		a/Desktop/Sci2-CL	/C-uata (Di	dist(CDC-Glants-2010	2015-Environnentan lear	×	<			
Sched	luler						O Workfle	Project Number	0	
Remove	From List	Remove complete	d automat	ically Remove all c	ompleted 🕞		∎ wa	Reference Number	0	
Ш								FOA Number	0	
	! Algorithm	Name [Date	Time	% Complete			FOA Title	0	
-	✓ Load	(06/08/2016	04:50:26 PM				Grantee Project Title	0	
								Category	0	
С Л								Sub-Category	0	1
								Award Class	0	
								Grantee Name	0	
								Primary Address	0	
								City	0.	4
								ОК	Cance	el

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

ing	Analysis Modeling	Vis	ualization R Help			
	Temporal	>	Burst Detection		1010 Data Manager	- E
	Geospatial	>		^	CSV file: C:\Users\mgin	da\AppData\Local\Temp\temp\Preproce
	Topical	>			with normalized Gra	antee Project Title
	Networks	>				
-1.2k	_20151218-win32.win	32.x86	/sci2/configuration/stopwords.t	txt		Burst Detection

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 Perform Burst Detection on time-series textual data.

Gamma	1.0	0
Density Scaling	2.0	0
Bursting States	1	0
Date Column	Year ~	0
Date Format	уууу	0
Burst Length Unit	Years ~	0
Burst Length	1	0
Text Column	Grantee Project Title \sim	0
Text Separator		0
Document Column	Project Number V	0
🗹 Ignore Input With	n Empty Text	0
	OK	Cancel

×

CNS Cyberinfrastructure for Network Science Center Burst Detection: CDC Grants Dataset

	👘 🖓 Data Manager
brace starts: 2013 ends: track starts: 2014 ends: adult starts: 2015 ends: cba starts: 2015 ends: reduc starts: 2011 ends: 2011 toxic starts: 2011 ends: 2011 relat starts: 2010 ends: 2010 evid starts: 2010 ends: 2010 brfs starts: 2010 ends: 2010 brfs starts: 2012 ends: 2011 substanc starts: 2011 ends: 2011 survey starts: 2015 ends: program starts: 2011 ends: 2011 surveil starts: 2010 ends: 2010 address starts: 2010 ends: 2010 address starts: 2013 ends: 2010 address starts: 2014 ends: 	Our Data Vianager CSV file: C:\Users\mginda\AppData\Local\Temp\temp\Preproces: With normalized Grantee Project Title Burst detection analysis (Year, Grantee Project Title): maxim
Image: Indice on price on price on optical conduction in the image: Indice on optical conduction in the image:	 Lowercase Tokenize Stem and Stonword Text Param Bursti Param Bursti Param Bursti Comparison Right-click on Burst detection analysis (Year, Grantee Project Title): maximum burst level 1 in data manager and view th file in the spreadsheet

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW ACF	OBAT Ginda, Mi
	- P
Get External Data → Refresh All → All → All → Filter Filter Filter Filter Out	line
Connections Sort & Filter Data loois	Analysis
G11 \cdot : $\times \checkmark f_x$	
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

ing	Analysis Modeling	Vis	ualization R Help				
	Temporal	>	Burst Detection		1010 Data Manager	- E	
	Geospatial			~	CSV file: C:\Users\mgin	s\mqinda\AppData\Local\Temp\temp\Preproce	
	Topical	>			with normalized Gra	antee Project Title	
	Networks	>					
-1.2k	-1.2b_20151218-win32.win32.x86/sci2/configuration/stopwords.tx			bit		Burst Detection	

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"

Perform Burst Detection on time-series textual data.

Gamma	1.0	@						
Density Scaling	2.0	0						
Bursting States	1	0						
Date Column	Year ~	0						
Date Format	ууууу	0						
Burst Length Unit	Years ~	0						
Burst Length	1	0						
Text Column	Grantee Project Title \sim	0						
Text Separator	1	0						
Document Column	Project Number 🛛 🗸	0						
Ignore Input With Empty Text								
OK Cancel								

×

Burst Detection: CDC Grants Dataset

Temporal > Burst Detection	🖞 Data Manager			
Geospatial > Topical > Networks >	CSV file: C:\Users\mginda	a\AppData\Local\Temp\te tee Project Title	mp\Preproce	
-1.2b_20151218-win32.win32.x86/sci2/configuration/stopwords.txt	Burst Detection			
		Perform Burst De	etection on time-series textual d	ata.
Highlight the table 'with norma	Gamma	2.0		
Grantee Project Title'	Density Scaling	1.75		
Select Analysis > Temporal > B	Bursting States	1		
Detection		Date Column	Year 🗸	
For the final burst analysis use	Date Format	ууууу		
	- //	Burst Length Unit	Years ~	
For Density Scaling enter "1.75	5″	Burst Length	1	
For the Date Dolumn select "Y	Text Column	Grantee Project Title ~		
For Text Column select "Grant Project Title"	Text Separator			
For Document Column select		Document Column	Project Number 🛛 🗸	
"Project Number"		🗹 Ignore Input With	h Empty Text	
			OK	Can

NS

Cyberinfrastructure for Network Science Center

 \times

0

0

0

0

0

0

0

0

0

0

0

Cancel
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.

Temporal Analysis: Burst Detection

	Α	В	С	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	
26						

Cyberinfrastructure for Network Science Center

	Α	В	С	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

Cyberinfrastructure for Network Science Center Burst Detection: CDC Grants Dataset

😵 Sci2 Tool		- 🗆 X	
File Data Preparation Preprocessing Analysis Modeling Visualization R Help			
Sci2 Tool		–	
File Data Preparation Preprocessing Analy	rsis Modeling Visu	alization R Help	
healthi starts: 2011 ends: 2011		General > Control Cont	
Childhoodo starts: 2010 ends: 2010 risk starts: 2012 ends: 2013 drink starts: 2012 ends: 2013 drink starts: 2019 ends: 2013 sthma starts: 2019 ends: 2013 factor starts: 2012 ends: 2013 drink starts: 2015 ends: ends: 2013 drink starts: 2015 ends: ends: 2014 drink starts: 2014 drink starts: 2015 ends: ends: 2014 drink starts: 2015 ends: ends: 2014 drink starts: 2015 ends: ends: 2014 drink starts: 2015 drink starts: 2014 drink starts: 2015 drink starts: 2014 drink starts: 2014 drink starts: 2014 drink starts: 2014 drink starts:		Temporal Temporal Bar Graph Geospatial > Topical > Detworks > Temporal Bar Graph File: C:\Users\mginda\AppData\Local.Temp\temp\Preproces: Burst detection analysis (Year, Grantee Project Title): maxim Burst detection analysis (Year, Grantee Project Title): maxim File: C:\Users\mginda\Sigma (Participation): Comparison (Partipation): Comparison (Participation): Comparison (Partipati	
poison starts: 2011 ends: 2011 factor starts: 2012 ends: 2013		CSV file: C:\Users\mginda\AppData\Loca\Temp\temp\Preproces:	
annound starts: 2011 ends: 2011 sol1-1101 starts: 2013 ends:			
brfss starts: 2012 ends: 2013 announc starts: 2011 ends: 2011	Temporal	l Bar Graph	×
substanc starts: 2011 ends: 2011 prevent starts: 2011 ends: 2011			
surveil starts: 2012 ends: 2012 substanc starts: 2011 ends: 2011		Takes tabular data and generates PostScript for a temporal bar	
address starts: 2013 ends: 2013 program starts: 2011 ends: 2011		graph.	
Loaded: C:\Users\mginda\Desktop\Sci2-Cl address starts: 2013 ends: 2012			
strategi starts: 2014 ends:	Subtitle	Generated from CSV file: C:\Users\mginda\AppData\Local\Temp\temp\Preprocessed-CDC-EnvGrants-Burst-Results-F-4354567135667837397.csv	
Documentation: http://wiki.cns.iu.edu/dis			
Load was selected.			-
Scheduler Loaded: C:\Users\mginda\Desktop\Sci2-CDC-c	Label	Word	1 😻
Remove From List Remove complete			
Scheduler	Start Date	Start	· 🙆
Remove From List Remove completed as			
! Algorithm Name D			
Load Of	End Date	End	1 😵
Burst Detection 06 L Algorithm Name Date			
View with 0t Algorithm with 06	Size By	Weight	. 💿
View with 06	5.20 59		
View with 06/0			
Under View with 06/0	Date Format	Month-Day-Year Date Format (U.S., e.g. 10/15/2010)	1 😧 🗌
Burst Detection 06/0	1		
Lowercase, Tokenize, St 06/0	Category	No Category Coloring	
Euga	category	No category coloring	
	Scale Out	put?	0
	Simplified	Layout?	•
		OK	Cancel

- 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

NC

Sci2: Converting the Output PostScript Bursts to PDF



```
Command Prompt
```

Cyberinfrastructure for Network Science Center

¢:%.

Desktop app

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.

Comm	and Prompt				_	×	
osoft 2015	: Windows [Ver Microsoft Cor	rsion 10.0.1 poration. A	10240] All rights re	eserved.		^	
lsers	mginda≻						
	Command	l Prompt				_	×
	Microsoft W (c) 2015 Mi	lindows [Ver .crosoft Cor	rsion 10.0.1 rporation. A	0240] ll rights reserved.			^
	C:\Users\mg Volume in Volume Ser Directory	inda>DIR drive C has ial Number of C:\Users	s no label. is 8222-07E s\mginda	a] tights reserved. - C × n 10.0.10240] ation. All rights reserved. label. 8222-07E9 inda OIR> . OIR> . OIR> OCUMENTS OIR> 			
	05/18/2016 05/18/2016 05/18/2016 03/10/2016 03/10/2016 06/11/2016 06/11/2016 06/09/2016 06/06/2016 06/06/2016 04/13/2016 01/14/2016 01/14/2016 04/13/2016 05/18/2016 04/13/2016 04/13/2016	11:59 AM 11:59 AM 03:00 PM 01:23 PM 02:51 PM 02:51 PM 02:51 PM 02:66 PM 06:45 PM 04:42 PM 03:51 PM 03:51 PM 03:51 PM 03:51 PM 03:51 PM 03:51 PM 03:51 PM 03:51 PM 03:51 PM	(01R) (01R)	 .continuum .oracle_jre_usage .sqlworkbench 3D Objects Contacts Desktop Documents DocumentsADNI Downloads Favorites Links logs Mallet Music OneDrive Pajek Pictures Refine			
	94/13/2016	03.51 PM	(DTR)	Saved Games			

Cyberinfrastructure for 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
```

Select Con	nmand Prompt			_	×
	27 Dir(s)) 134,791,598,	080 bytes free		^
C:\Users\mg	inda≻cd C:\l	Jsers\mginda\Do	cuments		
C:\Users\mg Volume in Volume Ser	inda\Documer drive C has ial Number i	nts>DIR no label. is 8222-07E9			
Directory	of C:\Users	\mginda\Documen	ts		
26/09/2016	06:45 PM	<dtr></dtr>			
26/09/2016	06:45 PM	<dir></dir>			
96/06/2016	04:16 PM	<dir></dir>	Admin		
05/17/2016	09:38 AM	<dir></dir>	ADNI		
06/06/2016	10:07 AM	13,40	<pre>ADNI-JournalImpactFactors.xlsx</pre>		
02/10/2016	05:40 PM	23,56	3 ADNI-JournalList-countsstats.xlsx		
02/10/2016	06:22 PM	81,73	3 ADNI-JournalsStats Query.xlsx		
01/08/2016	03:16 PM	<dir></dir>	Adobe		
91/04/2016	02:32 PM	<dir></dir>	BBSRC		
96/09/2016	06:45 PM	13,92	5 Burst Detection.docx		
94/28/2016	12:16 PM	<dir></dir>	Conference		
02/07/2016	05:26 PM	3,318,93	7 Copy of icd102010enMeta-csv.txt		
04/13/2016	11:39 AM	<dir></dir>	Custom Office Templates		
01/04/2016	02:36 PM	<dir></dir>	DocumentationPrep		
05/16/2016	10:29 AM	<dir></dir>	Endnote		
05/23/2016	12:14 PM	<dir></dir>	IAI		
05/25/2016	11:44 AM	<dir></dir>	IVMOOC		
36/07/2016	10:06 AM	<dir></dir>			
03/21/2016	09:27 AM	95,72	L LIGHIR-ESRICONFERENCENOTES.rtt		
03/08/2016	12:24 PM	36	ð modellCSH.tXt		\sim

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
```

Cyberinfrastructure for Network Science Center

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.



