Patent Network Visualisation with Gephi

Paul Oldham
One World Analytics & IAS-UNU

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

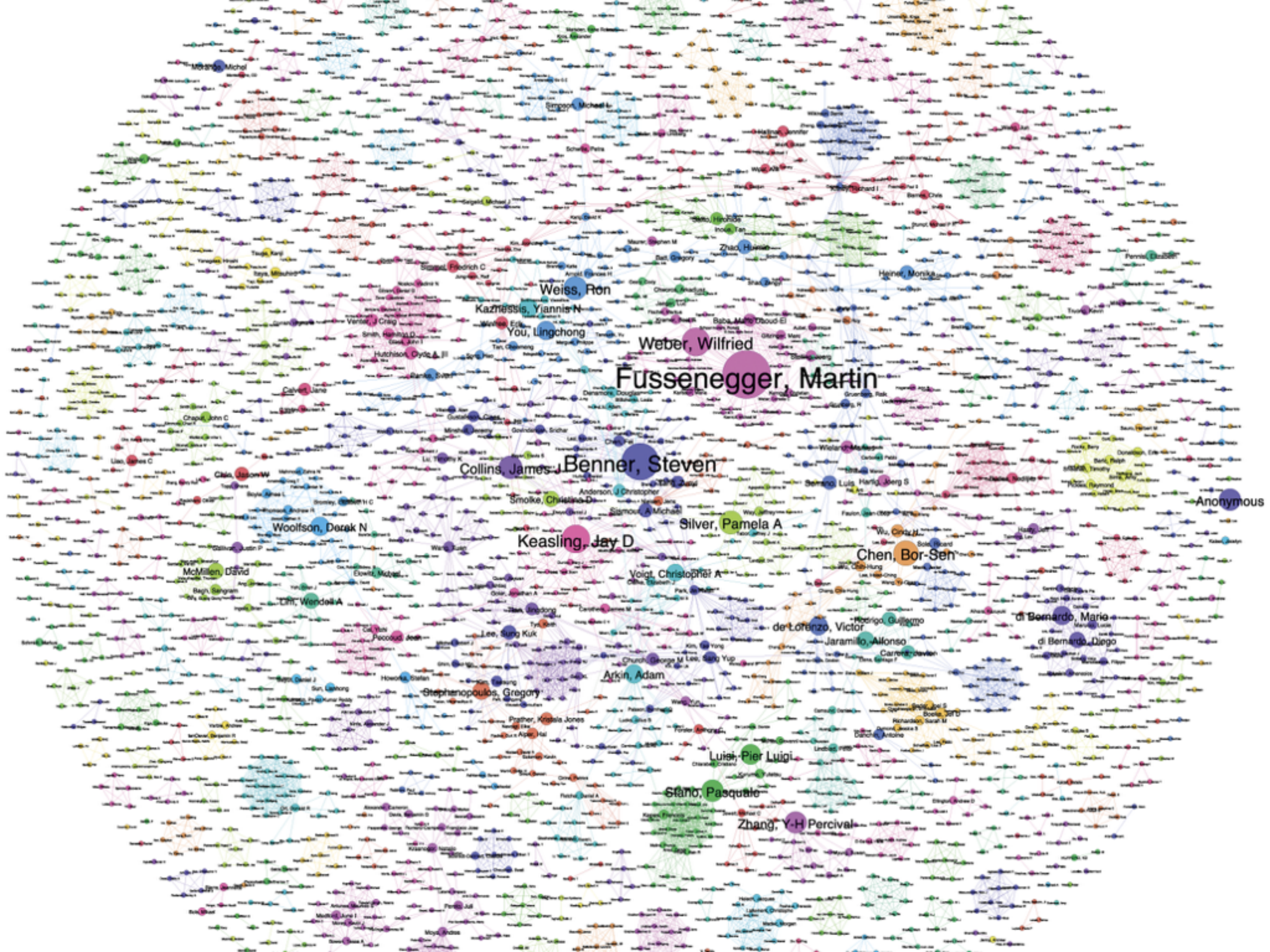
Synthetic Biology: Mapping the Scientific Landscape

Paul Oldham, Stephen Hall, Geoff Burton

Published: April 23, 2012 • DOI: 10.1371/journal.pone.0034368

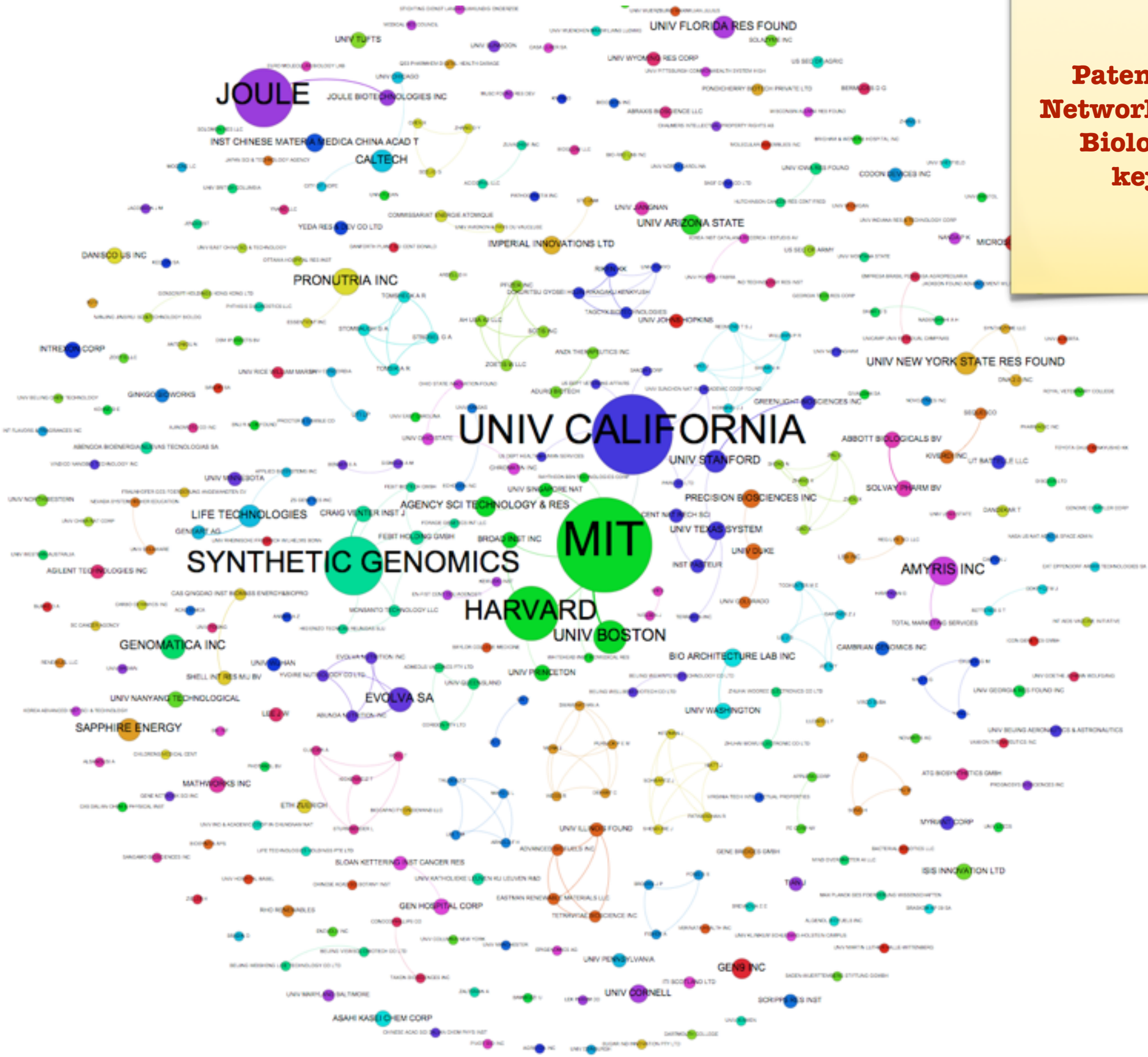
Abstract

This article uses data from Thomson Reuters *Web of Science* to map and analyse the scientific landscape for synthetic biology. The article draws on recent advances in data visualisation and analytics with the aim of informing upcoming international policy debates on the governance of synthetic biology by the Subsidiary Body on Scientific, Technical and Technological Advice (SBSTTA) of the United Nations Convention on Biological Diversity. We use mapping techniques to identify how synthetic biology can best be understood and the range of institutions, researchers and funding agencies involved. Debates under the Convention are likely to focus on a possible moratorium on the field release of synthetic organisms, cells or genomes. Based on the empirical evidence we propose that guidance could be provided to funding agencies to respect the letter and spirit of the Convention on Biological Diversity in making research investments. Building on the recommendations of the United States Presidential Commission for the Study of Bioethical Issues we demonstrate that it is possible to promote independent and transparent monitoring of developments in synthetic biology using modern information tools. In particular, public and policy understanding and engagement with



Fussenegger, Martin
Benner, Steven
Weber, Wilfried
Weiss, Ron
Kaznessis, Yiannis N
You, Lingchong
Collins, James J
Keasling, Jay D
Silver, Pamela A
Chen, Bor-Sen
Voigt, Christopher A
Woolfson, Derek N
McMillen, David
Arkin, Adam
de Lorenzo, Victor
Jaramillo, Alfonso
di Bernardo, Mario
di Bernardo, Diego
Anonymous
Luisi, Pier Luigi
Stano, Pasquale
Zhang, Y-H Percival
Stephanopoulos, Gregory
Lee, Sung Kuk
Smolke, Christina D
Anderson, J Christopher
Venter, J Craig
Calvert, Jane
Chaput, John C
Liao, James C
Heiner, Monika
Prather, Kristala Jones
Rodrigo, Guillermo

Patent Applicant Network in Synthetic Biology (simple keywords)



Gephi
makes graphs handy

Download Blog Wiki Forum Support Bug tracker

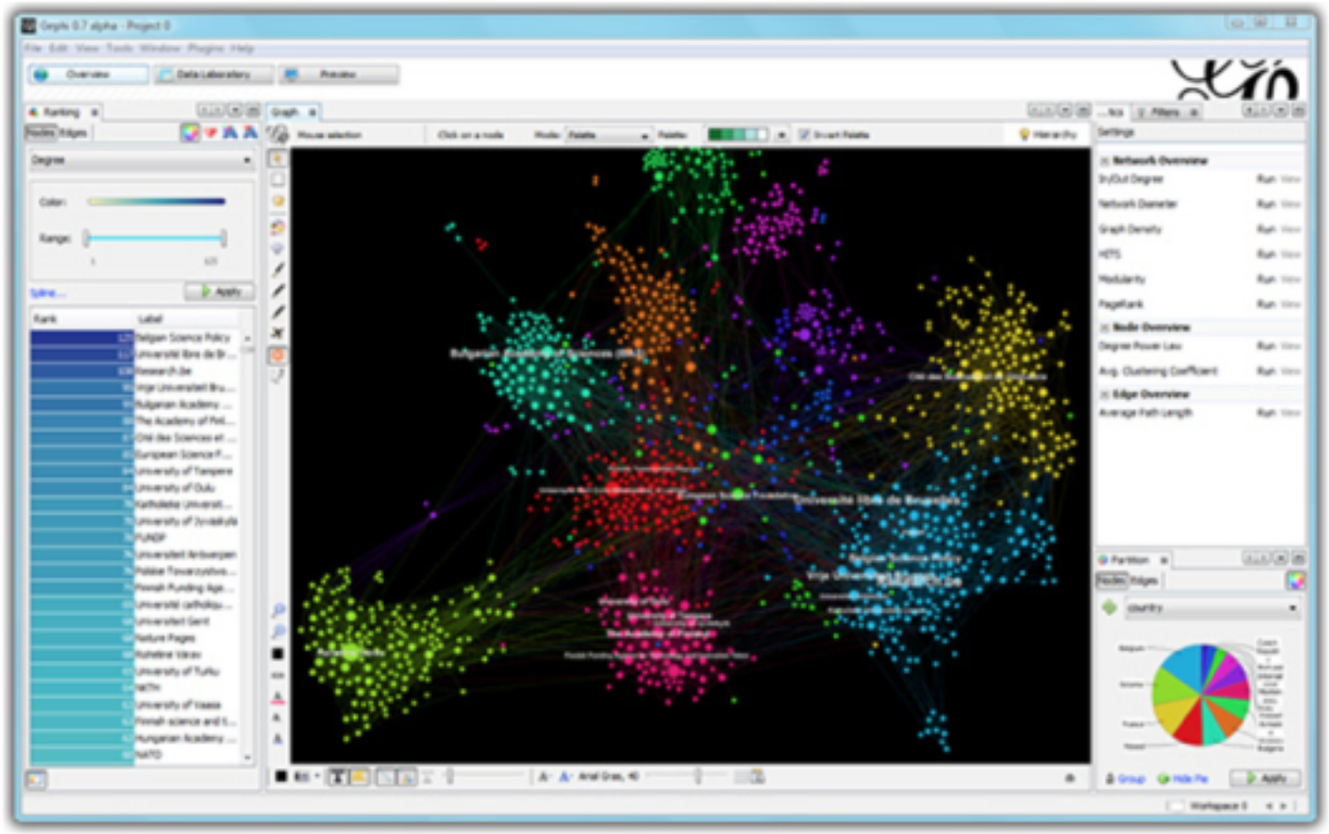
Home Features Learn Develop Plugins Services Consortium

The Open Graph Viz Platform

Gephi is an interactive visualization and exploration platform for all kinds of networks and complex systems, dynamic and hierarchical graphs.