Sci2: Converting the Output PostScript Bursts to PDF

You can view the PDF file with Adobe Reader. More information on this script tool is available in the <u>PS2PDF documentation site</u>.

Temporal Visualization

Legend

(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



CNS (cns.iu.edu)

0.52 Year(s)



Questions?



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

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

9:15 Tool Overview and Trouble Shooting

Cyberinfrastructure for

- Sci2 Overview scientometric analysis tool
- Open Refine data parsing, transformation, and editing tool

Tutorial Overview

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

CNS Cyberinfrastructure for Network Science Center Introduction to Network Analysis

What is a Network?

- Graph network visualized
- Nodes
- Edges
- Components

Representations

- Matrices
- Graphs
- Edge and Node Lists

Data Formats

- Tabular
- XML
- Text
- JSON



Cyberinfrastructure for Network Science Center 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

Other types of networks and graphs:

- Hierarchical networks (tree networks)
- o Bipartite Networks

Directed Networks

Ed Dir rep In Nu Ou Nu ed

Edge Direction: Directional relationship is represented by arrows

In-Degree: Number of incoming edges

Out-Degree: Number of outgoing edges

- o Multigraphs
- o Hypergraphs

Cyberinfrastructure for Network Science Center Introduction to Network Analysis

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



School Constructure for Introduction to Network Analysis

Node Metrics

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



Cyberinfrastructure for Introduction to Network Analysis

Graph Metrics - Edges

 Shortest paths – shortest distance between two nodes

- Weight strength of tie
- Directionality is the connection one-way or twoway (in-degree vs. out-degree)?
- Bridge deleting would change structure





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

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

9:15 Tool Overview and Trouble Shooting

Cyberinfrastructure for

- Sci2 Overview scientometric analysis tool
- Open Refine data parsing, transformation, and editing tool

Tutorial Overview

Gephi – network visualization

9:45 Geospatial Analysis

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

11:00 Topical/Temporal Analysis: Burst Detection

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

12:30 Lunch

1:30 Network Analysis

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

4:00 Wrap-up

4:30 Adjourn

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

Cuberinfrastructure for



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

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

CNS Cyberinfrastructure for Network Science Center Co-Author Network Analysis

Sci2	2 Tool	64					
File	Data Preparation Prepro	cessing Analy	/sis Modelir	ng Visualization R I	lelp		
EC	Convert to Generic Pu	Iblication			ita Manager	- 8	
Doc	Remove ISI Duplicate	Records			ISI Data: D:\Users\mginda\Deskto	p\sci2-IVMOOC\sampledata\scientometri	Load FourNetSciResearchers.isi
Four	2 Tool Data Preparation Preprocessing Analysis Modeling Visualiza Convert to Generic Publication Remove ISI Duplicate Records Remove Rows with Multitudinous Fields Extract Directed Network Extract Bipartite Network Extract Paper Citation Network Extract Co-Occurrence Network Extract Co-Occurrence Network Extract Co-Author Network Extract Reference Co-Occurrence (Bibliographic Coupling) Ne Extract Document Co-Citation Network Detect Duplicate Nodes Update Network by Merging Nodes ove From List ■ Remove completed automatically Remove all				361 Unique ISI Records.3		located in Sci2 Directory
Four	Extract Directed Netv	vork					
Four	Extract Bipartite Netv	vork Network					
ISI II	Extract Author Paper	Network					Select the data 361 Unique ISI
Miro	Extract Co-Occurrence	e Network					Records in the Data manager and the
D:\U	Extract Word Co-Occ	urrence Netwo	rk				
8715	Extract Co-Author Ne	twork					algorithm Extract Co-Author Network
Loac D:\U	Extract Reference Co-	Occurrence (B	ibliographic	Coupling) Network			in the Data Preparation menu.
urNe	Extract Document Co	-Citation Netw	ork				I
	Detect Duplicate Noo	des				4	
📮 S	Update Network by N	lerging Nodes			orkflow Manager		A pop-up window will appear; select
Rem	ove From List 🔲 Remov	e completed a	utomatically	Remove all comple	Workflow 1		the format ISI.
					😍 Load		
		D :		01 C 1 1			
	Algorithm Name	Date 04/06/2015	11me	% Complete			
		04/00/2013	09.09.40				
6							
						E Extract Co. Author N	
						Extract CO-Author No	etwork
						Extracts a co-authorship	network from one of several
						supported file types	
						supported me types.	
						File Format isi	▼ 00

CNS Cyberinfrastructure for Network Science Center Co-Author Network Analysis

Sci2	2 Tool	_						After creating your initial network, it				
File [Data Preparation Preproc	essing Analy	sis Modelin	g Visualization I	R Help			is a good idea to got a brief every				
©056 87150 Loade	2 Tool Data Preparation Preprocessing Analysis Modeling Visualizatio Disole Dista Preparation Preprocessing Analysis Modeling Visualizatio Disole Disole Dista Preprocessing Analysis Modeling Visualizatio Disole Dista Preparation Preprocessing Analysis Modeling Visualization Dista Preparation Preprocessing Analysis Modeling Visualization Dista Preparation Preprocessing Analysis Modeling Visualization Dista Preparation Preprocessing P)))))))))))))))))))	III Data Manager III ISI Data: D:\Users\mg III 361 Unique ISI Rec	ginda\Desktor	D\sci2-IVMOOC\sampledata\scientometri	of its statistical properties.						
D:\Use urNet! Extract Imple Integr Docur http:// %28T Input I File Fc	ers\mginda\Desktop\sci2-1 SciResearchers.isi tt Co-Author Network was : menter(s): Timothy Kelley rator(s): Timothy Kelley mentation: //wiki.cns.iu.edu/display/ /ext + Files%29 Parameters: ormat: isi neduler ove From List Remove	vMoo(t	ract + Co-Aut	Network Unweigh Weighte Unweigh Weighte Unweigh Weighte Unweigh Weighte The second seco	Analysis Toolkit (NAT) ted & Undirected d & Undirected ted & Directed d & Directed	III	work	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:				
4	! Algorithm Name ☑ Extract Co-Author ☑ Load	Date 04/06/2015 04/06/2015	Time 09:10:17 09:09:48	% Complete	© Parameters		This graph claims to be un Nodes: 247 Isolated nodes: 0 Node attributes present: Edges: 891 No self loops were discove No parallel edges were dis	directed. label, number_of_authored_works, times_cited ered. scovered.				
							Edge attributes: Did not detect any Numeric attributes number weight This network seems Average degree: 7.2146 This graph is not weakly on There are 3 weakly connect The largest connected comp Did not calculate strong of Density (disregarding weig Additional Densities by Nu	y nonnumeric attributes. min max mean 1 33 1.76094 1 33 1.76094 ; to be valued. connected. red components. (0 isolates) ponent consists of 194 nodes. connectedness because this graph was not directed. ghts): 0.0293 imeric Attribute				

~			_					×)	
🎐 Sci	2 Too							×	
File	Data	Preparation Preproce	essing Analy	sis Modeling	g Visualization	R	Help		One of th
E C	(Convert to Generic Pub	lication				ita Manager		author ne
8715		Remove ISI Duplicate F	Records				ISI Data: D:\Users\mginda\Desktop\sci2-IVMOOC\sampledata\scientor	netri	vour data
Loac		Remove Rows with Mu	Ititudinous Fie	elds			361 Unique ISI Records.3		your data
urNe	I	Extract Directed Netwo	ork				Author information 2		(e.g. Johr
	I	Extract Bipartite Netwo	ork				Author Mormation.2		(0.8.00)
Extra		Extract Paper Citation N	Vetwork						
Integ	1	Extract Author Paper N	etwork						To detect
Doc	I	Extract Co-Occurrence	Network						
http	I	Extract Word Co-Occur	rence Networ	k					want to s
%28	1	Extract Co-Author Netv	vork						data mar
Inpu	1	Extract Reference Co-C	occurrence (Bil	bliographic C	oupling) Netwoi	ſK			
File	I	Extract Document Co-0	Citation Netwo	ork					Data Pre
_	[Detect Duplicate Node	s				III	4	Dunlicate
🖳 S	l	Update Network by Me	erging Nodes				prkflow Manager		Dupilcult
Ren	ove	From List 🔲 Remove	completed au	Itomatically	Remove all com	nple	Workflow 1		
m							• Load		A non-un
							Extract Co-Author Network Deremeters		
	1	Algorithm Name	Date	Time	% Complete		V Palameters		this demo
		Extract Co-Author	04/06/2015	09:10:17					naramete
_		Load	04/06/2015	09:09:48					paramete
Û									
Ţ									
							Detect Duplicate Nodes		
							Detects duplicate no	odes by co	omparing their
Sol2 Tool The Data Preparation Preprocessing Analysis Modeling Visualization R Help Convert to Genetic Publication Remove Rows with Multitudinous Fields Convert to Genetic Publicate Network Extract Paper Clation Network Extract Paper Clation Network Extract Paper Clation Network Extract Paper Clation Network Extract Co-Author Network Extract Co-Author Network Extract Co-Author Network Remove from List Remove completed automatically Remove all complet Detect Duplicate Nodes Pletect Duplicate Nodes Pletect Co-Author Network Convert Co-Currence Network Extract Co-Author Network Extract Co-Author Network Extract Co-Author Network Convert Co-Currence Network Extract Co-Author Network Convert Co-Currence Network Extract Co-Author Network Convert Co-Currence Network Convert Co-									
								[] a la a l	
							Attribute to compare on	label	
								0.05	
							Merge when this similar	0.95	
								0.05	
							Create notice when this simila	ar 0.85	
								-	
							Number of shared first letters	2	

Cyberinfrastructure for Network Science Center

> One of the challenges of a coauthor 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.

X

- 0

0

0

0

Cancel

OK



Cyberinfrastructure for Network Science Center

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.

_ D X



Cyberinfrastructure for Network Science Center

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

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

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

Select open, and then OK.

Cancel



Cyberinfrastructure for Network Science Center

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

What changes do you notice to the network statistics?

This graph claims to be undirected. This graph claims to be undirected. **Revised Network Original Network** Nodes: 247 Nodes: 238 Isolated nodes: 0 Isolated nodes: 0 Node attributes present: label, number_of_authored_works, times_cited Node attributes present: label, number_of_authored_works, times_cited Edges: 891 Edges: 872 No self loops were discovered. No self loops were discovered. No parallel edges were discovered. No parallel edges were discovered. Edge attributes: Edge attributes: Did not detect any nonnumeric attributes. Did not detect any nonnumeric attributes. Numeric attributes: Numeric attributes: min mean max min max mean 33 1.76094 number_... 1 number_... 33 1.799311 weight 33 1.76094 33 weight 1 1.77408 This network seems to be valued. This network seems to be valued. Average degree: 7.2146 Average degree: 7.3277 This graph is not weakly connected. This graph is not weakly connected. There are 3 weakly connected components. (0 isolates) There are 3 weakly connected components. (0 isolates) The largest connected component consists of 194 nodes. The largest connected component consists of 187 nodes. Did not calculate strong connectedness because this graph was not directed. Did not calculate strong connectedness because this graph was not directed. Density (disregarding weights): 0.0293 Density (disregarding weights): 0.0309 Additional Densities by Numeric Attribute Additional Densities by Numeric Attribute



Cyberinfrastructure for Network Science Center

> 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

Visualization R Help		
General >	🗖 🗖 🕅 Data Manager	
vo Temporal > Geospatial > pt Topical > In Networks >	n CSV file: C:\Users\mginda CSV file: C:\Users\mginda Extracted Co-Authors GUESS Gephi	a\AppData\Local\Temp\ hip Network rk Analysis Log
ular+Hierarchy Mimport rep	port	×
Source: Graph	ml-for-gephi8262436818410951216.graphml ort edge type set as UNDIRECTED	Issues INFO
-CircHeir.ps Graph Type:	Undirected 🗸	Auto-scale
# of Nodes:	410	Create missing nodes
# of Edges:	1215	New graph
Dynamic Graph	i: no	 Append Graph
Hierarchical Gra	aph: no	◯ Time frame
		OK Cancel

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

CNS Cyberinfrastructure for Network Science Center



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

Navigate workspaces

CNS Cyberinfrastructure for Network Science Center Overview of Gephi: Data Laboratory

	🌃 Gephi 0.8.2 - Project 0								- 0) ×
	File Workspace View Tools Win	dow Plugins Help								. .
	Overview	Data Laboratory	Preview						\mathbf{X}	1
Г	🗉 Data Table 🛛 🕷								,	EDG
	Nodes Edges @ Configuration	🔂 Add node 🕕 Add edge	📸 Search/Replace 📳 Ir	nport Spreadsheet 📳 Export table 🎇 More actions	s 🗸		Filte	er:	Nodes	~ 🤗
	Nodes	10		Laper	number of authored	times cited blon	del commu		blondel community level 1	
						- 1389 comm	unity			
Sele	ct the node or	🛛 Add node e	edges, impo	orts data (nodes and	4	48 comm	unity /	A REGEX fil	ter for nodes	and
edge	e lists data.	edges lists)	to create r	new networks in blank	4	48 comm	unity e	edges table	columns in (data
and	configuro tho	workspace	and over	rts data tables for a	2	2 17 comm	unity +	abla		
anu	configure the	workspace	s, and expo	ints data tables for a	4	48 comm	unity L	.abie.		
shee	et	network.				48 comm	unity <u>-</u> o		ommanity_o	
	A Chaves Sandra S			Chavee Sandra S		365 comm	unity_0		community_0	
	Gubareva, Larisa	n10		Gubareva, Larisa		5 303 comm	unity_0		community_0	
	Hall, Henrietta	n12		Hall, Henrietta		2 17 comm	unity 0		community_0	
	Wallis, Teresa	n13		Wallis, Teresa	4	48 comm	unity 0		community 0	
	Villanueva, Julie	n14		Villanueva, Julie	4	48 comm	unity_0		.community_0	
	Xu, Xiyan	n15		Xu, Xiyan	4	48 comm	unity_0			
	Bresee, Joseph	n16		Bresee, Joseph	10) 1712 comm	unity_0		community_0	
	Cox, Nancy	n17		Cox, Nancy	10) 1995 comm	unity_0	c	:ommunity_0	
	 Tappero, Jordan W. 	n28		Tappero, Jordan W.	3	3 58 comm	unity_3	(:ommunity_3	
	Nyenswah, Tolbert G.	n30		Nyenswah, Tolbert G.	6	5 58 comm	unity_1	(.ommunity_1	
	Montgomery, Joel M.	n31		Montgomery, Joel M.	5	5 42 comm	unity_1	c	.ommunity_1	
	Neatherlin, John	n32		Neatherlin, John	3	3 25 comm	unity_1	c	.ommunity_1	
	Singleton, James A.	n35		Singleton, James A.	10	933 comm	unity_54	c	.ommunity_45	
	Flannery, Brendan	n36		Flannery, Brendan	2	2 51 comm	unity_2	C	.ommunity_2	
	Fry, Alicia	n37		Fry, Alicia	3	3 310 comm	unity_2	c	ommunity_2	
	Pesik, Nicki	n41		Pesik, Nicki	2	2 52 comm	unity_3	c	ommunity_3	
	Brown, Clive M.	n42		Brown, Clive M.	3	3 61 comm	unity_3	_	ommunity 3	<u> </u>
	Aranas, Aaron E	n43		Aranas, Aaron E	3	3 54 comm	unity_16	🚽 Convert	s data fields	
	Conen, Nicole J.	n48		Conen, Nicole J.	Craata colur	mac fitting		from ctr	ndard to	
	w Hale, Christa	1151	1	Haltan Kelly	Cleate colui	inis nung				
I A	Adds & Merge data column				Boolean crit	eria and regex		🗄 dynamie	c (temporal	
+	oolc		Column	aditing to alc	functions 11	coful for filtorin		data fio		
	.0015		Column		functions. O	serui for filterin	ıg.		lusj	
-	A Clarke Kavin D	n65		Clarke Kevin P		33 comm	unity 4		ommunity 4	~
										^
		<u> </u>	i i	i i i	10	i i i i i i i i i i i i i i i i i i i	P	İ		
		Add Merge	Delete Clear	Copy data to Fill column Duplicate	Create a boolean column Cr	eate column with list of N	legate	Convert column		
		column columns	column ~ column ~	other column ~ with a value ~ column ~	from regex match v reg	gex matching groups ~ boolea	in values ~	to dynamic ~		~
									Workspac	ce0 ∢ ▶

Cyberinfrastructure for Network Science Center Overview of Gephi: Preview



Workspace 0 🔺 🕨

Gephi: Initial Layout of Network

Eile Worksn	8.2 - Project 0 ace View Tools Window Plugins Help						
0	verview Data Laboratory	Preview					
Partition	Ranking % Clustering	Graph 38				() v	Context %
Nodes Edges	🕥 💎 🎘 🎘	G Rectangle selection				💡 Hierarchy	Nodes: 238
Choose a	rank parameter V	k					Edges: 872
							Undirected Graph
							Statistics Filters
							Library
		10 m					Attributes
	-	1					⊕- <u> </u> Edges
				Layout ∞			Operator Topology
	ForceAtlas 2		~	we have been defined			Saved queries
	•			fifanHu s Multilevel		~	
T T	0		👂 Run	•		0	
Layout %	Threads			U		Run	
ForceAtlas :	Threads number	2		Barnes-Hut's properties			
0	Behavior Alternatives			Quadtree Max Level	10		Queries
⊟ Threads	Dissuade Hubs			Theta	1.2		
Threads nur	LinLog mode			Multi-level			
Dissuade Hu	Prevent Overlan			Minimum level size	3		
Prevent Ove	Edge Weight Influence	10		Minimum coarsening rate	0.75		
Edge Weigh		1.0		Yifan Hu's properties			
Scaling	C running Seeling	30.0		Step ratio	0.97		
Gravity	Stranger Oravity	50.0		Optimal Distance	1 0.97		
Performant Tolerance (state)	Stronger Gravity	<u>≥</u>					
Approximate	Gravity	0.08					
Approximati	Performance	0.4					
	Tolerance (speed)	0.1					
ForceAtlas	Approximate Repulsion						
	Approximation	1.2					
Presets.							
Presets							
	Force Atlan 2		0	VifenHu's Multiloval		0	
	TOTCEALIAS Z			mannu s multilevel			
				_			
	Presets Reset			Presets Reset]

Cyberinfrastructure for Network Science Center

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

470

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

Signature for Network Science Center



Gephi: Edge Color



Cyberinfrastructure for Network Science Center

> 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



Cyberinfrastructure for Network Science Center

> 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.

🍘 Graph Distance settin	gs	\times	Network Diameter	3	Run	3
Distance			Graph Density	0.031	Run	3
The average graph-dista between any two nodes	nce between all pairs of nodes. Connected nodes have graph distance 1. The diameter is the longest graph distance in the network. (i.e. How far apart are the two most distant nodes).		HITS		Run	۲
			Modularity		Run	۲
ODirected	✓ Normalize Centralities in [0,1]		PageRank		Run	۲
UnDirected			Connected Components		Run	۲
Betweenness Centrality	r: Measures how often a node appears on shortest paths between nodes in the network.		Node Overview			
Closeness Centrality:	The average distance from a given starting node to all other nodes in the network.		Avg. Clustering Coefficient		Run	۲
Eccentricity:	The distance from a given starting node to the farthest node from it in the network.		Eigenvector Centrality		Run	۲
			Edge Overview			
	OK Cance	Avg. Path Length	2.315	Run	?	

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

CNS Cyberinfrastructure for Network Science Center





Gephi: Node Color Ranking & Scaling – Betweenness Centrality

Partition Ranking % Clustering		-						
Nodes Edges	🥥 💎	🔺 🔺						
times_cited		~						
Closeness Centrality		^						
Degree								
Degree								
Eccentricity								
Modularity Class								
Weighted Degree								
number_of_authored_works								
times_cited		\checkmark						
T T								



CNS Cyberinfrastructure for Network Science Center Gephi: Node Labels – Data Laboratory

9 Overview	ata Laboratory	Preview								\mathbf{X}
Data Table 🕺										
des Edges 🔮 Configuration 😌	dd node 🛨 Add edge 🏙	Search/Replace 📳 Imp	ort Spreadsheet	Export table	Rore actions	~			Filter:	Nodes 🗸 🢡
lodes	Id		Label			number_of_authored_		times_cited b	ondel_community_level_0	blondel_community_level_1
Finelli, Lyn	n1	F	inelli, Lyn				9	1389 cor	nmunity_0	community_0
 Blanton, Lenee 	n3	В	lanton, Lenee				4	48 cor	nmunity_0	community_0
Brammer, Lynnette	n4	В	rammer, Lynnette	2			4	48 cor	nmunity_0	community_0
 Smith, Sophie 	n5	s	mith, Sophie				2	17 cor	nmunity_0	community_0
Mustaquim, Desiree	n6	Μ	lustaquim, Desire	e			4	48 cor	nmunity_0	community_0
 Steffens, Craig 	n7	s	teffens, Craig				4	48 community_0		community_0
Leon, Michelle	n9	L	eon, Michelle				3	41 cor	nmunity_0	community_0
Chaves, Sandra S.	n10	C	haves, Sandra S.				5	365 cor	nmunity_0	community_0
Gubareva, Larisa	n11	6	ubareva, Larisa				5	307 cor	nmunity_0	community_0
	n12		all, Henrietta				Т			
								Betweenness C	Modularity Cl	
Add column - Settings ×							1	0.00	4 3	Wasserman, Ss
Add column to nodes							7	0.00	7 0	Vicsek, T
						в	0.26	7 7	Vespignani, A	
olumn Title: Label 2							4	0.0	2 7	Vazquez, A
							3	0.0	2 6	Stanley, He
Type: String 🗸						1	0.00	6 7	Pastor-satorras, R	
							9	0.01	3 0	Oltvai, Zn
							2	0.00	9 7	Munoz, Ma
		OK	C	Cancel			5	0.01	8 0	Kahng, B
V Koni, Kaulin	1155	N	oni, Kaulin				1	0.01	7 1	Garfield, E
Alvarado-ramy, Francisco Kilmarx, Peter H.	n58 n64	A	lvarado-ramy, Fra ilmarx, Peter H.	ancisco			2	0.00	7 7	Barrat, A
A Clarke Kevin D	n65		larke Kewin D				2	0.43	4 0	Barabasi, Al
	-		(11)	1951		64	H		-	

Gephi: Node Labels – Overview

Graph 🕺		4 +	🕼 Label text settings	×
Graph SQ Draggin C C C C C C C C C C C C C		€ Iterarchy	Nodes Edges Show p Select attributes to display as labels	× roperties ⊡
size A A A A A P P m ▼	Image: Second state Image: Second state Image: Second state Image: Second state <th>rial Bold, 32 Color:</th> <th>Size: Scaled ~ Color: Unique ~</th> <th>\</th>	rial Bold, 32 Color:	Size: Scaled ~ Color: Unique ~	\
	Size: Size;		Hide non-selected	
				143

Cyberinfrastructure for Gephi: Node Labels – Overview


CNS Cyberinfrastructure for Network Science Center Gephi: Final visualization – Node Parameters

🌃 Gephi 0.8.2 - Projec	-t 0				– 🗆 X
File Workspace View	Tools Window Plugins Help)			
Overview	Data Laboratory	Preview			XIN
Preview Settings 8		60 Preview XX			
🖌 Presets		Nodes			
Default	~	Border Width	0.025	_	
- Nodes		Border Color	custom [0.0.0]	Preview Settings - Border Color	×
Border Width	1.0		castom [ofofo]		
Border Color	custom [0,0,0]	opacity	100.0		
opacity	100.0	- Noda Labala		Border Color	
Show Labels		INDUC LADEIS		Configures the border color with either the	color of the node or a
Font	cruis Pue	Show Labels	\checkmark	custom color	
Proportional size		East		custoin color.	
Color	custom [0,0,0]	Font			
Shorten label		Proportional size			
Max characters	30	Color	custem [0, 0, 0]		
Outline size	0.0 custom [255 255 255]	Color	custom [0,0,0]		
Outline opacity	80.0	Shorten label			
Box					
Box color	parent	Max characters	30	💿 Custom 🛛 🗖	
Box opacity	100.0	Outline size	0.0		
🗆 Edges		Outline Size	0.0	-	
Show Edges		Outline color	custom [255,255,255]		
Descale weight		Outline enacity	80.0	=	
Color	mixed	Outline opacity	00.0		
Opacity	100.0	Box			
Curved	\checkmark				OK Cancel
Radius	0.0	Box color	parent .		OK Cancel
Edge Arrows		Box opacity	100.0		
Size	3.0		100.0		
Edge Labels					
Font	Arial 10 Plain			0	
Color	original				
Shorten label					
Max characters	30				
Outline size	0.0				
Outline color	custom [255,255,255]				
Outline opacity	80.0				
Preview ratio: 100%	🛞 Refresh				
Export: SVG/PDF/PNG		Background Reset zoom - +			
					Workspace 1 4