Runs on Windows, Linux and Mac OS X. Gephi is open-source and free.

Learn More on Gephi Platform »



Download FREE
Gephi 0.8.2-beta

Release Notes | System Requirements

- Features
- Quick start
- Screenshots
- Videos

Support us! We are non-profit. Help us to innovate and empower the community by donating only 8€:

Download Gephi (hosted on GitHub)

Follow the installation instructions. There can often be issues with the Java versions. Instructions for Windows are here. Instructions for Mac are here.

Install Java 1.7 & Configure

1. Go to http://www.oracle.com/technetwork/java/javase/downloads/jdk7-downloads-1880260.html and install Java 1.7

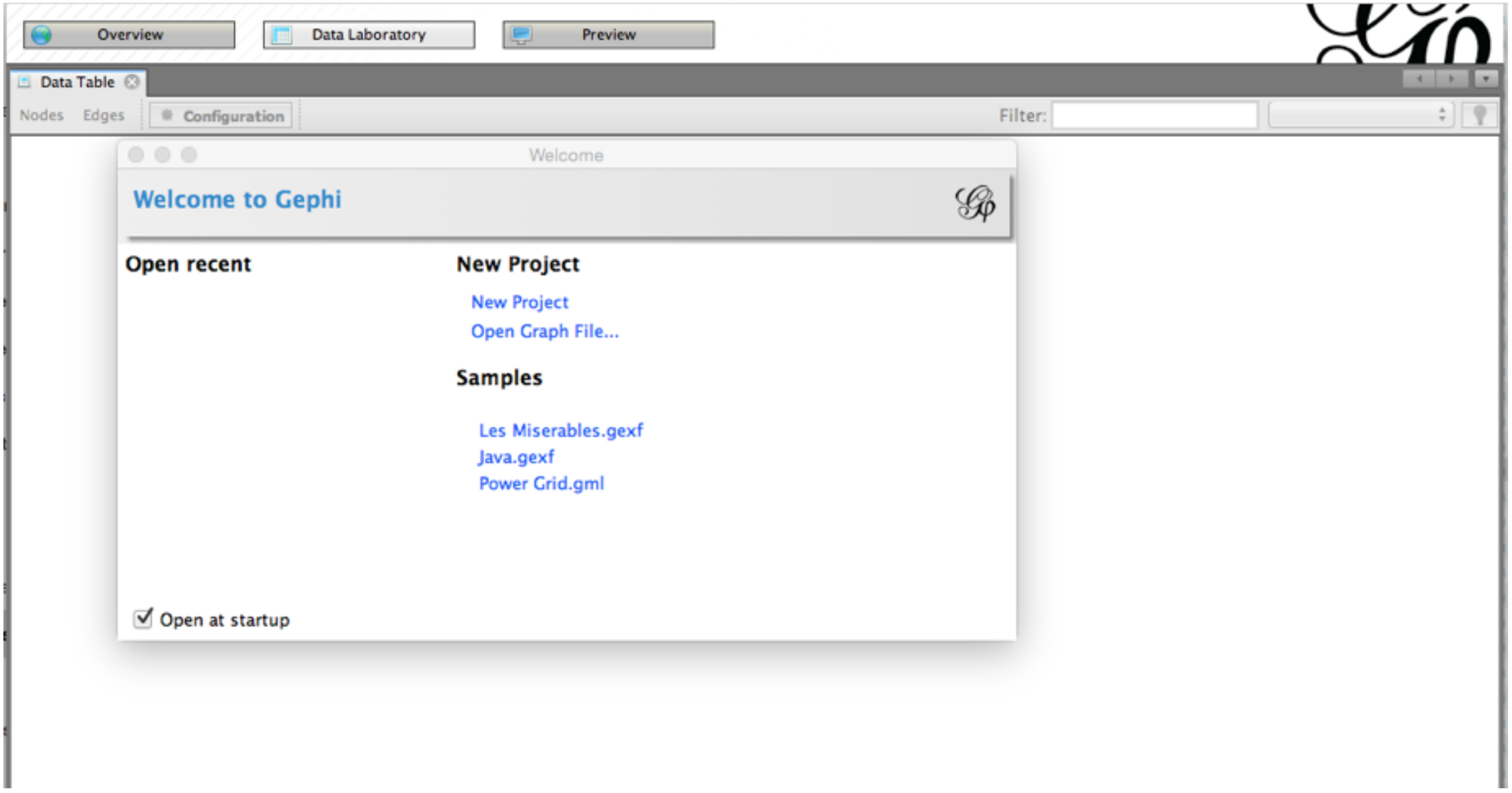
2. Go to your Gephi installation folder (probably C:\Program Files (x86)\Gephi-0.8.2) and locate the folder **etc**

3. Inside **etc** folder you will find a file named **gephi.config**, open this file with notepad.

4. Search for the keyword #jdkhome="/path/to/jdk"

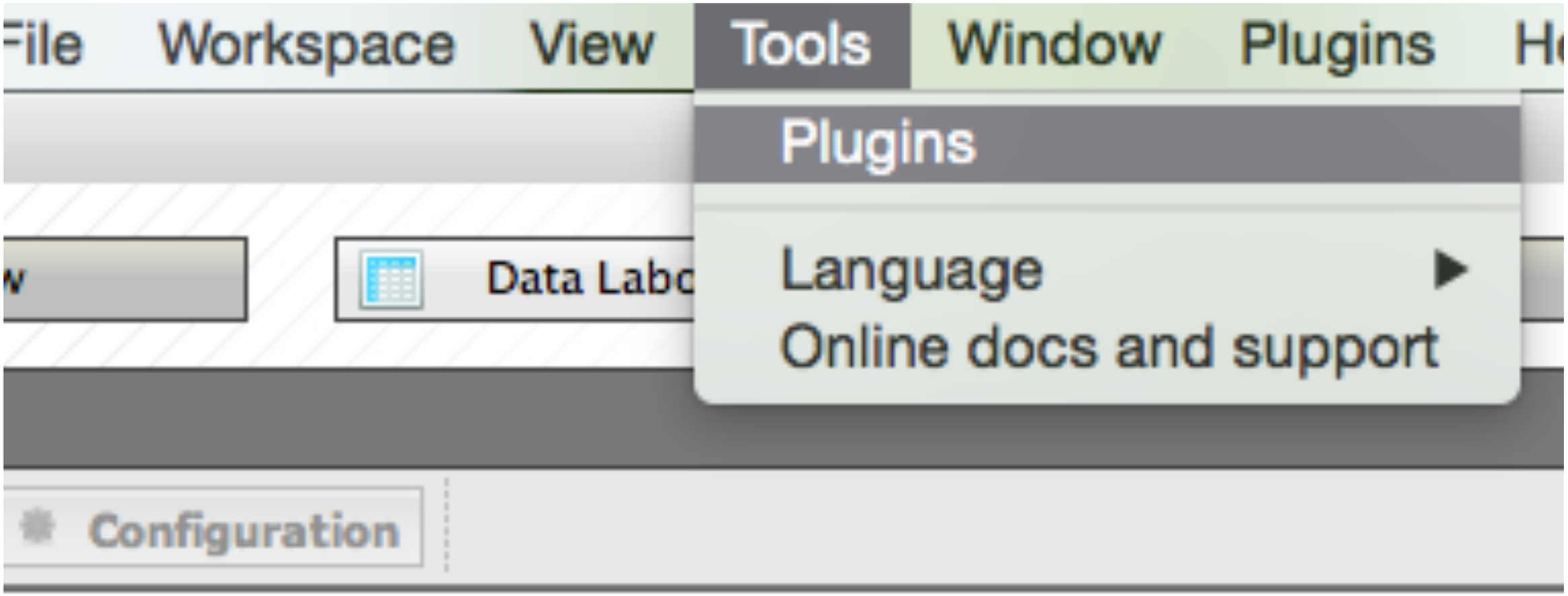
5. Once you find this part of code, remove the # from the beginning since the machine will not execute this code and it will consider it as comment.

Instructions from Gephi user Masoud on the Gephi Forum

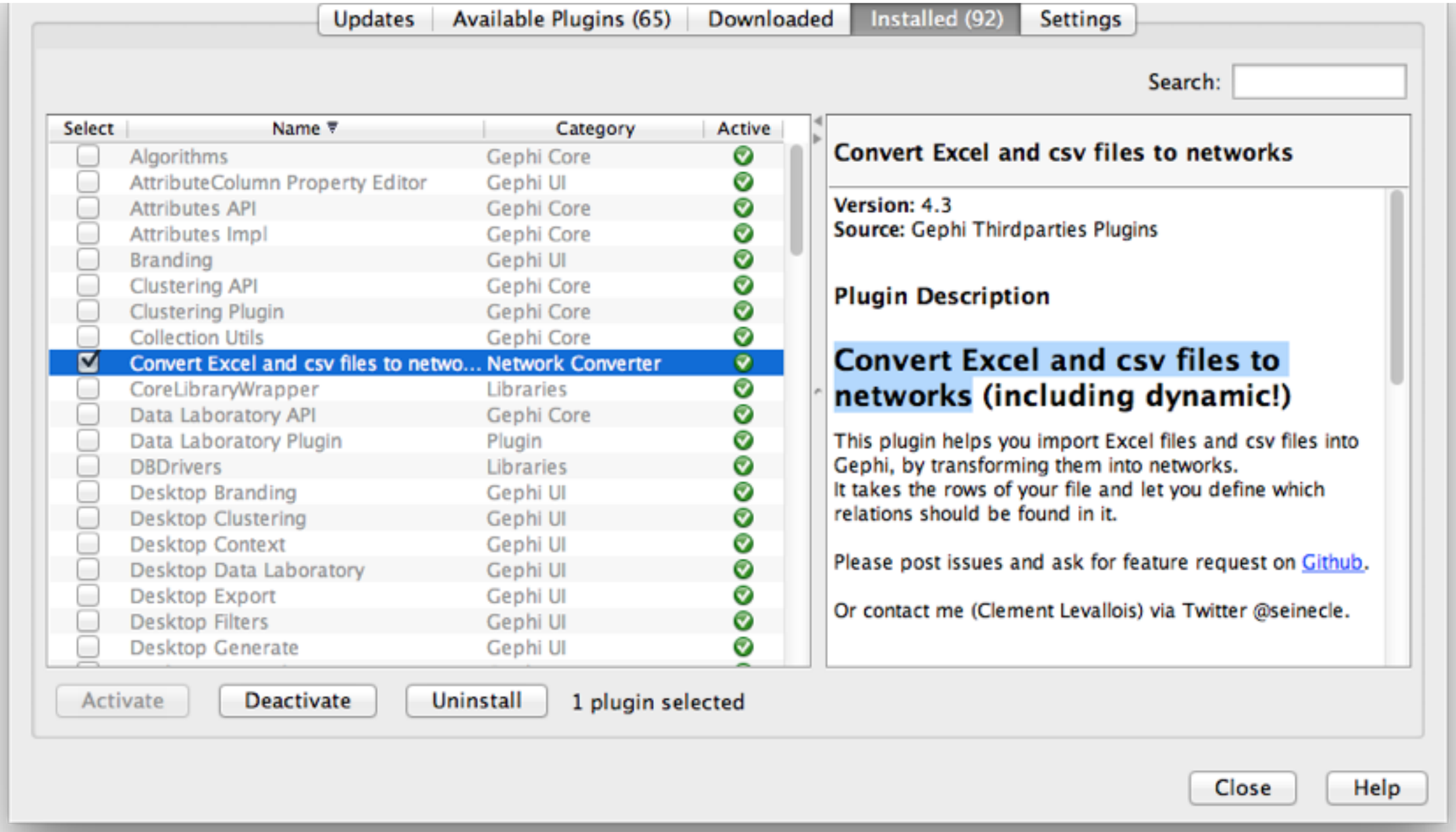


Click on New Project

This will create a blank project (if using import spigot, described below). If you have a .gexf file then choose Open Graph File