Gephi: Final visualization – Node Parameters

Overview	Data Laborat	ory Preview				
Preview Settings %		68 Preview				
📌 Presets		Node Labels		Preview Settings - Font		×
Default		Show Labels	\checkmark			
Nedes		Font	eruis Pule	··· East:	Foot Styles	Sizer
Border Width	1.0	Proportional size		Folic	Fond Style:	5126;
Border Color	custom [0,0,0]	Color	custom [0, 0, 0]	Arial	Plain	10
opacity	100.0	Color	custom [0,0,0]			
Node Labels		Shorten label		Agency FB	∧ Plain ∧	3 🔺
Show Labels		Max characters	30	Algerian	Bold	5
Font	CALIS PLD	Cutting size	0.0	Arial	Italic	8
Proportional size	✓		0.0	Arial Black	Bold Italic	10
Shorten label		Outline color	custom [255,255,255]	Arial Narrow		12
Max characters	30	Outline opacity	80.0	Arial Daundad MT Rold		14
Outline size	0.0	Box		Anal Rounded MT Bold		14
Outline color	custom [255,255,255]			Arial Unicode MS		18
Outline opacity	80.0	Box color	parent	Baskerville Old Face		24
Box		Box opacity	100.0	Bauhaus 93	× ×	36 🗸
Box color	parent	- Edges				
Box opacity	100.0	C Lugos		Preview		
Edges Show Edges		Show Edges	\leq			
Thickness	1.0	Thickness	1.0			
Rescale weight		Rescale weight		The quick brown fox jumps over the	e lazy dog	
Color	mixed					
Opacity	100.0	Color	mixed			
Curved		Opacity	100.0			
Radius	0.0	Curved				
Edge Arrows	2.0	Deditor	0.0		OK	Cancel
Size	3.0	Radius	0.0			
Show Labels				6.X		
Font	Arial 10 Plain				0~	
Color	original					
Shorten label		Preview ratio:	100%			
Max characters	30					
Outline size	0.0					
Outline color	custom [255,255,255]			@ Defrech		
Outline opacity	80.0			v⊘ Keiresn		
D : 11 10001		Export: SVG/P	DF/PNG			
Preview ratio: 100%						
	Ca Refresh					

Gephi: Final visualization – Edge Color Parameters

Gephi 0.8.2 - Project	: 0					- 0
File Workspace View	Tools Win	dow Plugins Hel	р			
Overview		Data Laboratory	/ Pre	view		
						~ 1
Preview Settings 8		-	60 Preview 8			
🕐 Presets						
Default		~				
- Nodes Reades Width	1.0					
Border Color	custom	[0.0.0]	Edges			
opacity	100.0	[0,0,0]	C Lugos		_	Preview Settings - Color X
Node Labels			Show Edges	\checkmark		
Show Labels			Thickness	20.0		Edge Color
Font	deb15 Pub		Deserte meiste			Configures the color of the odges. Edges can either a color on their own
Proportional size	\checkmark	[Rescale weight	L 🛛		(ariginal) or uses incident nodes calar
Color	custom	[0,0,0]	Color	original		(original) or uses incident hodes color.
Shorten label			Onacity	100.0		
Max characters	30		opuony			
Outline size	0.0		Curved	\leq	CI	Original
Outline color	custom	[255,255,255]	Radius	0.0		
Outline opacity	00.0					○ Mixed
Box color	narent		E Edge Arrows	5	_	0
Box opacity	100.0		Size	3.0		○ Source
Edges			Edge Labels			0
Show Edges					_	○ Target
Thickness	1.0		Show Labels			
Rescale weight			Font	Arial 10 Plain		O Curture
Color	mixed		Color	original		Custom
Opacity	100.0		COIOT	originai		
Curved			Shorten label			
Radius	0.0		Max characters	s 30		OK
Size	3.0		Outline size			OK Cancel
Edge Labels			Outline size	0.0	_	
Show Labels			1			
Font	Arial 10	Plain				
Color	original					
Shorten label						
Max characters	30		_			
Outline size	0.0		-			
Outline color	custom	[255,255,255] [
Outline opacity	00.0					
Preview ratio: 100%						
		🐵 Refresh				
Export: SVG/PDF/PNG			Background Reset	t zoom - +		

Gephi: Label Adjustments



Cyberinfrastructure for Network Science Center

Layout algorithm

Layout 🕺	-					
ForceAtlas 2	~					
Counter-Clockwise Rotate	~					
Expansion						
Force Atlas						
ForceAtlas 2						
Fruchterman Reingold						
Label Adjust						
Random Layout						
Yifan Hu	~					

Manual adjustments



Cyberinfrastructure for Network Science Center Gephi: Exporting the visualization and networks



Cyberinfrastructure for Network Science Center Gephi: Exporting the visualization and networks





Questions?



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

Web of Science™	InCites ™	Journal Citation Reports ®	Essential Science Indicators SM	EndNote ™			
WEB OF	= SCII	ENCE™					
Search	All Databas	es 🔽					
							Weld
Basic Search	~						
Example: oil s	spill* medi	terranean		8	Торіс	∽ Searc	h
			+ Add Another Field	Reset Form			
TIMESPAN							
All years	~					/ /	· /
C From 1964	to	2016		T			7
1004	• 10	2010					
MORE SETTIN	IGS						/
				V/			
							9
				\prod	/ /		/
				//		/ /	
				117			/
			CE	NTE	RS FOR	DIS	EAS
			Col	NTRO	DL AND F	REVEN	OITR

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. CNS Cyberinfrastructure for Network Science Center

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

te Project	« Star	t Over Configure Pars	sing Options		Proje	ect name CDC AIDS Diag	nosis CityLoo	ation 8196	Create Project
n Project	No	otes Location	Location Code	Year Diagnosed	Year Diagnosed Code	Race or Ethnicity	Race	or Ethnicity Code	Cases
rt Project	1.	Akron, OH	80	1982	1982	Black (and also not Hispa	anic) 2054	-5	1
it i roject	2.	Akron, OH	80	1984	1984	White (and also not Hispa	anic) 2106	-3	3
	3.	Akron, OH	80	1985	1985	White (and also not Hispa	anic) 2106	-3	4
	4.	Akron, OH	80	1986	1986	Black (and also not Hispa	anic) 2054	-5	3
	5.	Akron, OH	80	1986	1986	White (and also not Hispa	anic) 2106	-3	9
	6.	Akron, OH	80	1987	1987	Black (and also not Hispa	anic) 2054	-5	6
	7.	Akron, OH	80	1987	1987	White (and also not Hispa	anic) 2106	-3	14
	8.	Akron, OH	80	1988	1988	Black (and also not Hispa	anic) 2054	-5	11
	9.	Akron, OH	80	1988	1988	Hispanic	2135	-2	1
	10.	Akron, OH	80	1988	1988	White (and also not Hispa	anic) 2106	-3	15
	11.	Akron, OH	80	1989	1989	Black (and also not Hispa	anic) 2054	-5	7
	12.	Akron, OH	80	1989	1989	White (and also not Hispa	anic) 2106	-3	19
	13.	Akron, OH	80	1990	1990	Black (and also not Hispa	anic) 2054	-5	6
	14.	Akron, OH	80	1990	1990	White (and also not Hispa	anic) 2106	-3	23
	15.	Akron, OH	80	1991	1991	Black (and also not Hispa	anic) 2054	-5	11
	16.	Akron, OH	80	1991	1991	White (and also not Hispa	anic) 2106	-3	38
	17.	Akron, OH	80	1992	1992	Black (and also not Hispa	anic) 2054	-5	17
	18.	Akron, OH	80	1992	1992	White (and also not Hispa	anic) 2106	-3	41
	19.	Akron, OH	80	1993	1993	Black (and also not Hispa	anic) 2054	-5	28
	20.	Akron, OH	80	1993	1993	White (and also not Hispa	anic) 2106	-3	37
	21.	Akron, OH	80	1994	1994	Black (and also not Hispa	anic) 2054	-5	15
	22.	Akron, OH	80	1994	1994	Hispanic	2135	-2	1
	23.	Akron, OH	80	1994	1994	White (and also not Hispa	anic) 2106	-3	31
	24.	Akron, OH	08	1995	1995	Black (and also not Hispa	anic) 2054	-5	21
	25.	Akron, OH	80	1995	1995	Hispanic	2135	-2	1
	26.	Akron, OH	80	1995	1995	White (and also not Hispa	anic) 2106	-3	32
	27.	Akron, OH	80	1996	1996	Black (and also not Hispa	inic) 2054	-5	1/
	Pars	se data as	Charact	er encoding				U	pdate Preview
	CSV	/ TSV / separator-bas	ed files Column	s are separated b	у	Ignore first	0 line(:	s) at beginning of	file
	Line-	based text files	O cor	nmas (CSV)		Parse next	1 line(s	s) as column hea	iders
	Fixed	l-width field text files	tab	s (TSV)		Discard initial	0 row(s	s) of data	
	PC-A	vic toxt files	O cus	tom \t		Load at most	0 row(s	s) of data	
-	107		Escape	special characte	rs with \		- '	·	
	JSON	V files							
2	RDF/	N3 files				Parse cell text int	0	✓ Store blank	rows
sion 2.5 [r2407]	XML	files				numbers, dates, .	 are used	Store blank	cells as nulls
Help	Open	Document Format spre	eadsheets			to enclose cells o	are used ontaining	Store file so (file names	urce URLs)

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

) -	Application Tools Ref	ne	-			
File Hom	e Share View	Manage					
$\leftrightarrow \rightarrow \cdot \cdot$	> This PC > Lo	cal Disk (C:) → Users → r	nginda > Refine	✓ Ö Search Refine			
📌 Quick acc	C:\Users\mgi	nda\Refine\google-refine.exe				-	
Docume	13:38:23.373 13:38:23.380	[refine_ [refine_	server] Starting Server bound to server] Initializing context: '/	'127.0.0.1:3333' (Oms) ' from 'C:\Users\mginda\Refin	e\webapp'	(7ms)	
Desktop	13:38:24.175 ables (795ms)	[refine_	server] Failed to use jdatapath	to detect user data path: res	orting to	environm	ent vari
🕹 Downlo	ads 13:38:24.176 ables (1ms)	[refine_	server] Failed to use jdatapath	to detect user data path: res	orting to	environm	ent vari
Pictures	13:38:24.183	[refine] Starting Google Refine 2	.5 [r2407] (7ms)			
📙 mginda							
a OneDrive							
💻 This PC	(127.0.0.1:333	3	C ^ℓ Q OpenRefine youtube →	☆自 🗎 - Z + 合			
📃 Desktop	Google refine	A power tool for working with mess	data.				
🔮 Docume	000,000						
- Downlo	Create Project	Create a project by impo TSV, CSV, *SV, Excel (.xls	ting data. What kinds of data files can I im and .xlsx), JSON, XML, RDF as XML, and Go	<pre>,port? ogle Data documents are all supported. S[;]</pre>	L		
Music	Open Project	can be added with Google I	tefine extensions.				
Videos	1 2	Get data from	Locate one or more files on your com	puter to upload:			
Local Di		This Computer	Browse No files selected.				
KINGST		Web Addresses (URLs)	Next »				
👳 mginda		Cilpboard Google Data					
🛫 projects		Coogle Data					
KINGSTO							
i Network							
	Version 2.5 [r2407]						
	Help About						

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

127.0.0.1 :333	3	C OpenRefine youtub	be→☆自 🗎 - Z 🔸	n Onc	e Oper
g <mark>le</mark> refine	A power tool for working with messy data.			war	nt to cre
Project Project	Create a project by importing TSV, CSV, *SV, Excel (.xls and Refine extensions.	data. What kinds of data files ca .xlsx), JSON, XML, RDF as XML, an	n l import? nd Google Data documents are alli support	ed. Su	
Project	Get data from	Locate one or more files on you	r computer to upload:	First	t, you n
	This Computer	Browse CDC-articles-20041	4-isi-topcited.csv		rco wh
	Web Addresses (URLs) Clipboard Google Data	Next »		data C:/ "CD Onc	a file or ./Sci2-C C-article
		 ● File Upload ← → ~ ↑ □ → Sci2 Organize → New folder 	-CDC-data > Networks > coauthor		v 0
		mginda 🖈 ^	Name	Date modified	Туре
		a OneDrive	CDC-articles-200414-isi-topcited.csv	6/10/2016 9:33 AM	Microsoft Excel C
en 2.5 (r2407) Help About			CDC-articles-200414-isi-topcited.txt	6/8/2016 3:04 PM	Text Document
		File pa	me: CDC-articles-200414-isi-topcited csv		~

Cyberinfrastructure for Network Science Center

€ 0 Goo

Create Open F

Import

Versi

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.

III - 🔟 🕐

Search Networks

238 KB

189 KB

Size

All Files (*.*)

Open

× م

 \sim

Cancel



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

Create Project	«	Start Over	Configure Parsing	Options		Project name	e CDC WoS Arti	cles topcited csv	Create	Project »
Open Project		Abstract	Associated Group	Author Identifier	s Authors	Authors (Full Names)	Beginning Page	BIOSIS Citation Index	Book Author	Book Autho
mport Project	1.				Rolfes, Melissa[Blanton, Lenee]Brammer, Lynnette]Smith, Sophie]Mustaquim, Desiree]Steffens, Craig[Cohen, Jessica]Leon, Michelie]Chaves, Sandra S, JAbd Elal, Amwar Isa]Gubareva, Larisa]Hali, Henrietta]Wallis, Teresa]Wilanueva, Julie]Xu, Xiyan]Bresee, Joseph[Cox, Nancy[Fineli], Lyn		1189	9		
	2.				Ridpath, Alison Driver, Cynthia R. Nolan, Michelle L Karpati, Adam Kass, Daniel Paone, Denise Jakubowski, Andrea Hoffman, Robert S. Nelson, Lewis S. Kunin, Hillary V.		1195	2		
	3.				Summers, Aimee Nyenswah, Tolbert G. Montgomery, Joel M. Neatherlin, John Tappero, Jordan W.		1202	4		
	<									>
	L	ine-based t	ext files	• co	mmas (CSV)	\checkmark	Parse next	1 line(s) as colu	umn headers	
	F	ixed-width	field text files	O ta	os (TSV)		Discard initial	0 row(s) of data	ı	
	F	PC-Axis text files		U cu	stom ,		Load at most	0 row(s) of data	I.	
	J	SON files		Escap	e special characters with t					
	F	RDF/N3 files	3			\searrow	Parse cell text in numbers, dates	to Stor	re blank rows	
	Х	ML files					Quotation marks	are used	e plank cells a	as nulis
	(Open Docun .ods)	nent Format spread	sheets			to enclose cells column separato	containing (file rs in ea	names, URLs ach row)

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

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

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

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

cet / Filter Undo	Redo o	409 rows	409 rows						
	~	Show as: rows records	Show as: rows records Show: 5 10 25 50 rows						
Lising facets and	ilters	All Abstract	Associa	ted Gro 💌 Author Identifie 💌 Authors 💽 Au	uthc				
osing lacets and		Facet		Rolfes.					
Use facets and filters of your data to act on	^t Re-order / Re	emove Columns							
filter methods from the	9								
of each data column.	Drag columns to	o re-order	Drop columns here to remove						
Not sure how to get s Watch these screen			^	A h = h = = = +					
	Authors			Abstract					
	Cited Refere	nce Count		Associated Group					
	Cited Refere	nces		Author Identifiers					
	Conference [Dates		Authors (Full Names)					
	Conference	Host	1	Beginning Page					
	Conference I	ocation		BIOSIS Citation Index	Ī				
	Conference (2 control		Book Author					
	Conterence	sponsors	-	Book Author Fullname	1				
	Conference	litle		Book DOI					
	Document Ty	/pe		Pack Crown Authors					
	DOI								
	Edited by			Book Series Subtitle					
	Editors			Book Series Title					
	Ending Page)		Chinese Science Citation DB					
	File Type			Cited Patent					
	Funding Age	ncv and Grant Number	~	City of Publisher					
	in analisi Aigo	ney and orant namber			_				

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. Cyberinfrastructure for Network Science Center Articles from 2004-2014 – Editing Columns In Bulk

Google refine	CDC WoS Ar	ticles t		
Facet / Filter Und	o / Redo o	to remove	^ ^	
Facet / Filter	Undo / Redo 7	<u> </u>	409 row	
	Extract	Apply	Show as: n	
Con Filter	Book	Author Fullname		
Conference Title	Apply Operation Histo	rv		
Document Type		• 3		
DOI	Paste an extracted JSON his	story of operations to perfor	m:	
Edited by				
Editors				
Ending Page				
File Type				
Funding Agency and Grant Nu				
OK Cancel				
	Perform Operations Can	icel		

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

		Apply Operation History				To apply the relarder tasks pavigate
		Paste an extracted JSON history of operations to perform: "File Name", "Cite Me As"				to C://Sci2-CDC-data > Networks :
1 2	[1 }, {				coauthor > refine >
3 4 5	"op": " "descri; "column	"op": "core/column-reorder", "description": "Reorder columns", "columnNames": ["Authore",				CDC-WoS-JSON-1-Preprocessing.txt
6 7 8	"Auth "Cite "Cite "Conf	"DQI", "ISSN", "Journal Title (Full)", "Publication Date",				and open the file in Notepad.
10 11 12	"Docu "DOI" "Edit	"Publication Type", "Publication Year", "Publisher", "Times Cited",				Copy the text (Ctrl-c) and then Past
13 14 15	"Edit "Fund "Fund	"Title", "Total Times Cited", "Unique ID", "File Nere"				(Ctrl-V) in the Apply Operation
16 17 18	"ISI "ISBN "ISSN	"Cite Me As"				History window in OpenRefine.
19 20 21	"Jour "Jour "Jour	"op": "core/column-removal", "description": "Remove column Publisher", "columnName": "Publisher"				Then select the Perform Operations
22 23 24	"Lang "New . "New]				button.
25 26	"Orig "Publ	Perform Operations Cancel	Google refine CDC WoS Articles t	opcited cs	V Per	rmelink
28	"Publi	ication Type", ication Year",	Facet / Filter Undo / Redo 7	409 ro	ws	Presente Charles C 40 20 00 million
29 30	"Publi "Resea	isher", arch Field",	Extract Apply		. TOWS	Authors
31	"Resea	archer ID",	Filter:	☆ <i>되</i>	1. Ro	offes, MelissajBlanton, LeneejBrammer, LynnetteijSmith, SophiejMustaquim, DesireejSteffens, Craig(Cohen,
33	"Subje "Times	s Cited",	0. Create project		Ter	rssical,een, micheliejonaves, Sandra S.JAdo Elai, Aniwar Isa(Subareva, Lansajnai, Hennettajiwaiiis, aresa[Villanueva, Julie]Xu, Xiyan Bresee, Joseph Cox, Nancy Finelli, Lyn
34	"Title	e", I Times Cited"	1. Reorder columns			
36	"Uniqu	le ID",	2. Reorder columns			
37 38	"File "Cite	Name", Me As"	4. Reorder columns	84	2. Ric An	idpath, AlisonJDrver, Cynthia R. Nolan, Michele L. Karpat, AdamKass, DanielPaone, DeniseJakubowski, ndrealHoffman, Robert S. Nelson, Lewis S. Kunin, Hillary V.
40),		5. Reorder columns	-		
41	{ "op": "(core/column_reorder"	e. Reorder columns	~ ~	2 0	anna Aine Nearaich Tabad O Nachanna Iadh Nachada Isla Taona Iada M
16	ο μ . (7. Remove column Publisher	2 4	3. SU	ummers, Ameejnyens wan, robert os,mongomery, Joer muneatherim, Johni lappero, Jordan W.

CNS Cyberinfrastructure for Network Science Center

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

Cluster & Ed	lit column '	'Authors''		
This feature hel york" are very li	lps you find gro ikely to refer to	ups of different cell values that might be alter the same concept and just have capitalization	rnative repre on difference	esentations of the same thing. For example, the two strings "New York" and "new es, and "Gödel" and "Godel" probably refer to the same person. Find out more
Method neares	st neighbor \vee	Distance Function levenshtein V	Rad	ius 1.0 Block Chars 6 67 clusters found
Cluster Size	Row Count	Values in Cluster PPM	Merge?	New Cell Value A # Rows in Cluster
2	2	• Jiles, Ruth (1 rows) • Jiles, Ruch (1 rows)		Jiles, Ruth
2	6	 Hlavsa, Michele C. (4 rows) Hlavsa, Michele C (2 rows) 		Hlavsa, Michele C. Average Length of Choices
2	2	 Onwen, Diana H (1 rows) Onweh, Diana H (1 rows) 		Onwen, Diana H 8 - 27
2	2	• Blau, Dianna M. (1 rows) • Blau, Dianna M (1 rows)		Blau, Dianna M.
2	4	 Richardson, Lisa C (2 rows) Richardson, Lisa C. (2 rows) 		Richardson, Lisa C 0 - 0.5
2	3	Meltzer, Martin I. (2 rows) Meltzer, Martin I (1 rows)		Meltzer, Martin I.
2	4	 Frieden, Thomas R (3 rows) Frieden, Thomas R. (1 rows) 		Frieden, Thomas R
Select All	Deselect All			Merge Selected & Re-Cluster Merge Selected & Close Close

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."



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.

ethod neares	st neighbor 🗸	Distance Function levenshtein 🗸	Rad	ius 1.0 Block Chars 6	67 clusters found
Cluster Size	Row Count	Values in Cluster PPM	Merge?	New Cell Value	# Rows in Cluster
2	2	 Jiles, Ruth (1 rows) Jiles, Ruch (1 rows) 		Jiles, Ruth	
2	6	 Hlavsa, Michele C. (4 rows) Hlavsa, Michele C (2 rows) 		Hlavsa, Michele C.	2 — 12 Average Length of Choices
2	2	 Onwen, Diana H (1 rows) Onweh, Diana H (1 rows) 		Onwen, Diana H	₿ — 27
2	2	 Blau, Dianna M. (1 rows) Blau, Dianna M (1 rows) 		Blau, Dianna M.	Length Variance of Choices
2	4	 Richardson, Lisa C (2 rows) Richardson, Lisa C. (2 rows) 		Richardson, Lisa C	0 — 0.5
2	3	 Meltzer, Martin I. (2 rows) Meltzer, Martin I (1 rows) 		Meltzer, Martin I.	
2	4	 Frieden, Thomas R (3 rows) Frieden, Thomas R. (1 rows) 		Frieden, Thomas R	

Cyberinfrastructure for Network Science Center Articles from 2004-2014 – Clustering and Editing

409 rows	5		
Show as: ro	ws records Show	v: 5 10 25 50 rows	
	Unique ID	Authors	
☆ 듸 1.	WOS:000346946700002	Facet Text filter	enee Brammer, Lynnette Smith, Sophie Mu aves, Sandra S. Abd Elal, Anwar Isa Gub ı, Xiyan Bresee, Joseph Cox, Nancy Fine
		Edit cells	Transform
		Edit column	Common transforms
☆ 듸 2.	WOS:000346946700003	Transpose 🕨	Fill down lamK
		Sort	Blank down
		View 🕨	Split multi-valued cells
		Reconcile 🕨 🕨	Join multi-valued cells
☆ 🗐 3.	WOS:000346946700005	Summers, Aimee Nyensw	Cluster and edit eathe



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**.