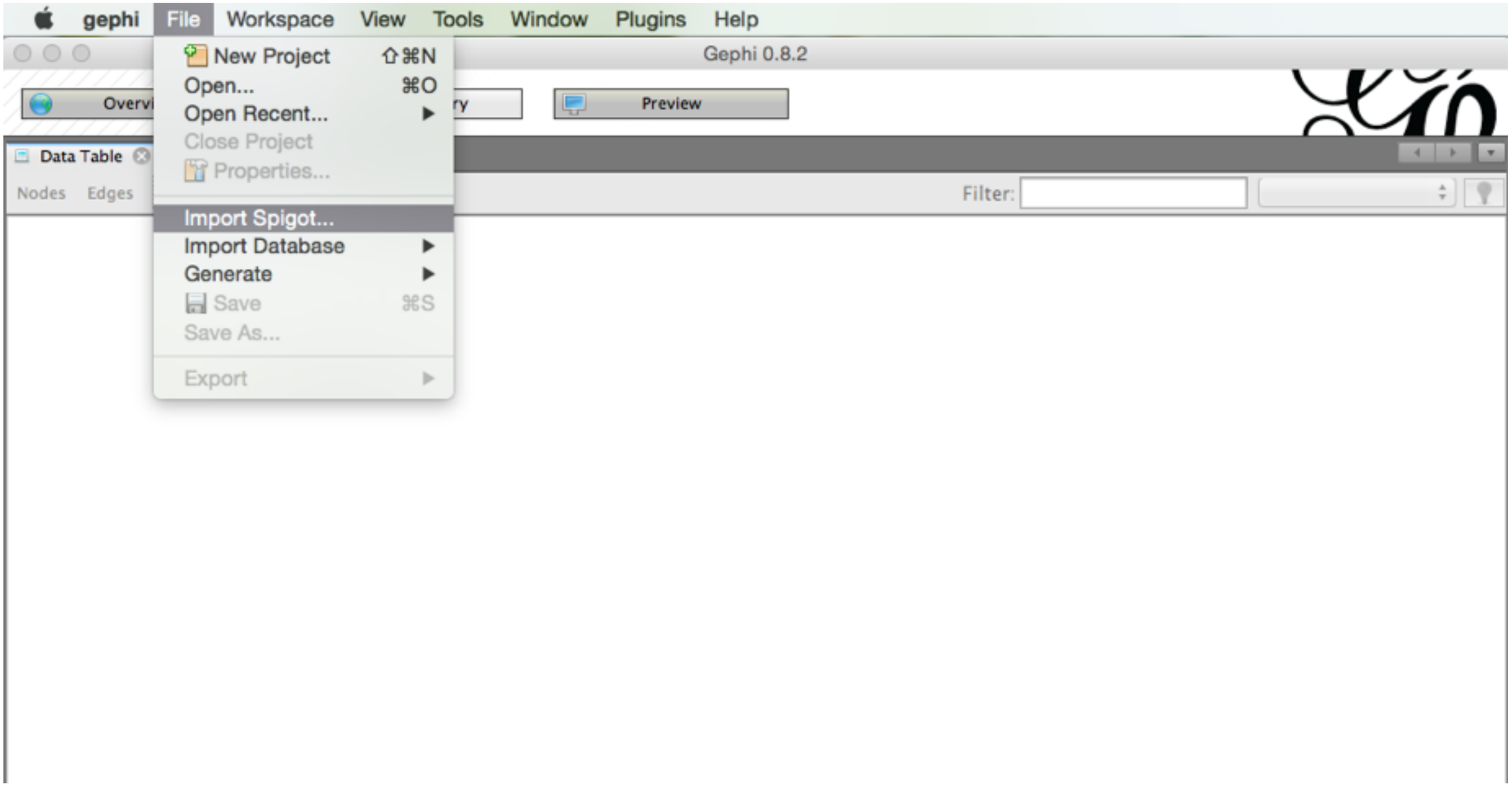


Go to Plugins

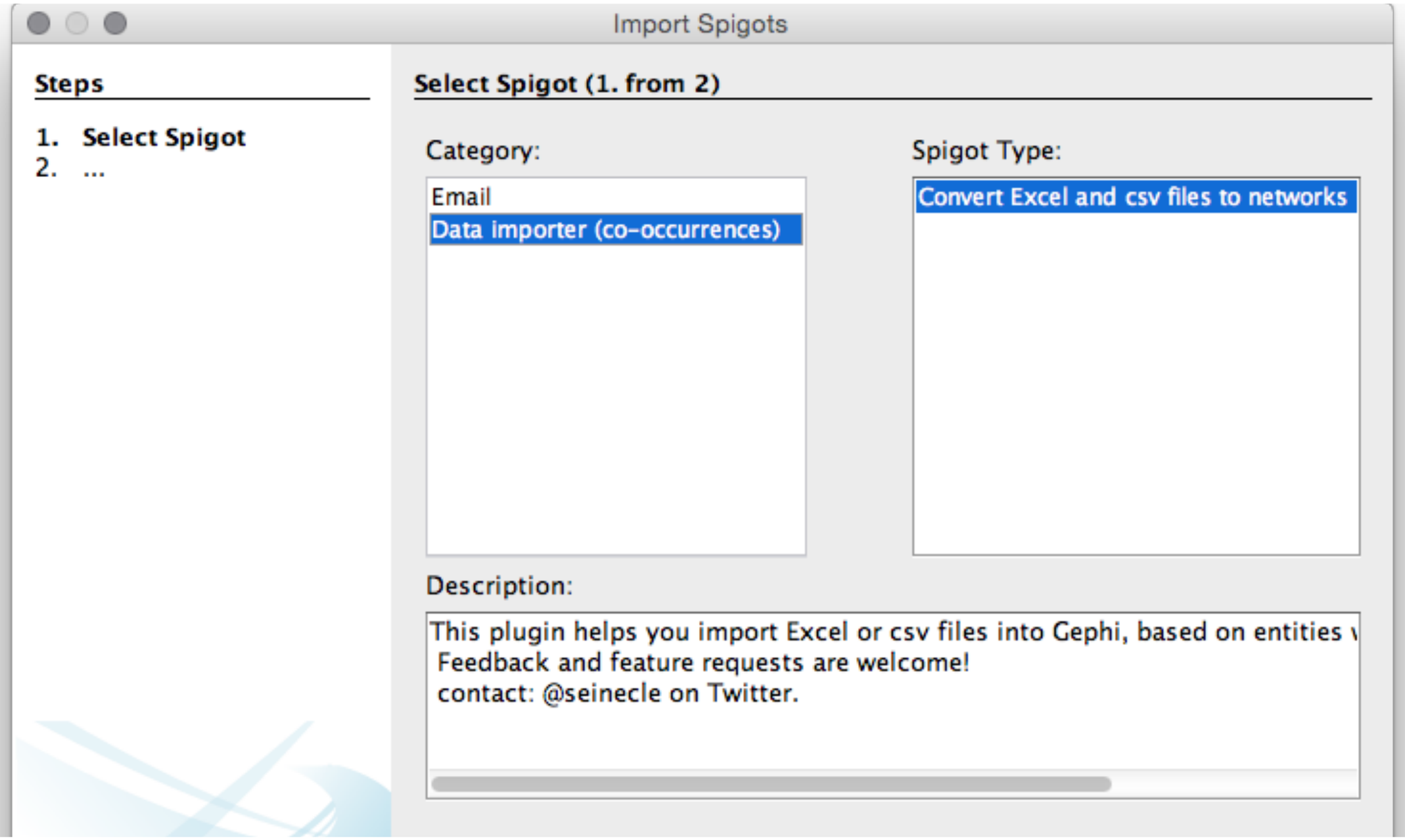


Install Plugins

Look for Convert Excel. Also look for No Overlap and Expansion (if not already installed by default).



From the File menu select Import Spigot



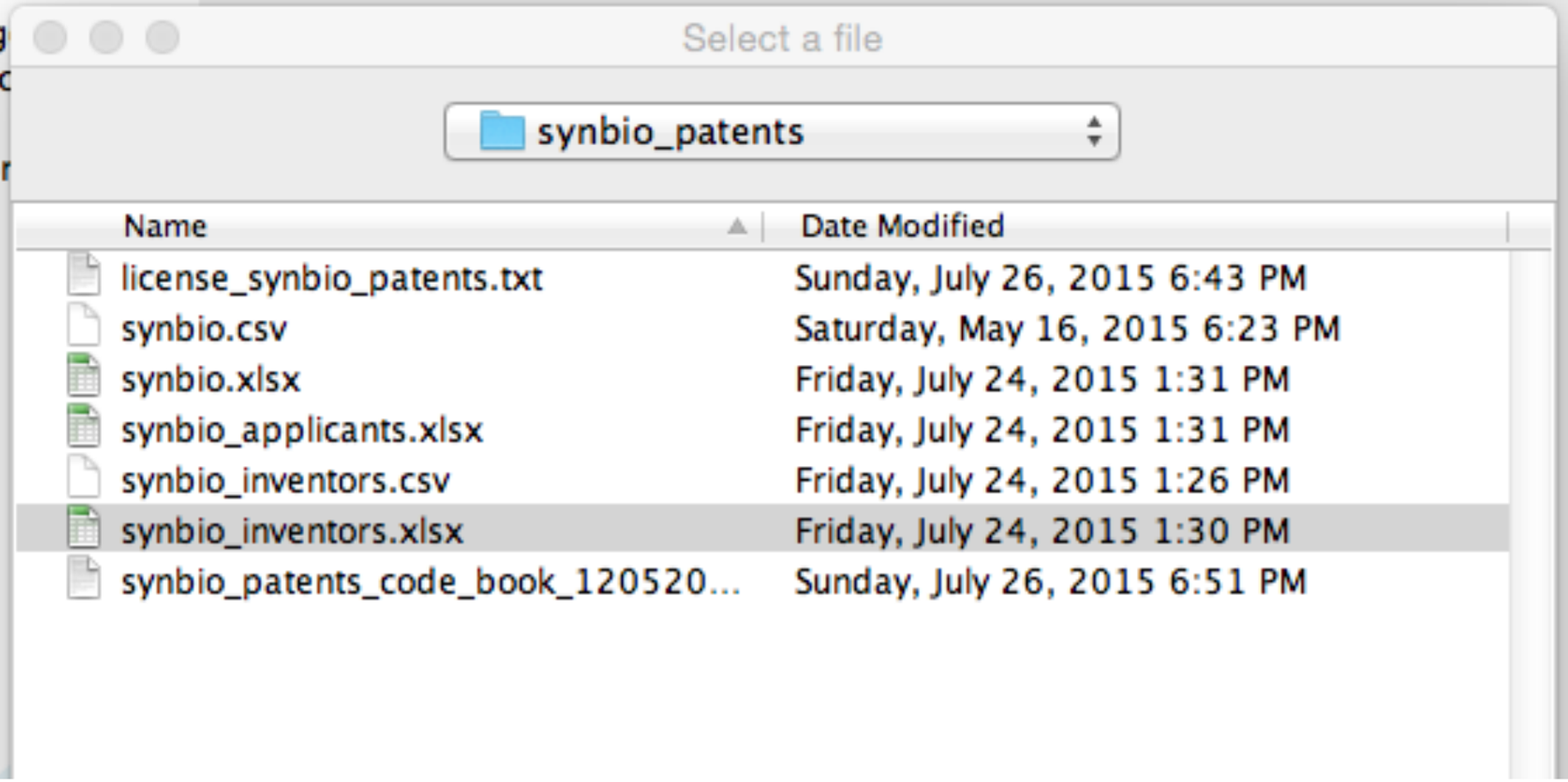
Data Importer

Steps

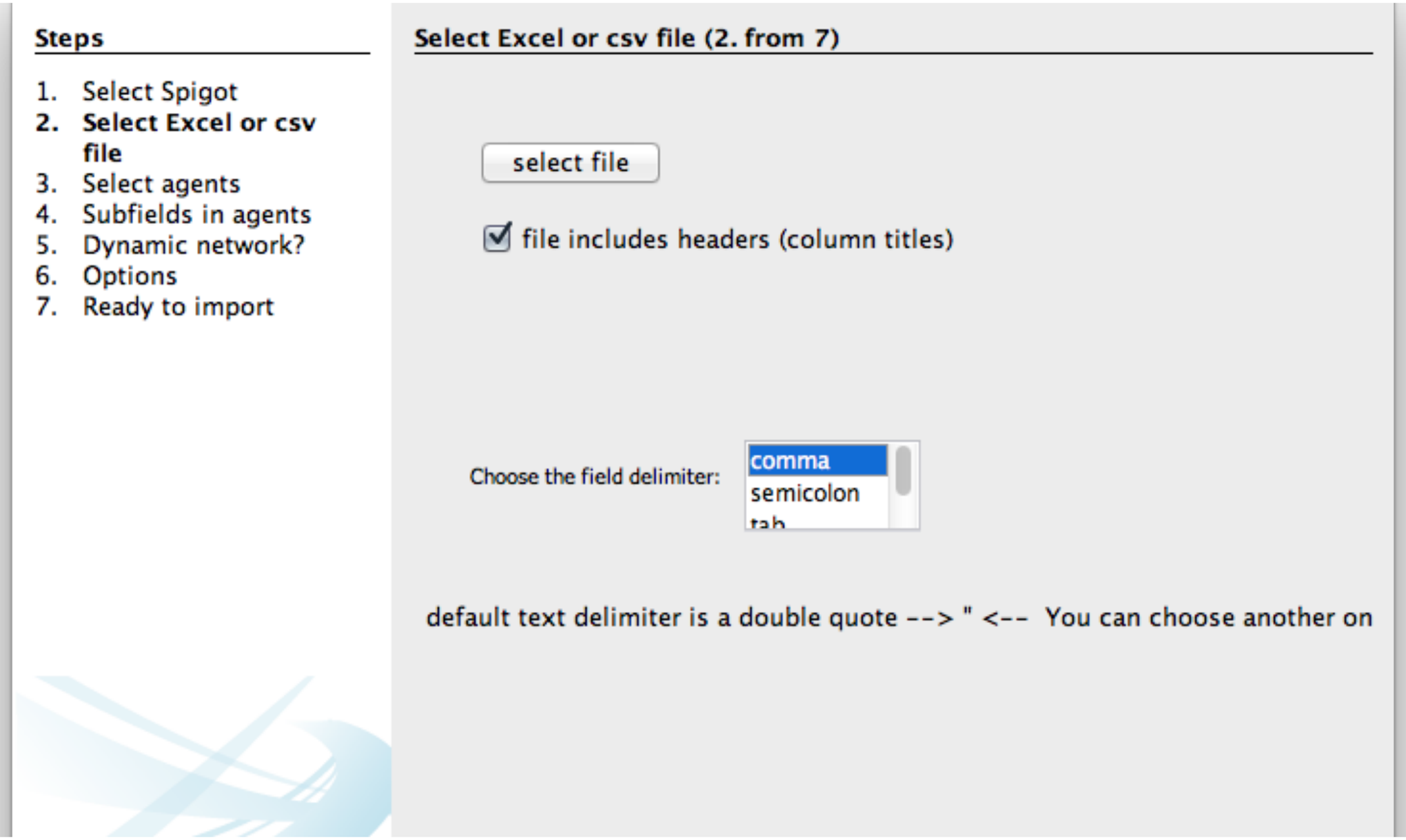
1. Select Spigot
2. **Select Excel or csv file**
3. Select agents
4. Subfields in ag
5. Dynamic netwc
6. Options
7. Ready to impor