40)9 re	eco	ords						
Sh	iow as	s: ro	ows records Show	v: 5 10 25 50 rec	ords				
	All		V nique ID	 Authors 	DOI	T ISSN	Journa		
☆		1.	WOS:000346946700002	Rolfes, Melissa		0149-2195	MMWR-MOR MORTALITY		
숤				Blanton, Lenee					
				Brammer, Lynnette					
				Smith, Sophie			9		
				Mustaquim, Desiree					
				Steffens, Craig					
				Cohen, Jessica					
				Leon, Michelle					
				Chaves, Sandra S.					
				Abd Elal, Anwar Isa					
				Gubareva, Larisa			******		
				Hall, Henrietta			******		
				Wallis, Teresa			******		
1	n			Villanuova Iulia					

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

3216 rov	vs		Facet / Filter	Undo / Redo s
Show as: ro	ows records Sho	w: 5 10 25 50 rows	Refresh	Reset All Remove All
	Authors	DOI 💌 ISSN 💌 Journal Titl	× Authors	change
☆ 듸 1.	Facet 🕨	Text facet	2404 choices Sort	by: name count Cluster
	Text filter	Numeric facet	Abd Elal, Anwar 1	<u> </u>
	Edit cells	Timeline facet	Abd Elal, Anwar Is	a 3
숫 디 2.	Edit column	Scatterplot facet	Abellera, John P 2	
☆ 듸 3.	Transpose 🕨	Custom text facet	Aberle-Grasse, Joh	nn M. 1
숤 듸 4.	Sort	Custom numeric facet	Abma, Jovce C. 1	
☆ 🎝 5.	View	Customized facets	Abouzelof, Rouett	H. 1
☆ 듸 6.	View		Abramova, L. 1	
☆ 듸 7.	Reconcile 🕨 🕨		Abrams, Steven 1	
☆ 🏹 8.	Leon, Michelle		Abrar, Leila 1	
☆ 듸 9.	Chaves, Sandra S.		Ackelsberg, Joel 1	~
52 57 10.	Abd Elal, Anwar			

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

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.

32	3216 rows										
Sho	Show as: rows records Show: 5 10 25 50 rows										
	AII		Authors	uthors		ISSN	Journal Ti	itle (l			
☆		1.	Facet Text filter	•		0149-2195	MMWR-MORBIE WEEKLY REPOR	nty A Rt			
			Edit cells	►	Tra	nsform					
☆	5	2.	Edit column Common transforms								
☆		3.	Transpose	►	Fill	down					
☆	5	4.	Sort		Blank down						
☆		5.	View	•							
হি	5	6.		-	Spi	it multi-valu	ed cells				
☆		7.	Reconcile	•	Joir	n multi-valu	ed cells				
ਨਿ	9	8.	Leon, Michelle		Chu	eter and ed	i+				
☆		9.	Chaves, Sandra S.		Ciu	ster allu eu	n				
☆	57	10.	Abd Elal, Anwar Isa								
25	173	11	Gubareva Larica								

Open Refine: Clustering and Editing – Key Collision



Cyberinfrastructure for

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.



Cyberinfrastructure for

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

ппрегриптинен

Cuberinfrastructure for

- 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.

					1
Aethod b	ey collision	 Keying Function Ingram-fin 	gerprint 🗠	Ngram Size 1	31 clusters
		 Miles, Isa (1 rows) 			# Choices in Cluster
2	2	Biermann, Janis (1 rows) Miner, James B (1 rows)		Biermann, Janis	
2	2	Kallen, A. (1 rows) Keenan, N. L. (1 rows)		Kallen, A.	2 − 4 # Rows in Cluster
2	2	Rosenman, Kenneth (1 rows) Thomas, Karen E. (1 rows)		Rosenman, Kenneth	2-6
2	2	Kahn, Katherine E. (1 rows) Tan, Kathrine R. (1 rows)		Kahn, Katherine E.	Average Length of Choices
2	2	Carter, Marion (1 rows)		Carter, Marion	7 — 20

Phonetic fingerprint

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

ппрегриптинет

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.

Clus	Cluster & Edit column "Authors"										
This feature helps you find groups of different cell values that might be alternative representations of the same thing. For exampl york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refe											
Meth	od key co	llision ~	Keying Function met	aphone3	\sim						
Clu	ster Size	Row Count	Values in Cluster	Merge?	New Cell Value	^	# (
3	3 3 • Freedman, David (1 rows) • Freedman, David O. (1 rows) • Freedman, David S. (1 rows)										

Cluster & Ed	Cluster & Edit column "Authors"											
This feature helps you find groups of different cell values that might be alternative representations of the same thing. For examply york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably references (and "Godel" probably references) (and "Godel" probably references) (and "Godel" probably references) (and "Godel" probably references) (begin to be alternative capitalization (cologne-phonetic)) (begin to be alternative capitalization) (cologne-phonetic)) (cologne												
Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value	^	#						
2	3	 Lu, Hua (2 rows) Luo, Yao-Hua (1 rows) 		Lu, Hua								

Cluster & E	dit column '	'Authors"			
This feature he york" are very	elps you find gro likely to refer to	oups of different cell values that might be alte the same concept and just have capitalizati	rnative repre on differenc	esentations of the same thing. For exa es, and "Gödel" and "Godel" probably	ample, the two strings "New York" and "new refer to the same person. Find out more
Method neare	est neighbor $$	Distance Function levenshtein ~	Rad	ius 1.0 Block Chars 6	67 clusters found
Cluster Size	Row Count	Values in Cluster PPM	Merge?	New Cell Value	# Rows in Cluster
2	2	 Jiles, Ruth (1 rows) Jiles, Ruch (1 rows) 		Jiles, Ruth	
2	6	 Hlavsa, Michele C. (4 rows) Hlavsa, Michele C (2 rows) 		Hlavsa, Michele C.	2 — 12 Average Length of Choices
2	2	 Onwen, Diana H (1 rows) Onweh, Diana H (1 rows) 		Onwen, Diana H	8-27
2	2	 Blau, Dianna M. (1 rows) Blau, Dianna M (1 rows) 		Blau, Dianna M.	Length Variance of Choices
2	4	 Richardson, Lisa C (2 rows) Richardson, Lisa C. (2 rows) 		Richardson, Lisa C	0 — 0.5
2	3	 Meltzer, Martin I. (2 rows) Meltzer, Martin I (1 rows) 		Meltzer, Martin I.	
2	4	 Frieden, Thomas R (3 rows) Frieden, Thomas R. (1 rows) 		Frieden, Thomas R	
Select All	Deselect All			Merge Selected & Re-Clu	Ister Merge Selected & Close Close

Nearest Neighbor

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

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

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



Levenshtein Distance

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

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

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

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

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

PPM

<u>Prediction by Partial Matching</u> 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.

09 reco	ords					321	6 rov	vs						
Show as: ro	ows records Show	v: 5 10 25 50 red	ords			Show	vas: ro	ows records Show	r: 5 10 25 50 ro	ows				
All	Unique ID	Authors	💌 DOI	💌 ISSN	Journal		11	▼ Unique ID	Authors		DOI		Journal T	itle (Full)
<u>}</u> ק 1.	WOS:000346946700002	Rolfes, Melissa		0149-2195	MMWR-MOR MORTALITY	\$		WOS:000346946700002	Facet Text filter	١		0149-2195	MMWR-MORBI WEEKLY REPO	DITY AND MORTALITY RT
									Edit cells	•	Tra	nsform]
						~	2 2		Edit column	•	Cor	mmon trans	forms 🕨	
		Blanton, Lenee			······································		-7 - . -7 - 3		Transnose					-
		Smith Sonhie			······				manopooe	-	Fill	down		
		Mustaquim Desiree					-7 ⁻ .		Sort		Bla	nk down		
		Steffens, Craig					-7 J.		View	•	Spli	it multi-value	ad cells	-
		Cohen, Jessica				23	~ •.		Desensile		opi			
		Leon, Michelle				23	J 1.		Reconcile	•	Joir	n multi-value	ed cells	
		Chaves, Sandra S.				23	- 8. 		Leon, Michelle		Clu	ster and ed	t	
		Abd Elal, Anwar Isa			······	23	9.		Chaves, Sandra S.	-		_		
		Gubareva, Larisa				53	5 10.		Abd					
		Hall, Henrietta				57	11 .		Gut					
		Wallis, Teresa					1 12		Hall	hat se	epara	ator curren	tly separates	the values?
ted csv	Permalink				Join n	ulti-valued cells in	colum	n Authors Undo					ОК	Cancel
09 rec	ords													
now as: j	rows records	Show: 5 10 25	50 records						rendan Finelli,	Lyn Fr	у,		0149-21	AND MORTAL ITY
AII	Unique ID	Author	s				💌 DOI							
<i>ि</i> 1.	WOS:000346946700	0002 Rolfes, Melis Desiree Stef Isa Gubarev Joseph Cox,	Rolfes, Melissa Blanton, Lenee Brammer, Lynnette Smith, Sophie Mustaquim, Desiree Steffens, Craig Cohen, Jessica Leon, Michelle Chaves, Sandra S. Abd Elal, Anwar Isa Gubareva, Larisa Hall, Henrietta Wallis, Teresa Villanueva, Julie Xu, Xiyan Bresee, Joseph Cox, Nancy Finelli, Lyn											
	WOS:000346946700	0003 Ridpath, Alis Denise Jakul	on Driver, Cynthia R. oowski, Andrea Hoffr	Nolan, Michelle L. I nan, Robert S. Nel:	Karpati, Adam∣Ka son, Lewis S.∣Kı	s, Daniel Paone, in, Hillary V.		014	is					
														171

💌 All 🔍 Uni	que ID	Authors					
Facet >	346946700002	Rolfes, Melissa Blanton, L Desiree Steffens, Craig C Isa Gubareva, Larisa Hall, JosenblCox, Nancv/Finelli					
Edit rows							
Edit columns	Re-order / remove columns						
View 🕨							
☆ ק 2. WOS:00	0346946700003	Ridpath, Alison Driver, Cyr Denise Jakubowski, Andre					

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

rag columns to re-order	Drop columns here to remove
Authors	
DOI	
ISSN	
Journal Title (Full)	
Publication Date	
Publication Type	
Publication Year	
Times Cited	
Title	
Total Times Cited	
File Name	
Cite Me As	
Unique ID	
	Ĭ

Open Refine: Exporting of CDC Web of Science Cyberinfrastructure for Network Science Center Publications

Export project Tab-separated value Comma-separated value HTML table Upc Act Sta Sep ODF spreadsheet	Open Export • Export project • Tab-separated value • Comma-separated value • HTML table • Excel • ODF spreadsheet •				In Operight of the Ex Comm as bel	enRefine, corner of cornt drop na-separa ow.	navigate the scree o down m i ted value	to the u n, and se enu, and . Save th	pper elect d select he file
28- 201 Triple loader MQLWrite Illne Dea Custom tabular expor Per Atti Templating Electronic Dance Music	ter VCDC article		nter name o nter name o anize This PC Desktop Downloa Downloa Downloa Downloa Downloa Comparison Nusic Pictures Videos Local Dis KINGSTO mginda projects KINGSTON File na Save as t Save as t	f file to sav	re to ici2-CDC-data > Net ter Name Name -WoS-Articles-topcite osoft Excel Comma Se	tworks > coauthor > r ^ ed-processed-disambig.cs sparated Values File (*.csv)	esults Date modified	v ð Type	Search Networks

 \sim \sim

Cancel

×

Q

?