Select Excel or csv file (2. from 7)

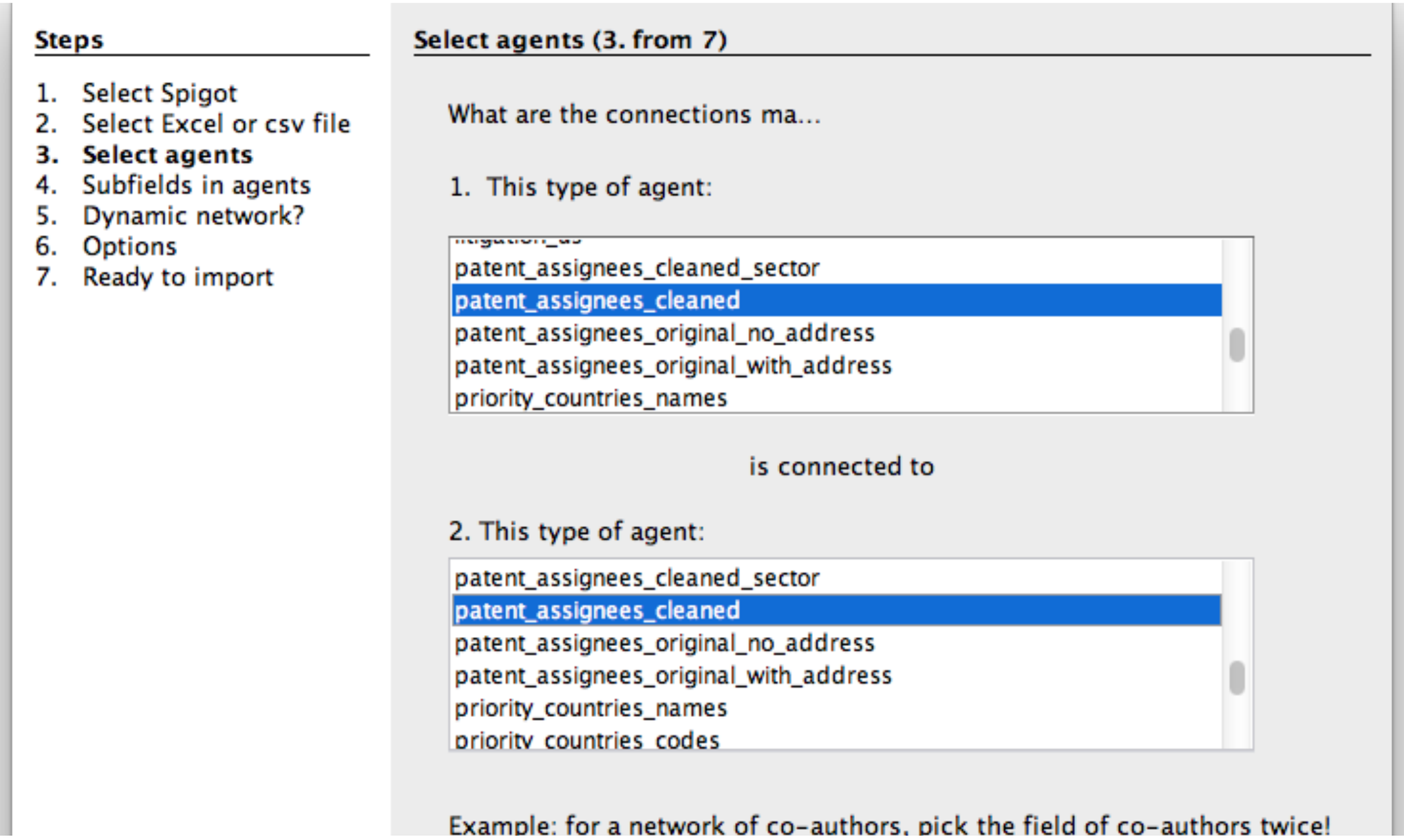
select file



Select the synbio_inventors file



Select comma as the field delimiter



Select patent_assignees_cleaned

For co-occurrence choose the same field in both boxes

Steps

1. Select Spigot
2. Select Excel or csv file
3. Select agents
4. **Subfields in agents**
5. Dynamic network?
6. Options
7. Ready to import

Subfields in agents (4. from 7)

patent_assignees_cleaned

Is this field made of subfields? If so, what is the delimiter?

comma
semicolon
tab
space

Example:
Let's imagine you selected "authors" for agents in the previous screen. Suppose that the field for "authors" is made of several co-authors, separated by semicolon.
=> Please choose "semicolon" as a delimiter.
(leave the selection empty if no delimiter applies).

Help | < Back | Next > | Finish | Cancel

Use semicolon (subfield separator)

Steps

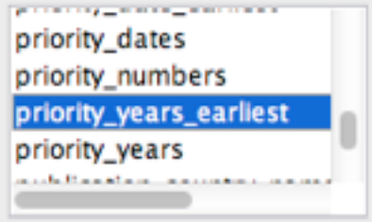
1. Select Spigot
2. Select Excel or csv file
3. Select agents
4. Subfields in agents
5. **Dynamic network?**
6. Options
7. Ready to import

Dynamic network? (5. from 7)

If you want to create a dynamic network, where is the field for time?

Leave blank if the network is not dynamic

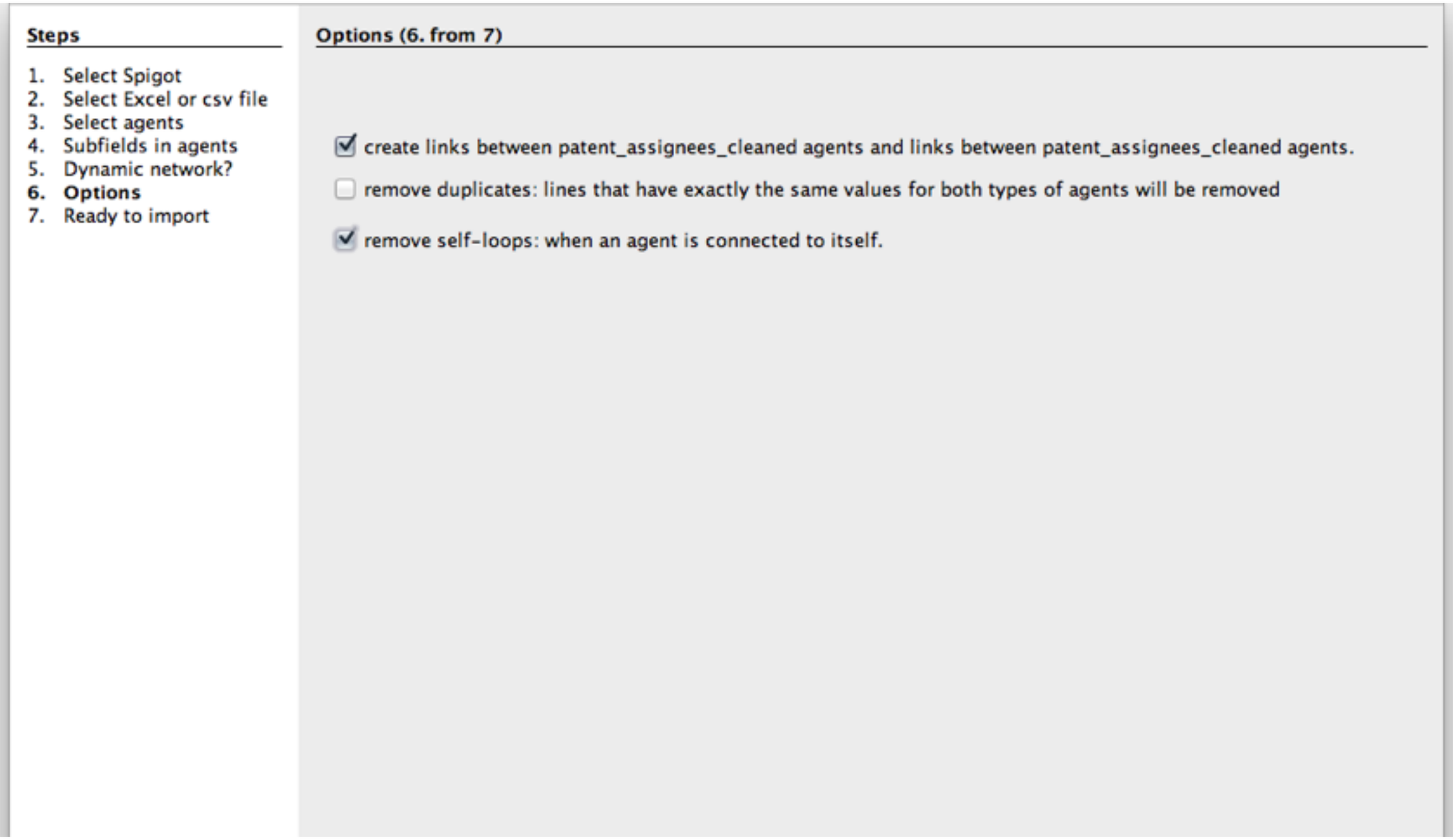
! Works only for Excel files at the moment (not csv!).



Let's imagine you each of your data entry has a time stamp. You could create a network changing in time, where nodes and their connections appear and disappear along a timeline.
2 time formats allowed: 1984, 2014-09-17 (but not 2014-17-09).
Dates (1984) but also durations (1984,1986) are allowed in this field. For duration, use the comma to separate the start and end dates.

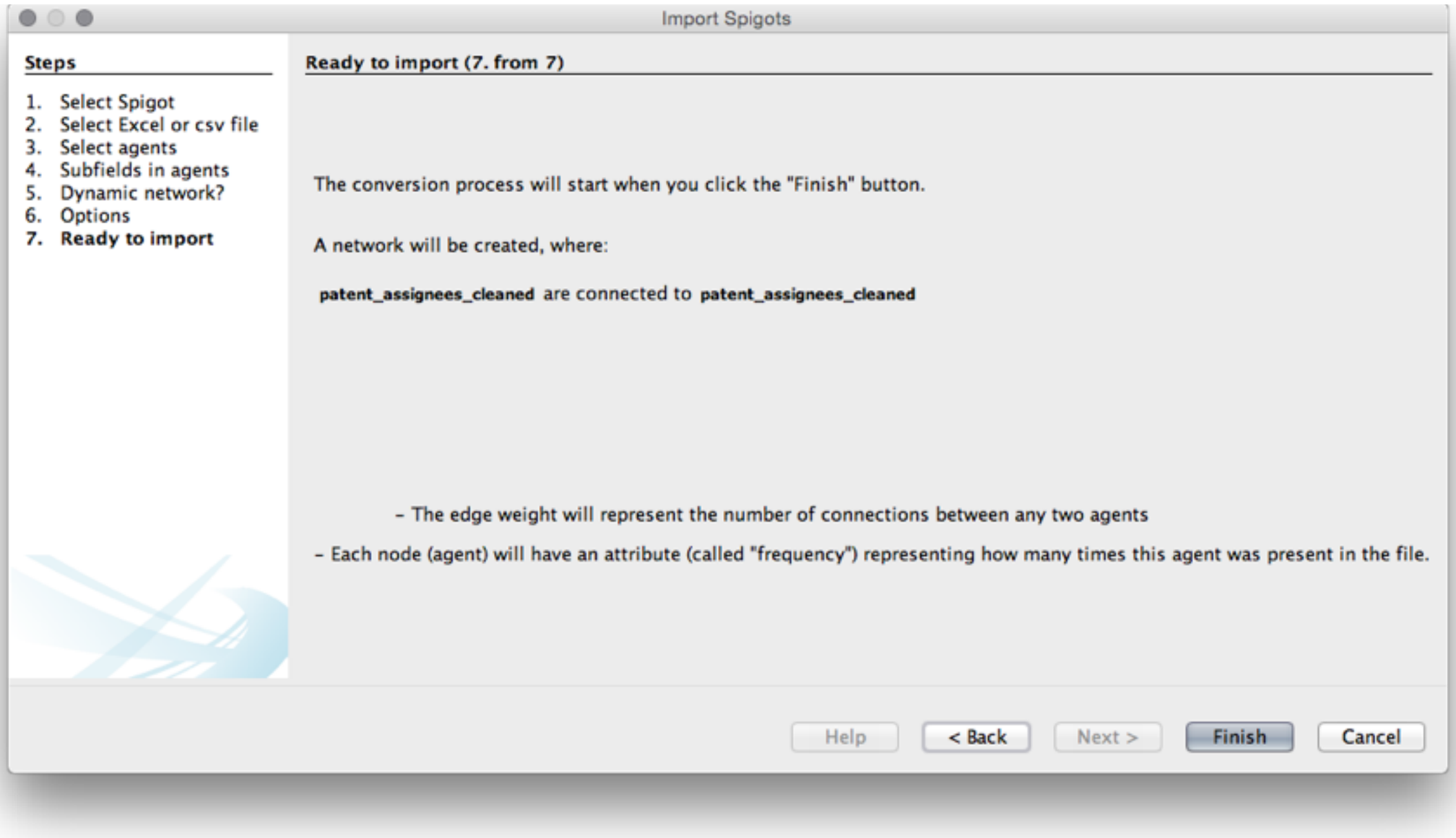
Dynamic network session - priority_year_earliest

Leave blank or choose a year field. Note that gephi can struggle with date formats and so this does not always work correctly (because of problems in the underlying data format).



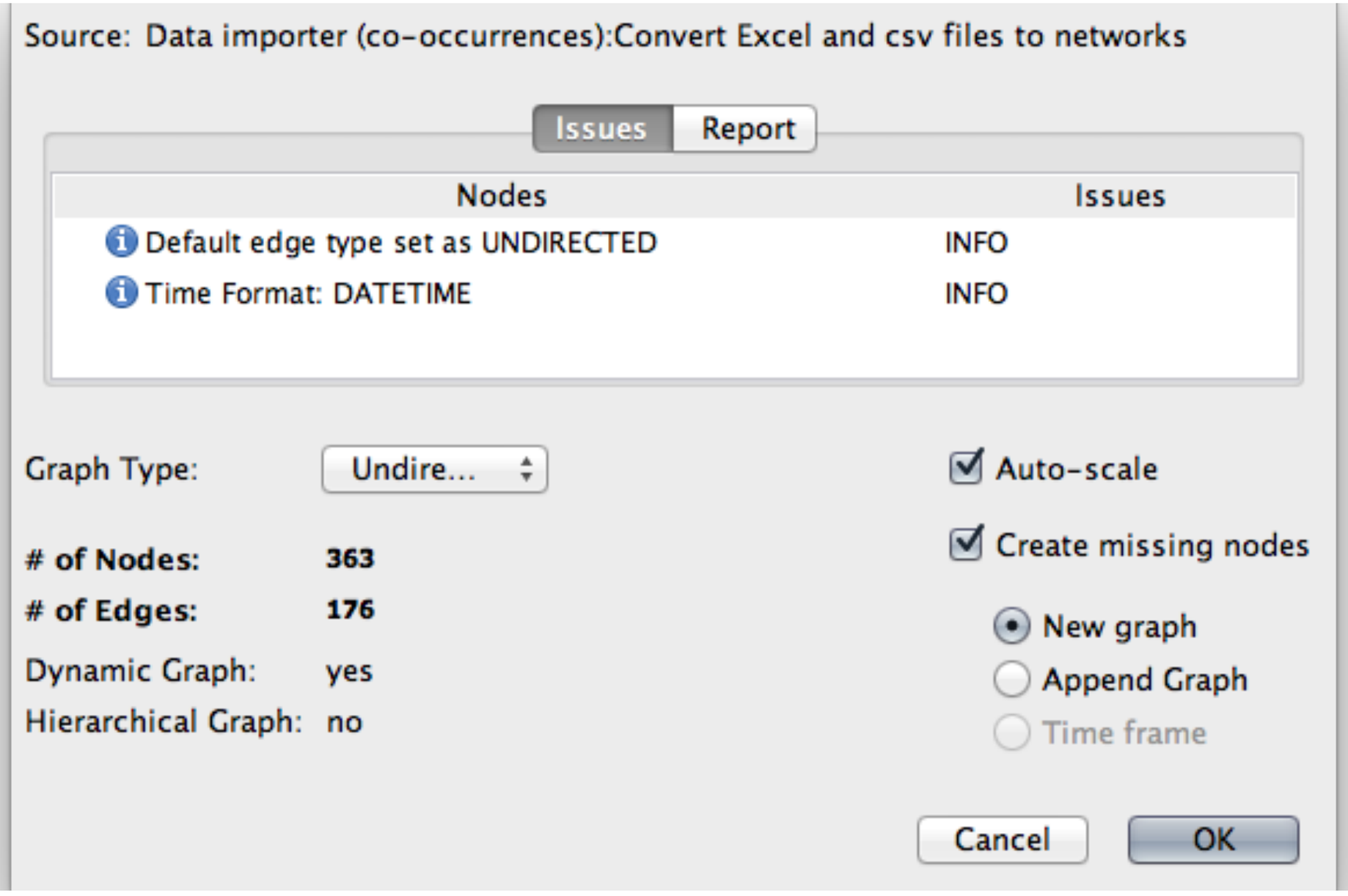
Create Links and Remove Self Loops

Removing self loops is removing the diagonal from the matrix where an agent is connected to itself. If self loops are not removed gephi creates a handle mark that we do not normally want.

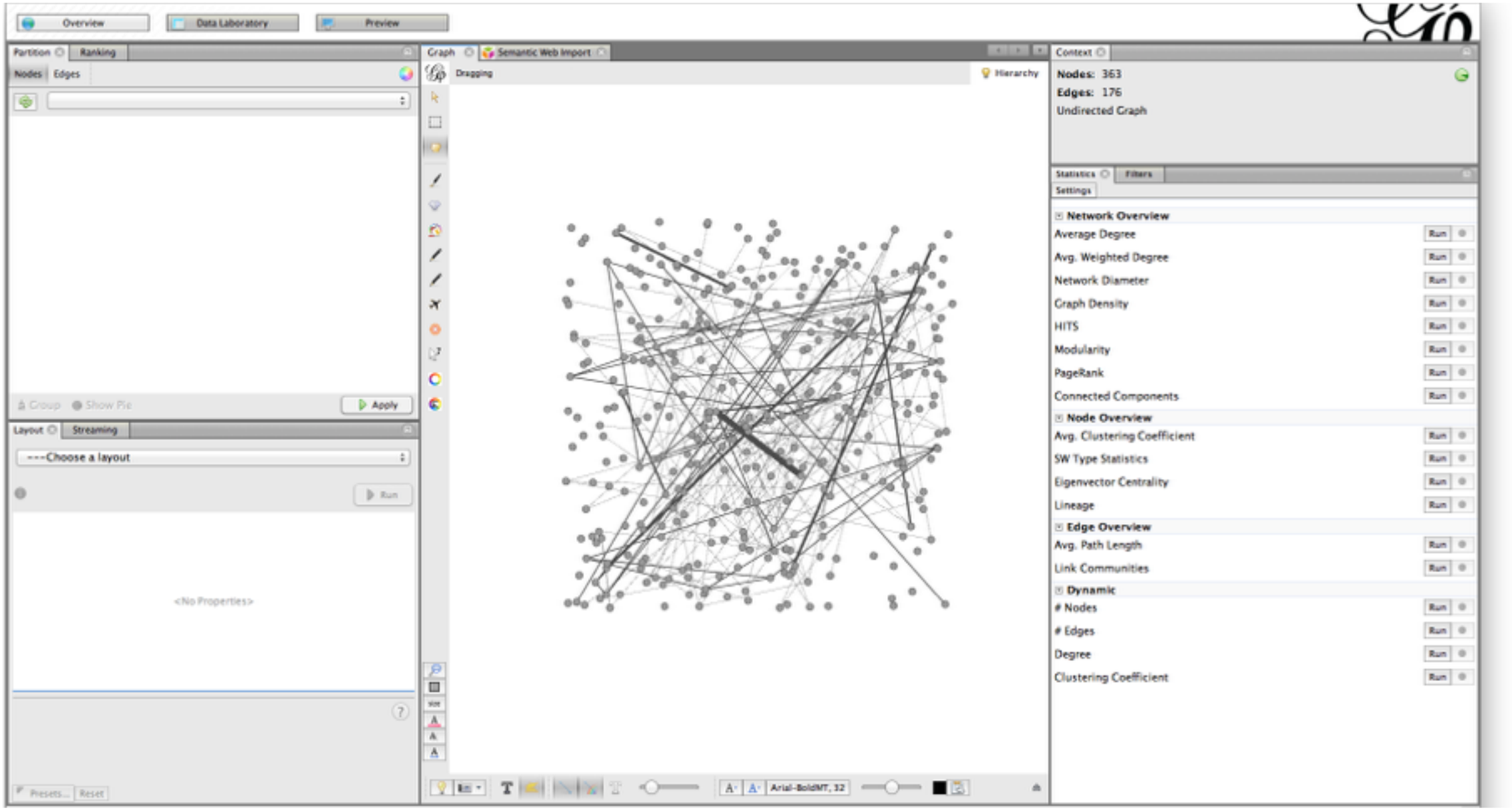


Conversion

Counts the number of times a node or agent appears in the data as the frequency score (note that can vary slightly from record counts but normally by a very small amount)

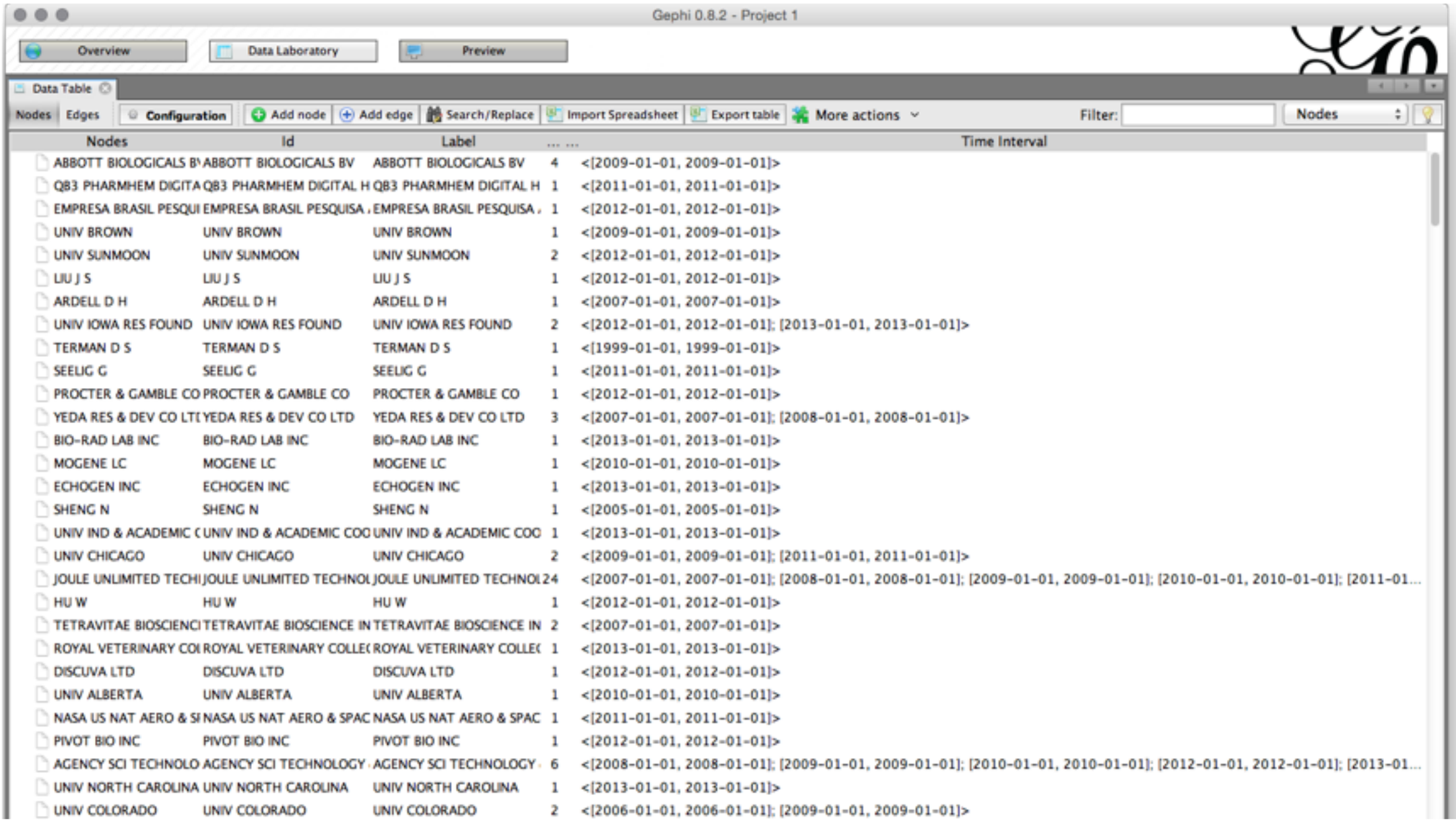


Generally we want an undirected graph



The raw network view

This network has not yet been laid our or coloured. Note the panels and tabs to the right and the left (statistics and filters). Note the