•== **-**

C	G <mark>oogle</mark> re	fine CDC	WoS Articles	t		
1	Facet / Filter	Undo / Red	l o o			
			Facet / Filter	Undo / Redo 7		409 row
	Using face	nd filtere to ex		Extract	Apply	Show as: n
	Use lacets a	na iliters to se	Filter			



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. Cyberinfrastructure for Open Refine: Cluster & Editing the Authors column in bulk

🔚 Redshi	ift-pw.txt 🗵 [anew 3 🔀 📇 ADNI-2016-Journal Analysis. R 🗷 🔚 CDC-WoS-JSON-2-ClusterEdit.txt 🗵
1758		1.
1759		To". "Ibluwalia Indu B "
1760	1	
1761		
1762		
1762	11	
1703	1	
1/64	"op"	: "core/mass-eqit",
1765	"des	cription": "Mass edit cells in column Authors",
1766	"eng	ineConfig": {
1767	"n	ode": "row-based",
1768	"f	acets": []
1769	},	
1770	"col	
1771	"exp	Apply Operation History
1772	"ed:	
1773	{	
1774		Paste an extracted JSON history of operations to perform:
1775		"expression": "value",
1776		"edits": [
1777		
1778		"ITOMBIANK": Talse,
1779		
1780		"Thomas, Ann R.",
1701	,	"Thomas, Ann"
1701],
1702	. 1	"to": "Thomas, Ann R."
1/63	11	
1784	۲	"fromBlank": false.
1785	"op	"fromError": false,
1786	"des	"from": [
1787	"eno	"Moran, John",
1788	"I	"Moran, John S."
1789	"1	J, "", "Moran John"
1790	},	
1791	"col	
1792	"exp),
1793	"ed:	
1794	{	"op": "core/multivalued-cell-join",
1795		"columname": "Authors".
		"keyColumnName": "Unique ID",
		"separator": " "
		3
		[1
		Defense Occupiers Consul

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

😵 Sci2 Tool	– 🗆 X			
File Data Preparation Preprocessing Analysis Modeling Visualization R He	p			
E Console	🗖 🗖 1889 Data Manager 🔤 🗖	Το ον	tract a co-a	withor notwork f
vereicome to the science of science lool (sci2). The development of this tool is supported in part by the Cyberinfrastructure for Networ center and the Department of Information and Library Science in the School of Information Cyberinfrastructure portal (http://sci2.cns.iu.edu) for documentation, screenshots, and IIS-0715303, and the James S. McDonnell Foundation. See the Science of Science Cyberinfrastructure portal (http://sci2.cns.iu.edu) for documentation, screenshots, and expert. The primary investigators are Katy Börner, Indiana University and Kevin W. Boyack, SciT. Strategies Inc. The Sci2 tool was developed by Daniel Halsey, Adam Simpson, Saumya P Sumit Samant, Vivek Karihaloo, Chin Hua Kong, Steven Corenflos, Joseph Biberstine, Th Smith, David M. Coe, Micah W. Linnemeier, Patrick A. Phillips, Chintan Tank, and Russ It uses the Cyberinfrastructure Shell (http://cishell.org) developed at the Cyberinfrastructure Metwork Science Center (http://cishell.org) at leavelue). Please cite as follows:	s Science is cand 0738111 to ask an andey, omas G. Li Duhon. ture for is were	navig Netw CDC- proce	gate to C:/ vorks > coar WoS-Article essed-disar	./Sci2-CDC-data = uthor > results > <i>es-topcited-</i> <i>mbig.csv</i>
Sci2 Team. (2009). Science of Science (Sci2) Tool. Indiana University and SciTech Strateg https://sci2.cns.iu.edu. Loaded:	← → ~ ↑ 📙 « Sci2-CDC-data > Networks > coauthor > result:	ن ~ ت	Search results	م
C·Users\mginda\Desktop\Sci2-CDC-data\Networks\CDC-WoS-Articles-topcited-proce	Organize 🔻 New folder		:== :==	- II ?
Remove From List Remove completed automatically Remove all completed	ConeDrive Name	Date modified	Туре	Size
	CDC-WoS-Articles-topcited-processed-d	6/10/2016 10:42 AM	Microsoft Excel C	178 KB
! Algorithm Name Date Time % Con	This PC	6/10/2016 11:45 AM	Microsoft Excel C	67 KB
	 Desktop Documents Downloads Music Pictures Videos Local Disk (C:) KINGSTON (D:) mginda (\\smb.) KINGSTON (D:) 	6/10/2016 11:43 AM	Text Document	2 KB
	File name: CDC-WoS-Articles-topcited-processed-disa	mbig.csv ~	**	~
			Open	Cancel

Cyberinfrastructure for Network Science Center Sci2: Extracting a co-author network

🤣 S	ici2 Tool	
File	Data Preparation Preprocessing Analysis Modeling Visualization R Help	p
Welc Welc The cent Corr and Cybi expe	Convert to Generic Publication Remove ISI Duplicate Records Remove Rows with Multitudinous Fields Extract Directed Network Extract Bipartite Network Extract Paper Citation Network	Science s and 738111 ask an
The Strat Sum Smit It us Netv deriv	Extract Co-Occurrence Network Extract Word Co-Occurrence Network Extract Co-Author Network Extract Co-Author Network Extract Reference Co-Occurrence (Bibliographic Coupling) Network Extract Document Co-Citation Network	h ndey, mas G. J. Duhon. Jre for were
Plea	Data at Dualizate Made	

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.

Extract C	o-Author Network	×					
Extracts a co-authorship network from one of several supported file types.							
File Format	isi	~ 📀					
	[OK Cancel					

🚻 Data Manager		
CSV file: C	:\Users\mginda\AppData\Local\Temp\temp\Preprocesse	d-CDC-V
🖏 Extract	ted Co-Authorship Network	
Author	r information	

Sci2: Viewing the Authors list

1010 Data Mana	ger	- 8
∽ I CSV fil S Ext	le: C:\Users\mginda\Appl tracted Co-Authorship Ne	Data\Local\Temp\temp\Preprocessed-CDC-V twork
At At	Save	
	View	
	View With	
	Rename	
	Discard	
ľ		A1 - :

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

A	L 🕘 E 🗶 🗸	f_x label			
	A	В	С	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	*

Cyberinfrastructure for Sci2: Applying the Network Analysis Toolkit

Analysis	Modeling	Vi	ualization R Help	
Tem	poral	>	🗖 🗖 🔤 🖬 🖬 🖬	
Geos	spatial	>	CSV file: C:\Users\mginda\AppData\Lo	CZ
Торі	cal	>	🕵 Extracted Co-Authorship Network	
Netv	vorks	>	Network Analysis Toolkit (NAT) Author information	
dicelaw/CIS		rt.	Unweighted & Undirected >	
uispiay/Ci3			Weighted & Undirected >	
			Unweighted & Directed >	
			Weighted & Directed >	
ad				

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 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...**

1919 Data Manager				
 CSV file: C:\Users\mginda\AppData\Local\Temp\t Extracted Co-Authorship Network 				
Graph and Network Analysis Log				
Author inform	Save			
	View			
	View With			
	Rename			
	Discard			
Sci2: Detecting Duplicate Nodes in Sci2

File	Data	Preparation	Preprocessing	Analysis	Modeling	Visualization	n R	Help		
۵		Convert to G	eneric Publicatior	n				Ì	-	
Attri		Remove ISI D	uplicate Records							^
Nun		Remove Row	s with Multitudin	ious Fields						
		Extract Direct	ted Network							
Dete		Extract Bipar	tite Network							
Impl		Extract Paper	Citation Network	k						
Integ Doci		Extract Authority	or Paper Network							
		Extract Co-O	ccurrence Netwo	rk						
Mer		Extract Word	Co-Occurrence N	Vetwork						
Attri		Extract Co-A	uthor Network							
Crea		Extract Refer	ence Co-Occurrer	nce (Biblio	graphic Cou	pling) Netwo	rk			
 Dete		Extract Docu	ment Co-Citation	Network						
Auth		Detect Dupli	cate Nodes							
Integ		L 🔳 Detec	t Duplicate Nod	les					\times	
	inen		Detects duplica	ate nodes	by compar	ring their att	ribute	5		>
		Attribute	to compare on	la	ibel			~	0	
		Merge w	hen this similar	0	.999				0	
		Create no	otice when this s	similar 🛛	.85				0	
		Number	of shared first le	etters 2					0	
							Oł	< (Cancel	

Cyberinfrastructure for Network Science Center

> 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



Cyberinfrastructure for

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

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

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

```
Noteworthy nodes that will NOT be merged33075575398
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"
```

Cyberinfrastructure for 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.



Cyberinfrastructure for Network Science Center Sci2: Updating a merge list

x	5	· @ · =					based o	n label7724	4149648	50423875.c	sv - Excel	
F	ILE HO	OME INSERT PAGE LAYOU	FORMUL	AS DATA	REVIEW	VIEW	ACROBAT					
Fre	om From tess Web	From From Other Text Sources *	Refresh All -	Connections Properties dit Links	ti AAZ AZ AZ Sort	Filter	Clear Reapply Advanced	Text to Columns	Flash Fill	Remove Duplicates	Data Validation =	Consolidate A
		Get External Data	Conne	ctions			ler				Data R	1015
A	2	\cdot : $\times \checkmark f_x$	Finelli, Ly	n								
	Α	В	С	D	E	F	G	н	1	J	К	L
1	label	number_of_authored_works	times_cited	uniqueInc cor	nbineValu	es						
2	Finelli, Lyı	9	1389	1 *								
3	Rolfes, Me	1	10	2 *								
4	Blanton, L	4	48	3 *								
5	Brammer,	4	48	4 *								
6	Smith, So	2	17	5 *								
7	Mustaquir	4	48	6 *								
8	Steffens,	4	48	7*								
9	Cohen, Je	2	15	8*								
10	Leon, Micl	3	41	9*								
11	Chaves, Sa	5	365	10 *								
12	Abd Elai, /	3	30	11 *								
13	Gubareva,	5	307	12 *								
14	Hall, Henr	2	17	14 *								
10	Villopuou	4	40	24 Card								2 ~
10	Vilianuevo Vii Vivon	4	40	5011								. ^
18	Brosoo In	4	1711	* <u>A</u> dd Level	× Delet	e Level	Copy Level		Optio	ons	My dat	a has <u>h</u> eaders
19	Cox, Nanc	10	1995	Caluma			- 5++0+			Order		
20	Kunin, Hil	10		Sort by			Valuer			A to 7		
21	Ridpath, A	1	-	Ishal			values		\sim	A 10 Z		~
22	Driver, Cv	1	5	numi	per_of_autho	ored_wor	ks					
23	Nolan, Mi	1	5	times	_cited ueIndex							
24	Karpati, A	1	5	comb	ineValues							
25	Kass, Dani	1	5									
26	Paone, De	1	5									
27	Jakubows	1	5	L						_		
28	Hoffman,	1	5								OK	Cancel
29	Nelson, Le	1	5	28 *								
	< →	based on label77244149648	504238	+								: •
REA	DY										AVERA	5E: 398.2777778

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

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

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

In Excel, search for the authors listed by using the Find tool

(Ctrl-F), and copy (Ctrl-C) the name from the list and select (Ctrl-C) **OK.**

Sci2-NoteworthyNodesthatwillmerged.txt - Notepad
File Edit Format View Help
0.9333333373069763 similar:
"Paulozzi, Leonard J."
"Paulozzi, Len J."
0.9298245906829834 similar:
"Abd Elal, Anwar Isa"
"Abd Elal, Anwar"
0.9215686321258545 similar:
"Srinivasan, Arjun"
"Srinivasan, A"
0.9215686321258545 similar:
"Kleinman, Ken"
Kleinman, Kenneth
U.920034925305446 similar:
Krieg, Edward F., Jr.
0 9166666865348816 cimilan
"Malarcher Ann M"
"Malarcher A"
0.9136905074119568 similar:
"Hadler. James L."
"Hadler, Jim L."
0.9037036895751953 similar:
"Harriman, Kathleen"
"Harriman, Kathy"
0.9019607901573181 similar:
"Trosclair, Angela"
"Trosclair, A"
0 895833313/65118/ similar

Sci2: Updating a merge list

Cyberinfrastructure for

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

1	label	number_of_authored	_works	times_cited	uniqueInd	combineV	alue:
-	Finalli I m		0	1000	4	*	

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 Nangy	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	*

Cyberinfrastructure for Network Science Center 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



CNS Cyberinfrastructure for Network Science Center Sci2: Updating the co-author network

😵 Select Files				×	
\leftarrow \rightarrow \checkmark \uparrow \blacksquare « Desi	ktop > Sci2-CDC-data > Networks > Results	√ Č	Search Results	م	
Organize 👻 New folder			: = = = : = = =	- 🔳 🕐	
CneDrive	Name	Date modified	Туре	Size	
This PC	🚯 CDC-WoS-Coauthor-MergeList.csv	6/10/2016 11:45 AM	Microsoft Excel C	67 KB	To update the co-author network
Desktop	CDC-WoS-Articles-topcited-processed-d	6/10/2016 10:42 AM	Microsoft Excel C	178 KB	with the duplicate node merges,
Documents	Sci2-NoteworthyNodesthatwillmerged.txt	6/10/2016 11:43 AM	Text Document	2 KB	first load the Merge list as the
Downloads					Standard CSV format.
b Music					
Pictures					
Videos					
🏪 Local Disk (C:)					
KINGSTON (D:)					
🛫 mginda (\\smb.)					
🛖 projects (\\smb.					
KINGSTON (D:)	Load				×
File nar	ne: C Phe file 'C:\Users\mginda\Desktop\S Please select the format you would li	Sci2-CDC-data\Network: ike to try.	s\Results\CDC-WoS-Co	author-MergeList.c	csv' can be loaded using one or more of the following formats.
	Standard csv format NSF csv format Scopus csv format				
					Select Cancel Details >>