Nodes	Id	Label	... ...	Time Interval
ABBOTT BIOLOGICALS B'	ABBOTT BIOLOGICALS BV	ABBOTT BIOLOGICALS BV	4	<[2009-01-01, 2009-01-01]>
QB3 PHARMHEM DIGITA	QB3 PHARMHEM DIGITAL H	QB3 PHARMHEM DIGITAL H	1	<[2011-01-01, 2011-01-01]>
EMPRESA BRASIL PESQUI	EMPRESA BRASIL PESQUISA	EMPRESA BRASIL PESQUISA	1	<[2012-01-01, 2012-01-01]>
UNIV BROWN	UNIV BROWN	UNIV BROWN	1	<[2009-01-01, 2009-01-01]>
UNIV SUNMOON	UNIV SUNMOON	UNIV SUNMOON	2	<[2012-01-01, 2012-01-01]>
LIU J S	LIU J S	LIU J S	1	<[2012-01-01, 2012-01-01]>
ARDELL D H	ARDELL D H	ARDELL D H	1	<[2007-01-01, 2007-01-01]>
UNIV IOWA RES FOUND	UNIV IOWA RES FOUND	UNIV IOWA RES FOUND	2	<[2012-01-01, 2012-01-01]; [2013-01-01, 2013-01-01]>
TERMAN D S	TERMAN D S	TERMAN D S	1	<[1999-01-01, 1999-01-01]>
SEELIG G	SEELIG G	SEELIG G	1	<[2011-01-01, 2011-01-01]>
PROCTER & GAMBLE CO	PROCTER & GAMBLE CO	PROCTER & GAMBLE CO	1	<[2012-01-01, 2012-01-01]>
YEDA RES & DEV CO LTI	YEDA RES & DEV CO LTD	YEDA RES & DEV CO LTD	3	<[2007-01-01, 2007-01-01]; [2008-01-01, 2008-01-01]>
BIO-RAD LAB INC	BIO-RAD LAB INC	BIO-RAD LAB INC	1	<[2013-01-01, 2013-01-01]>
MOGENE LC	MOGENE LC	MOGENE LC	1	<[2010-01-01, 2010-01-01]>
ECHOGEN INC	ECHOGEN INC	ECHOGEN INC	1	<[2013-01-01, 2013-01-01]>
SHENG N	SHENG N	SHENG N	1	<[2005-01-01, 2005-01-01]>
UNIV IND & ACADEMIC C	UNIV IND & ACADEMIC COO	UNIV IND & ACADEMIC COO	1	<[2013-01-01, 2013-01-01]>
UNIV CHICAGO	UNIV CHICAGO	UNIV CHICAGO	2	<[2009-01-01, 2009-01-01]; [2011-01-01, 2011-01-01]>
JOULE UNLIMITED TECHI	JOULE UNLIMITED TECHNOL	JOULE UNLIMITED TECHNOL	24	<[2007-01-01, 2007-01-01]; [2008-01-01, 2008-01-01]; [2009-01-01, 2009-01-01]; [2010-01-01, 2010-01-01]; [2011-01...
HU W	HU W	HU W	1	<[2012-01-01, 2012-01-01]>
TETRAVITAE BIOSCIENCI	TETRAVITAE BIOSCIENCE IN	TETRAVITAE BIOSCIENCE IN	2	<[2007-01-01, 2007-01-01]>
ROYAL VETERINARY COI	ROYAL VETERINARY COLLE	ROYAL VETERINARY COLLE	1	<[2013-01-01, 2013-01-01]>
DISCUVA LTD	DISCUVA LTD	DISCUVA LTD	1	<[2012-01-01, 2012-01-01]>
UNIV ALBERTA	UNIV ALBERTA	UNIV ALBERTA	1	<[2010-01-01, 2010-01-01]>
NASA US NAT AERO & SI	NASA US NAT AERO & SPAC	NASA US NAT AERO & SPAC	1	<[2011-01-01, 2011-01-01]>
PIVOT BIO INC	PIVOT BIO INC	PIVOT BIO INC	1	<[2012-01-01, 2012-01-01]>
AGENCY SCI TECHNOLO	AGENCY SCI TECHNOLOGY	AGENCY SCI TECHNOLOGY	6	<[2008-01-01, 2008-01-01]; [2009-01-01, 2009-01-01]; [2010-01-01, 2010-01-01]; [2012-01-01, 2012-01-01]; [2013-01...
UNIV NORTH CAROLINA	UNIV NORTH CAROLINA	UNIV NORTH CAROLINA	1	<[2013-01-01, 2013-01-01]>
UNIV COLORADO	UNIV COLORADO	UNIV COLORADO	2	<[2006-01-01, 2006-01-01]; [2009-01-01, 2009-01-01]>

Select Data Laboratory Then Nodes

Overview | Data Laboratory | Preview

Data Table

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

Source	Target	Type	Id	Label	Weight	Time Interval
ANZA THERAPEUTICS INC	ADURO BIOTECH	Undirected	161			2 <[2008-01-01, 2008-01...
BEIJING WEISHENG LIDE TE...	BEIJING VIEWSOLIDBIOTECH...	Undirected	68			1 <[2011-01-01, 2011-01...
BEIJING WELLSENS BIOTECH...	BEIJING WEIXINYU BIOTECH...	Undirected	141			1 <[2010-01-01, 2010-01...
CITY OF HOPE	CALIFORNIA INST OF TECH...	Undirected	15			1 <[2009-01-01, 2009-01...
CORIDON PTY LTD	ADMEDUS VACCINES PTY LTD	Undirected	3			1 <[2007-01-01, 2007-01...
EASTMAN RENEWABLE MAT...	ADVANCED BIOFUELS INC	Undirected	159			2 <[2007-01-01, 2007-01...
EVOLVA NUTRITION INC	ABUNDA NUTRITION INC	Undirected	121			2 <[2010-01-01, 2010-01...
EVOLVA SA	ABUNDA NUTRITION INC	Undirected	79			2 <[2010-01-01, 2010-01...
EVOLVA SA	EVOLVA NUTRITION INC	Undirected	53			2 <[2010-01-01, 2010-01...
FEBIT HOLDING GMBH	FEBIT BIOTECH GMBH	Undirected	160			1 <[2006-01-01, 2006-01...
FISHER A	BROOKS J P	Undirected	162			1 <[2012-01-01, 2012-01...
FONG S S	BROOKS J P	Undirected	26			1 <[2012-01-01, 2012-01...
FONG S S	FISHER A	Undirected	140			1 <[2012-01-01, 2012-01...
HARVARD COLLEGE	BROAD INST INC	Undirected	64			2 <[2011-01-01, 2011-01...
INST PASTEUR	CENT NAT RECH SCI	Undirected	49			1 <[2012-01-01, 2012-01...
JEE N Y	GARTNER Z J	Undirected	51			1 <[2012-01-01, 2012-01...
JOULE UNLIMITED TECHNO...	JOULE BIOTECHNOLOGIES INC	Undirected	52			4 <[2007-01-01, 2007-01...
KEMIJSKI INST	EN-FIST CENT ODLIAOENOSTI	Undirected	82			1 <[2012-01-01, 2012-01...
KICKENWEIZ T	GLIEDER A	Undirected	118			1 <[2013-01-01, 2013-01...
KITZMAN J	HIATT J	Undirected	32			1 <[2011-01-01, 2011-01...
LEE T M	ARNOLD F H	Undirected	99			1 <[2013-01-01, 2013-01...
LIFE TECHNOLOGIES CORP	APPLIED BIOSYSTEMS INC	Undirected	86			1 <[2007-01-01, 2007-01...
LIFE TECHNOLOGIES CORP	GENEART AG	Undirected	98			3 <[2011-01-01, 2011-01...
LIU C	CAI Y	Undirected	104			1 <[2013-01-01, 2013-01...
LIU J S	GARTNER Z J	Undirected	70			1 <[2012-01-01, 2012-01...
LIU J S	JEE N Y	Undirected	10			1 <[2012-01-01, 2012-01...
LIU P	HU W	Undirected	137			1 <[2012-01-01, 2012-01...
MALI P G	CHURCH G M	Undirected	174			1 <[2012-01-01, 2012-01...
MASSACHUSETTS INST TEC...	BROAD INST INC	Undirected	45			2 <[2011-01-01, 2011-01...
MASSACHUSETTS INST TEC...	HARVARD COLLEGE	Undirected	148			3 <[2010-01-01, 2010-01...
MAYO S L	ARNOLD F H	Undirected	90			1 <[2013-01-01, 2013-01...
MAYO S L	LEE T M	Undirected	31			1 <[2013-01-01, 2013-01...
MONK J	DEHART C	Undirected	92			1 <[2008-01-01, 2008-01...

Now Select the Edges

Run the Modularity Algorithm

This will produce a pop up window that calculates the number of communities. Notice the reference article by Blondel et al at the bottom and read it later (it is very good).

Physics > Physics and Society

Fast unfolding of communities in large networks

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre

(Submitted on 4 Mar 2008 (v1), last revised 25 Jul 2008 (this version, v2))

We propose a simple method to extract the community structure of large networks. Our method is a heuristic method that is based on modularity optimization. It is shown to outperform all other known community detection method in terms of computation time. Moreover, the quality of the communities detected is very good, as measured by the so-called modularity. This is shown first by identifying language communities in a Belgian mobile phone network of 2.6 million customers and by analyzing a web graph of 118 million nodes and more than one billion links. The accuracy of our algorithm is also verified on ad-hoc modular networks. .

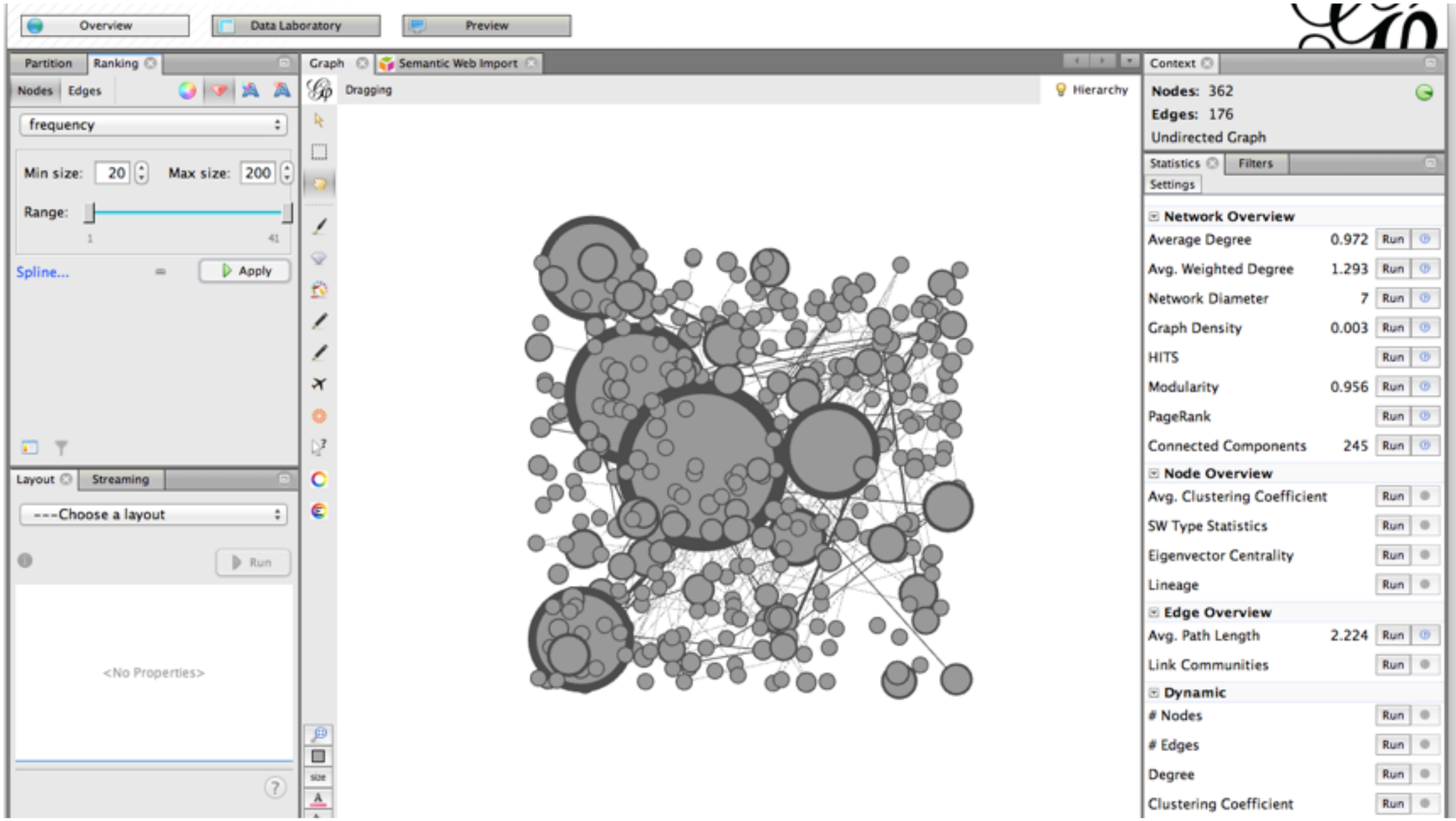
Comments: 6 pages, 5 figures, 1 table; new version with new figures in order to clarify our method, where we look more carefully at the role played by the ordering of the nodes and where we compare our method with that of Wakita and Tsurumi

Subjects: **Physics and Society (physics.soc-ph)**; Statistical Mechanics (cond-mat.stat-mech); Computers and Society (cs.CY); Data Structures and Algorithms (cs.DS)

Journal reference: J. Stat. Mech. (2008) P10008

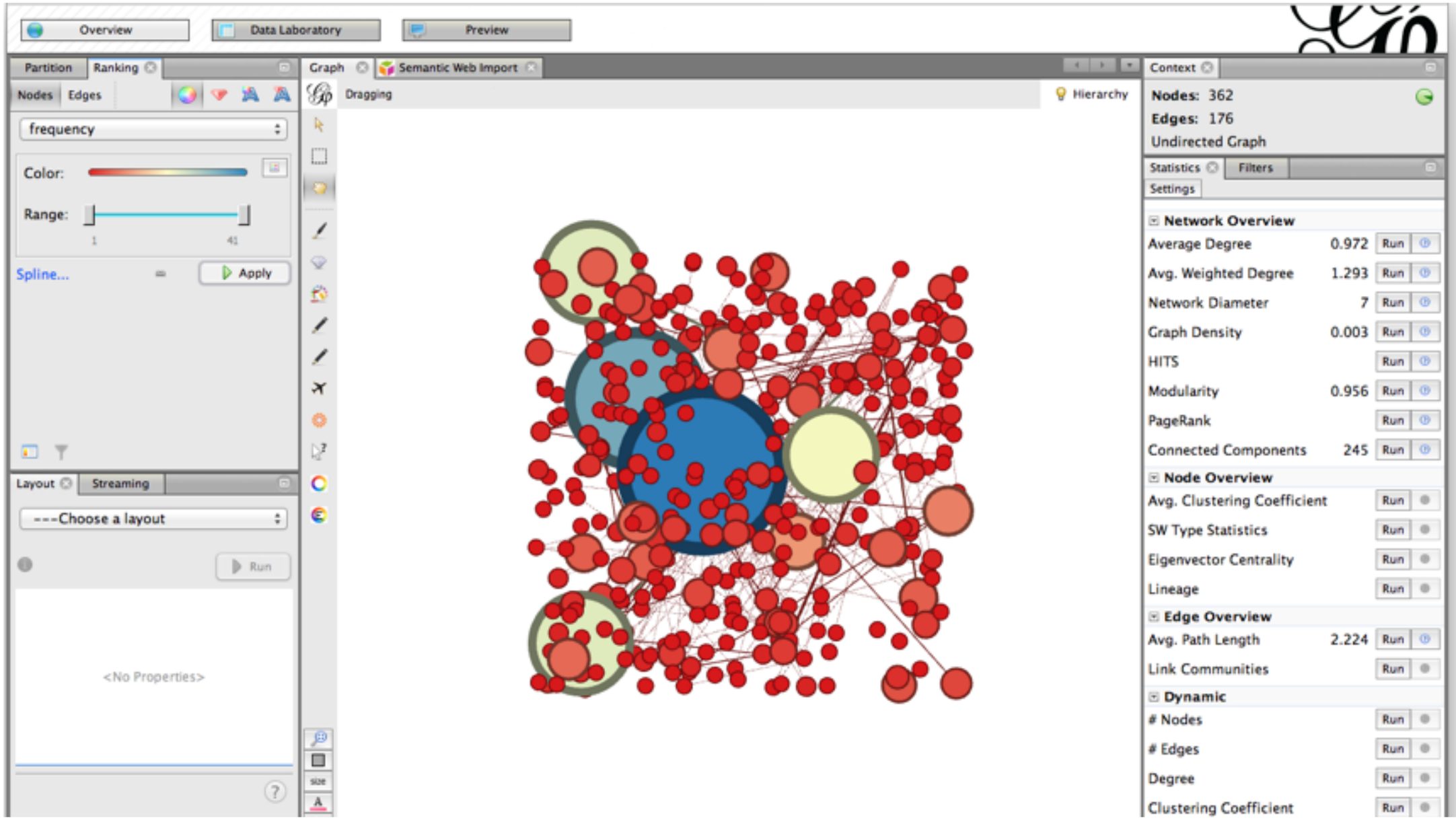
Modularity Class Algorithm

Allocates nodes to communities iteratively on the strength of the links between then (allocates to one community only). Mainly used on large networks but extremely useful for identifying clusters of interlinked actors



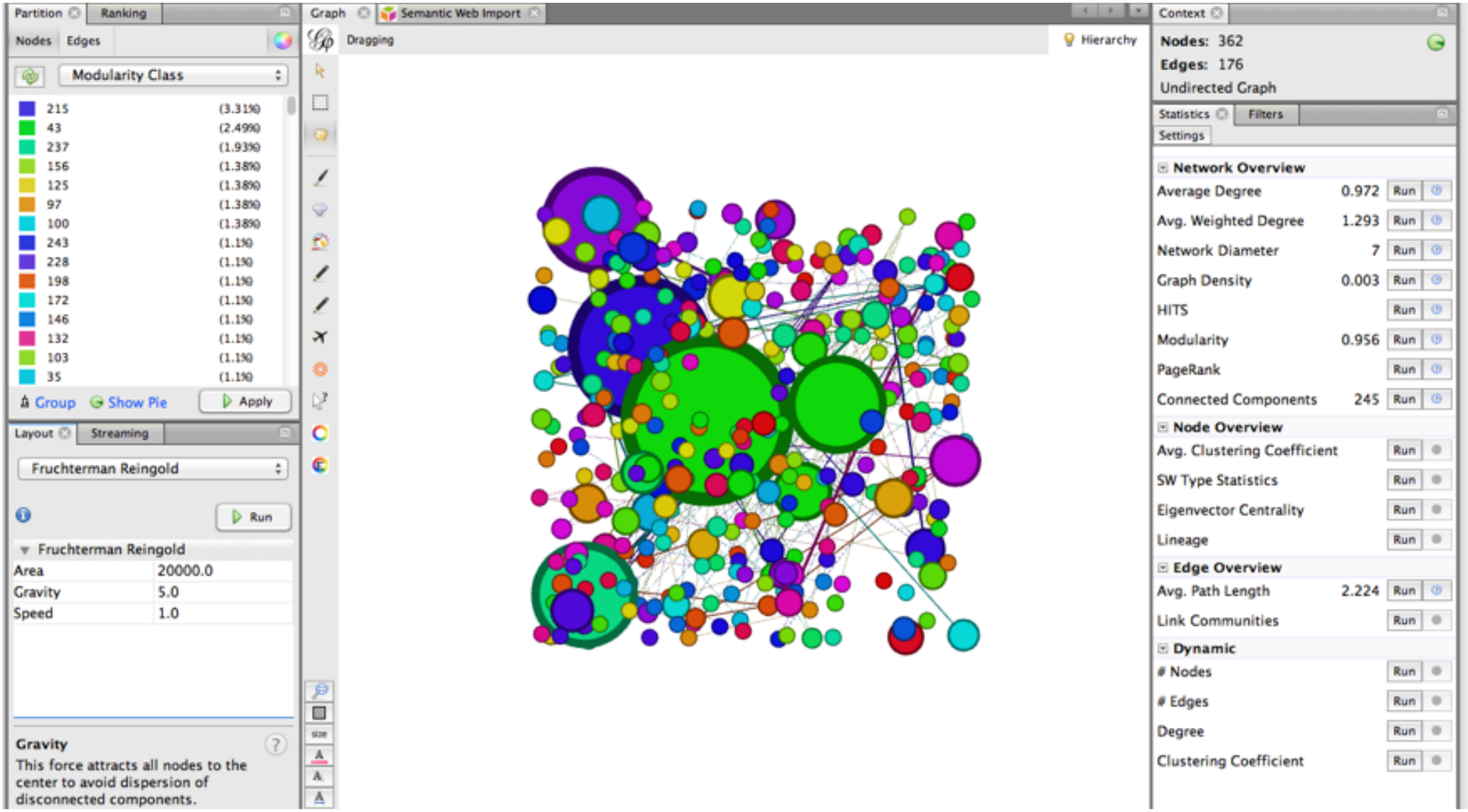
Size the Nodes - Ranking Tab

Size the nodes with the red inverted triangle. Choose limits such as 20 and 200 (smaller Min size has an impact on ease of viewing labels later)



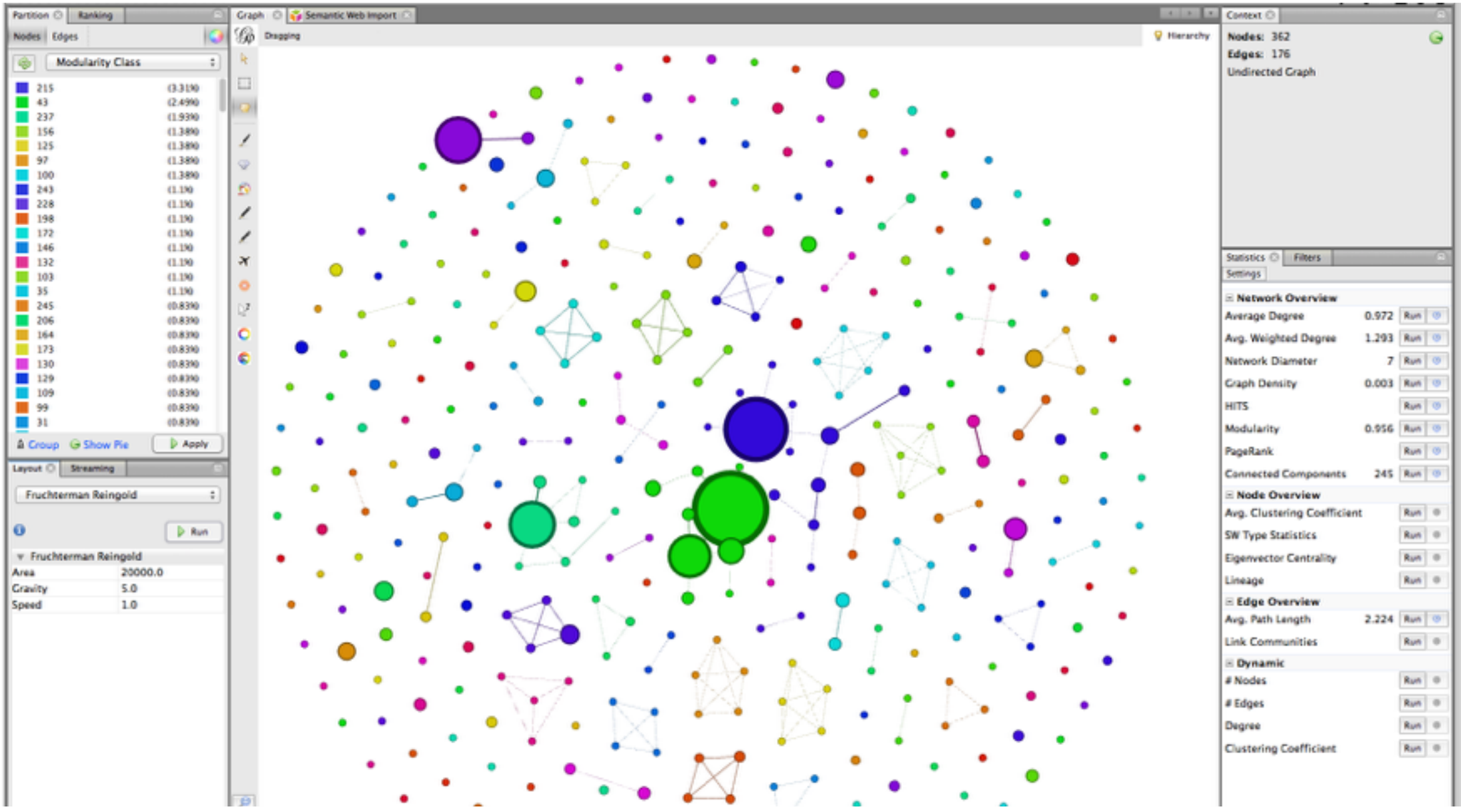
Choose the Colour Wheel Button. Then Apply

You can right click on the small box next to the colour wheel to choose different colours. Also try using the slider.



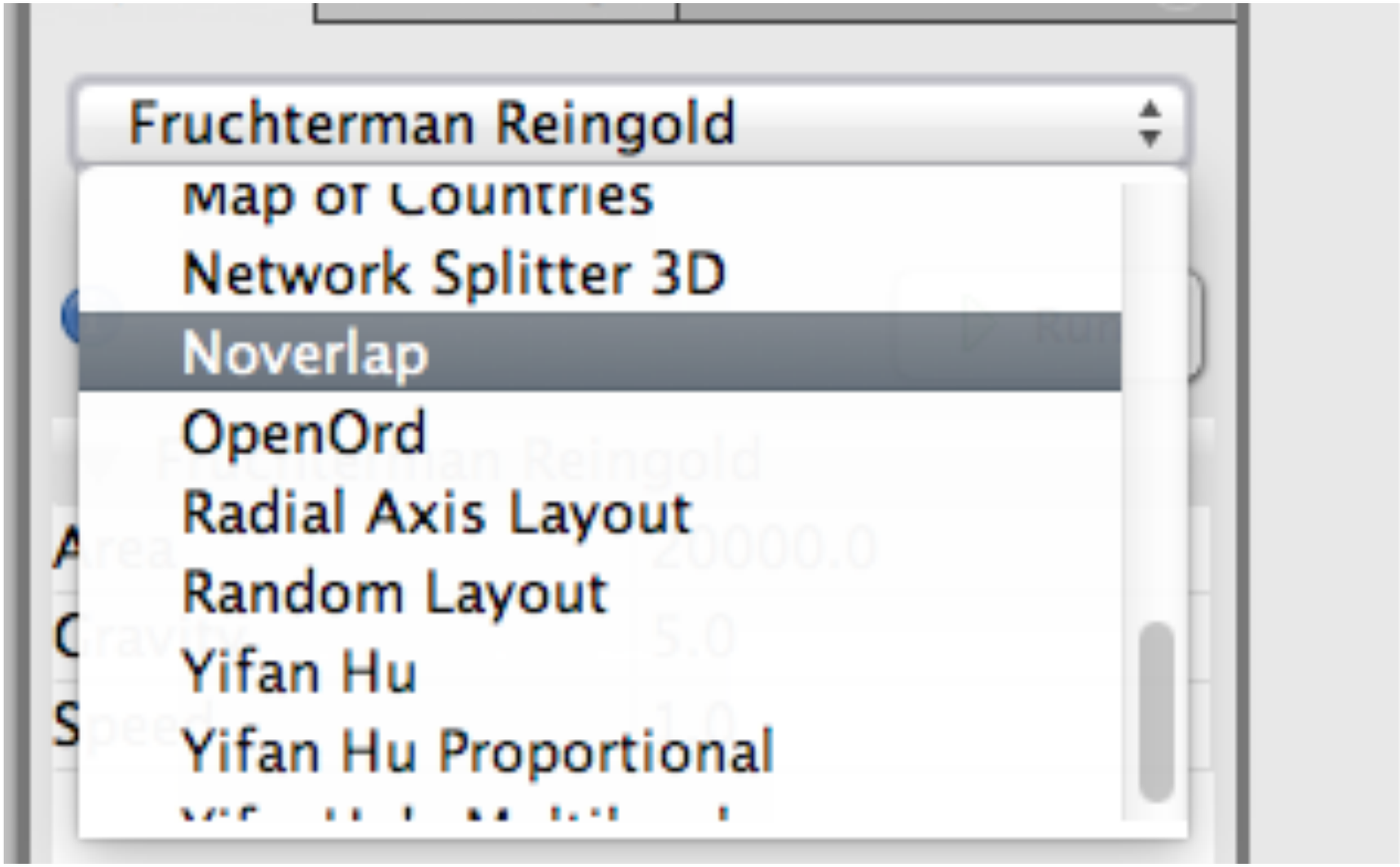
Choose Partition - Nodes

Press the green arrows and then choose Modularity Class. Note that our nodes are now coloured by clusters or communities.