Sci2: Updating the co-author network

😵 Sci2 T	[ool				_	\Box ×			
File Dat	a Preparation Preprocessing Analysis Modeling Visualization R Help								
	Convert to Generic Publication		1010 Data Manag	jer					
Men	Remove ISI Duplicate Records	^	V CSV file	: C:\Users\maind	a\AppData\Local\Temp\temp\Prepro	cessed-CDC-V			
Attri	Remove Rows with Multitudinous Fields		V 🖏 Ext	acted Co-Authors	hip Network				
Nun	Extract Directed Network			Graph and Netwo	rk Analysis Log				
	Extract Binartite Network			Merge Table: base	ed on label.3				
Dete	Extract Dipartice Network			Text Log: Nodes t	hat will be merged.3				
Autr				hor information	rthy hodes that will NOT be merged.3				
Integ	Extract Author Paper Network		CSV file	: C:\Users\mgind	a\AppData\Local\Temp\temp\Prepro	cessed-CDC-V			
Doci	Extract Co-Occurrence Network								
Inpu	Extract Word Co-Occurrence Network								
Mer	Extract Co-Author Network								
Nun	Extract Reference Co-Occurrence (Bibliographic Coupling) Network			Update Network by M	erging Nodes				
Crea	Extract Document Co-Citation Network					Update a network by m	erging nodes		
Loac	Detect Duplicate Nodes		Ag	gregation Function File	C:/Users/mginda/Desktop/Sci2/sci2-1.2b_2015	i1218-win32.win32.x86/sc	i2/		
Doci	Undate Network by Merging Noder								
C:\Usersy	mgmua/Desktop/Sci2-CDC-data/ivetworks/Nesuits/CDC-wos-Coauthor-ivierge	List.csv	0						_
		×	< 57	Select a File				×	
			<i>~</i>	→ ~ ↑ <mark>-</mark> « s	sci2 > sampledata > scientometrics > prope	rties 🗸 Č	Search properties	م	
			O	ganize 👻 New fol	der			- 🔳 🔞	
Colo	ating both the original patwork and the l	aadad		OneDrive	Name	Date modified	Туре	Size ^	
Sele	cting both the original network and the i	oaueu		This PC	bibtexCoAuthorship.properties	1/4/2016 2:09 PM	PROPERTIES File	1 KB	
mer	ge list CSV, navigate to Data Preparation	> Upda	te	Desktop	endnoteCoAuthorship.properties	1/4/2016 2:09 PM	PROPERTIES File	1 KB	
Nati	vark by Marging Nadaa	•		Documents	isiCoAuthorship.properties	1/4/2016 2:09 PM 1/4/2016 2:09 PM	PROPERTIES File	1 KB	
neu	work by Merging Nodes.			Downloads	isiPaperCitation.properties	1/4/2016 2:09 PM	PROPERTIES File	1 KB	
				b Music	mergeBibtexAuthors.properties	1/4/2016 2:09 PM	PROPERTIES File	1 KB	
Dress	use for the excrements function file in the	c: ว		E Pictures	mergeEndnoteAuthors.properties	1/4/2016 2:09 PM	PROPERTIES File	1 KB	
BLOA	vse for the aggregate function file in the	SCIZ		Videos	mergelsiAuthors.properties	1/4/2016 2:09 PM 1/4/2016 2:09 PM	PROPERTIES File	1 KB	
sam	ple data directory. Load the			🏪 Local Disk (C:)	mergeNsfPls.properties	1/4/2016 2:09 PM	PROPERTIES File	1 KB	
()				KINGSTON (D:)	mergeNsfPls-pre2014.properties	1/4/2016 2:09 PM	PROPERTIES File	1 KB	
"me	rgeisiAuthors.properties" file, and then	select		🛖 mginda (\\smb.)	mergeScopusAuthors.properties	1/4/2016 2:09 PM	PROPERTIES File	1 KB	
OK.				projects (\\smb.	nsfCoPI.properties	1/4/2016 2:09 PM	PROPERTIES File	1 KB 🗸	
	Update Network by Merging Nodes			KINGSTON (Dr)	X			>	
	Undate a network by	v meraina nodes			erties		×.*	~	
		,					Open	Cancel	
	Aggregation Function File C:/Users/mginda/Desktop/Sci2/sci2-1.2b_20151218-win32.win32.x86	j/sci2/sampledata/	cientometrics/prope	ties/mergelsiAuthors.p	roperties Browse				

OK Cancel

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

Preprocessing Analysis	Modeling Visualization R Help			Preprocessing Analysis	Modeling Visualization R Help
General >	- 8	រពិរំ Data Manager	-	General >	
Temporal >	^	Extract Edges Above or Below Value	CC	Terreneral	
Geospatial >				iemporai /	^
Topical >		Extract all edges with an attribute strictly above a certain number		, Geospatial >	
Networks >	Extract Top Nodes	Extract from this number 1.0		Topical >	ates)
connected	Extract Nodes Above or Below Value			Networks >	Extract Top Nodes
nected components. (4 isc	Delete Isolates	Below?		(initia): 0.0067	Extract Nodes Above or Below Value
omponent consists of 170 connectedness because the	Extract Top Edges	Numeric Attribute number of coauthored works		Numeric Attribute	Delete Isolates
	Extract Edges Above or Below Value				Extract Top Edges
eights): 0.0067 Numeric Attribute	Remove Self Loops	OK Cancel -	-		Extract Edges Above or Polow Value
Vullenc Attribute	Trim by Degree		1	Below Value was selected.	Extract Edges Above of Below Value
	MST-Pathfinder Network Scaling			is G. Smith	Remove Self Loops
Below Value was selected. hith	Fast Pathfinder Network Scaling			. Smith	Trim by Degree
			_		MST-Pathfinder Network Scaling

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.

luito Brobu citatino	to be unuri	ecteu.					
Nodes: 410					This graph claims to be	undir	ected.
Isolated nodes: 0					N-1 2101		
Node attributes p	resent:				Nodes: 2191		
					Node attributes present	• 1abo	
Edges: 1215					node accribuces present	. Iave	I, HUMDEI
No self loops were	e discovere	t.			Edges: 16042		
No parallel edges were discovered.			No self loops were disc	overed			
					No parallel edges were	discov	ered.
Edge attributes:							
Did not de Numeric at	etect any no ttributes:	onnumeric	attribu	tes.			
		min	max	me	ean		
n	umber	2	10	2.	.50041		
We	eight	1	10	2.	.43457		
This netwo	ork seems to	o be valu	ed.				
Average degree: 5 This graph is not There are 53 weak	.9268 weakly con ly connected	nected. d compone	nts. (0	isola	ates)		
The largest conne	cted compone	ent consi	sts of 7	4 noo	des.		
Did not calculate	strong con	nectednes	s becaus	e thi	is graph was not direct	ed.	
Density (discussion	ding unight	-) • 0 01/	5				

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.

E LEXT LOC: IN	odes that will be merded.4
Clustering Coe	fficient
Nearest Neigh	bor Degree
Strength vs De	gree
Degree & Stree	ngth
Average Weigl	nt vs End-point Degree
Strength Distri	bution
Weight Distrib	ution
Randomize We	eights
Node Between	ness Centrality
Blondel Comm	nunity Detection
Louvain Comr	nunity Detection (with resolution parameter)
Louvain Multil	evel Refinement Community Detection
SLM Commun	ity Detection
HITS	



Cyberinfrastructure for Network Science Center

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.

	Analysis Mode	eling Vis	ualization	R	Help					
-	Temporal	>					1010 0101	Data I	Mana	
t	Geospatial	>				^	~		SV fil	
	Topical	>						~ 5	🔪 Ext	
	Networks	>	Netv	vork A	Analysis Too	olkit (NAT)			N
	lang		Unw	eighte	ed & Undire	ected		>		1
(2002). Statistical mechanic			Weig	ghted	& Undirect	ted		>		I
Unweighted & Di				ed & Direct	ed		>		ł	
ľ	uispiay/ cisricce/	Degree	Weig	ghted	& Directed			>		١
										۱
										Ì
ſ	eam									ć
										1
										'
D	CIShell-Session-2670277137484061540\StaticExecutableRu							١		
ľ	vi>wgnupiot.exe -ioad usegnupiot.txt -persist							(
							0	Work		E
-										1
le	leted automatically Remove all completed								F	
										1
										1
	Date	Time		%	Complete					0
	06/10/2016	11:55:54	AM							
	06/10/2016	11:55:13	AM							ł
		11.64.77	0.0.4				10 C			

Cyberinfrastructure for Network Science Center

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

No	Node Degree					
De	Degree Distribution					
K-N	Nearest Neighbor					
Wa	atts-Strogatz Clustering Coefficient					
Wa	atts Strogatz Clustering Coefficient over K					
Dia	ameter					
Ave	erage Shortest Path					
Sho	ortest Path Distribution					
No	ode Betweenness Centrality					
We	eak Component Clustering					
Glo	obal Connected Components					
Ext	tract K-Core					
An	inotate K-Coreness					
Blo	ondel Community Detection					
Lou	uvain Community Detection (with resolution parameter)					
Lou	uvain Multilevel Refinement Community Detection					
SLN	M Community Detection					
HI	TS					



Questions?

Sci2: Visualizing Co-Author Network in Gephi



Cyberinfrastructure for Network Science Center

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

CNS Cyberinfrastructure for Network Science Center



Gephi: Initial Layout of Network

Cyberinfrastructure for Network Science Center



196

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

Cyberinfrastructure for Network Science Center



Gephi: Showing and Updating Node Labels



Cyberinfrastructure for Network Science Center

Cancel

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.



Cyberinfrastructure for Network Science Center 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.

CNS Cyberinfrastructure for Network Science Center Gephi: Filtering Networks



CNS Cyberinfrastructure for Network Science Center Gephi: Final visualization



Cyberinfrastructure for Network Science Center Gephi: Final visualization – Node Parameters



CNS Cyberinfrastructure for Network Science Center Gephi: Final visualization – Edge Color Parameters

- Edges		Dis de dell'activent i	Channel Calar		
Show Edges		26 Preview Settings - Color	× Choose a Color		
Thickness	15.0	Edge Color			
Rescale weight	\checkmark	Configures the color of the edges. Edges can either a color on their own			
Color	custom [92, 148, 83]	(original) or uses incident nodes color.			
Opacity	100.0				
Curved	\checkmark	Original			
Radius	0.0				
Edge Arrows		Mixed			Hue: 117 C
Size	3.0	O Source			
Edge Labels		0			Sat: 44 🖨 🔿
Show Labels		○ Target			Brit 69 📥 @
Font	Arial 10 Plain	Custom		0	
Color	original				Red: 109 - C
Shorten label					
Max characters	30	ОК	Cancel		Green: 176 🜩 🔿
Outline size	0.0				Blue: 00 🛋 C
Outline color	custom [255,255,255]		tenary		
Outline opacity	80.0		HI WINK SA		How EDR062
			-16		- HEX: ODDU03
Provine ratio: 100%					Cancel OK

Preview ratio: 100%		
	🍘 Refresh	Background Reset zoom - +
Export: SVG/PDF/PNG		

Color	original						
Opacity	100.0	🖆 Export					×
Curved	\checkmark	Savein	- Peculta		~		_
Radius	0.0	Save In:	Results		~		=
∃Edge Arrows		0	— h				
Size	3.0						
Edge Labels		Recent Items					
Show Labels							
Font	Arial 10 Plain						
Color	original	Deckter					
Shorten label		Desktop					
Max characters	30						
Outline size	0.0						
Outline color	custom [255,255,255]	. Documents					
Outline opacity	80.0	This PC					
Preview ratio: 100%		9	File name:	·			Save
	🗞 Refresh	Network	Files of type:	PDF Files (*.pdf)		~	Cancel
Export: SVG/PDF/PNG							Options



Questions?