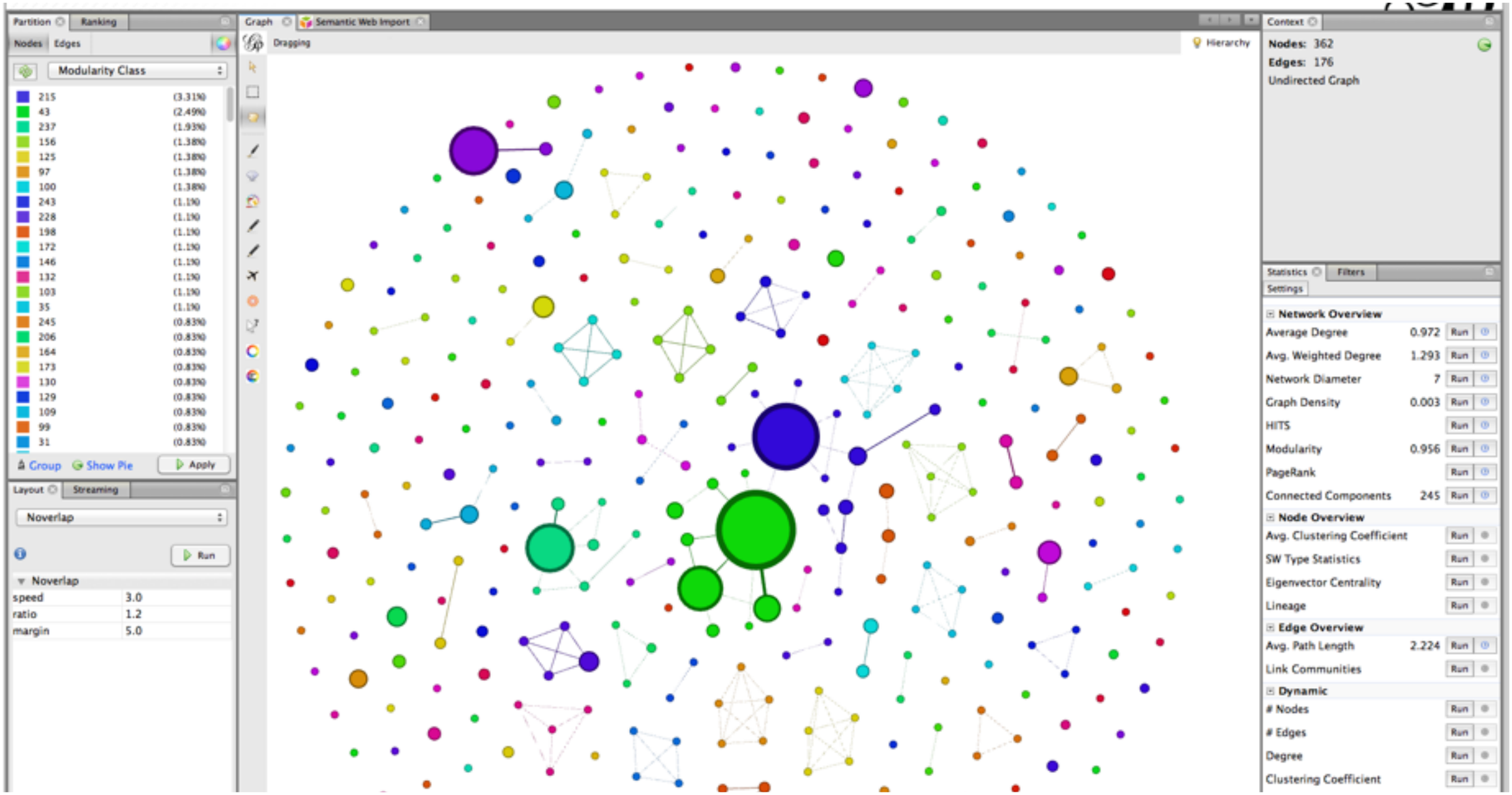
Layout Tab

Choose Fruchterman Reingold. Use 15000 for Area and 5.0 for gravity. Then press Run.
Do not press stop yet. (Illustrate manual adjustment to layout while running)

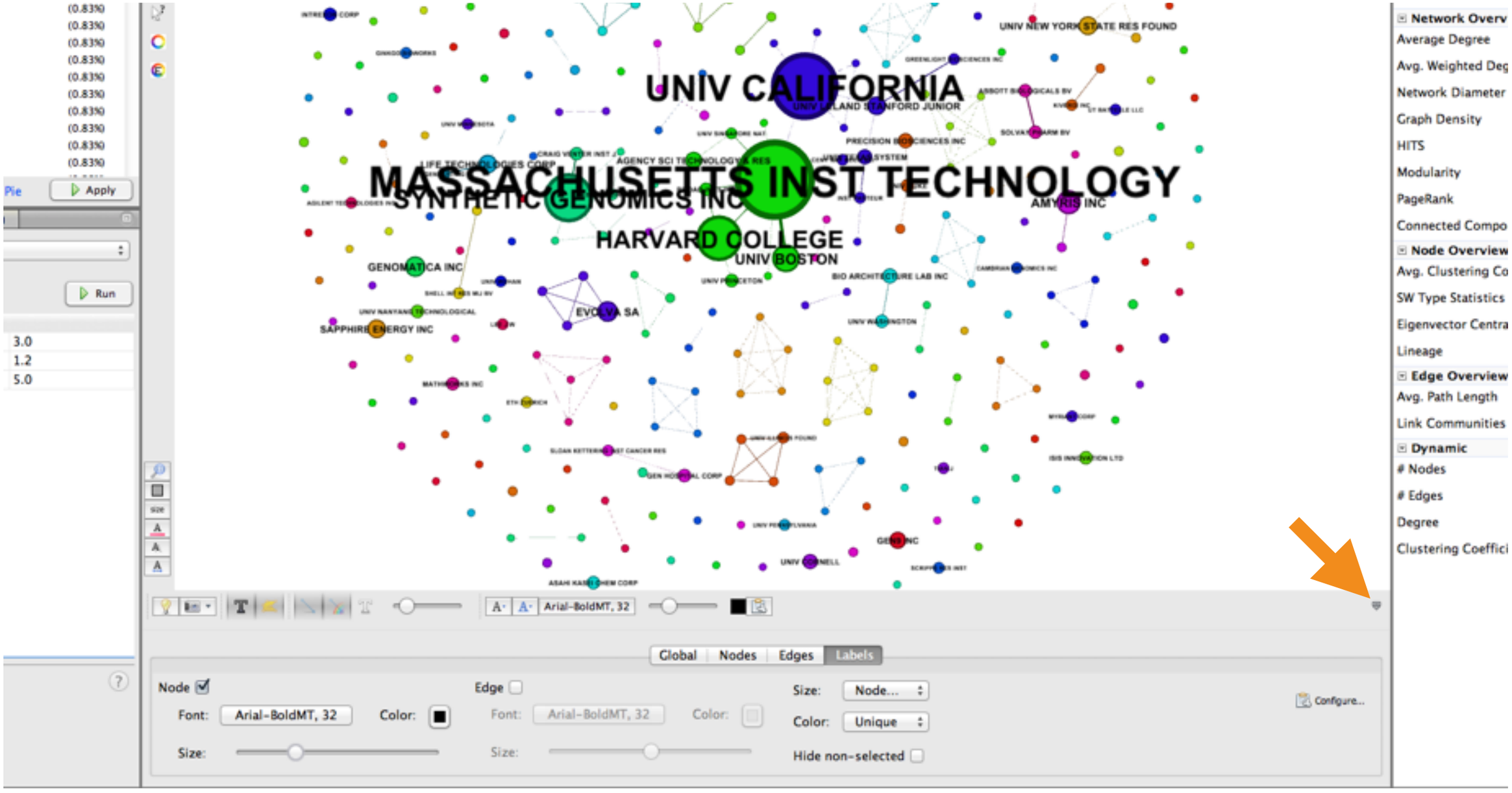


Use No overlap to prevent nodes overlapping

Make sure you have installed this plugin (also LabelAdjust is helpful)

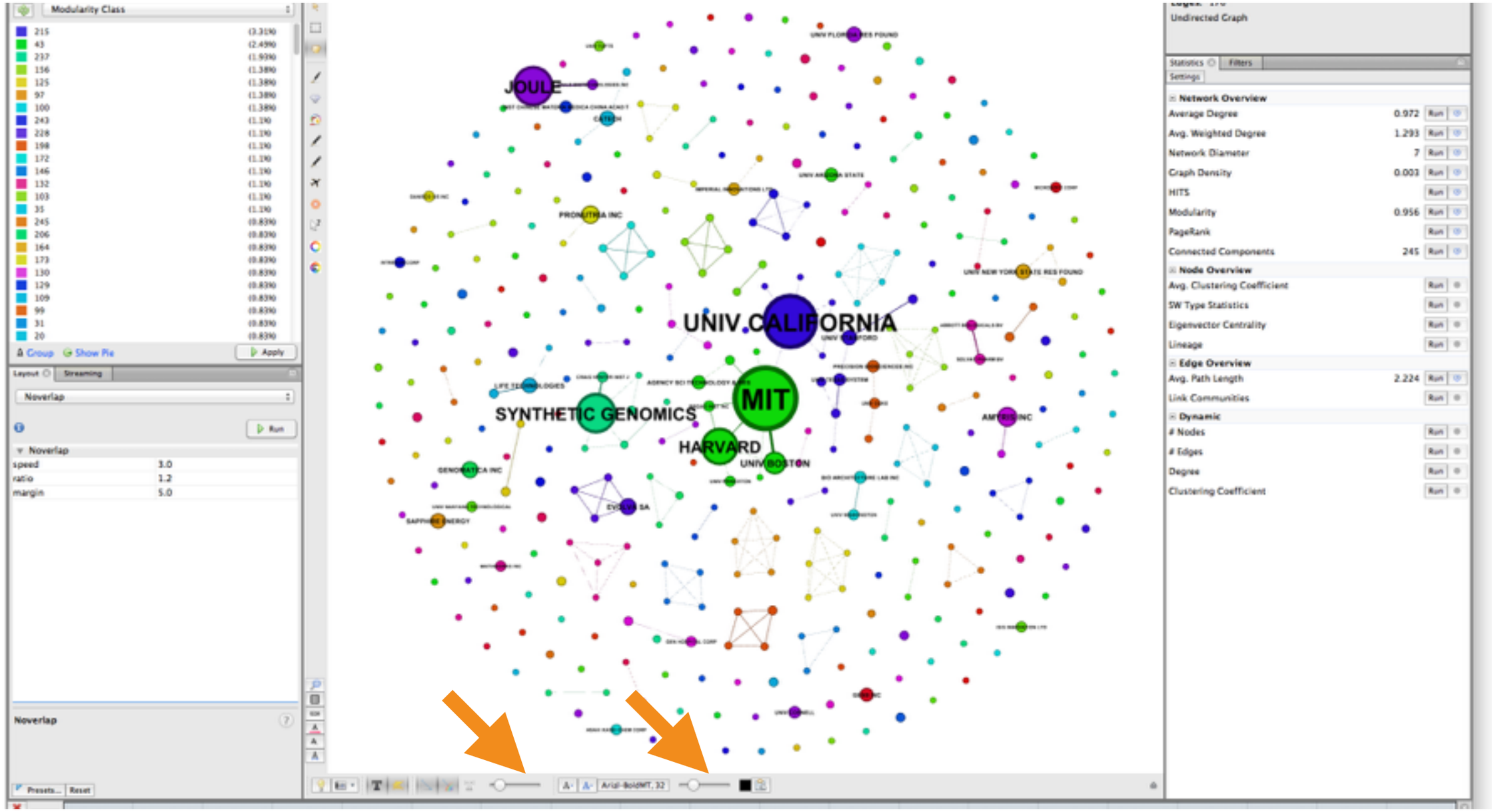


No overlapping nodes



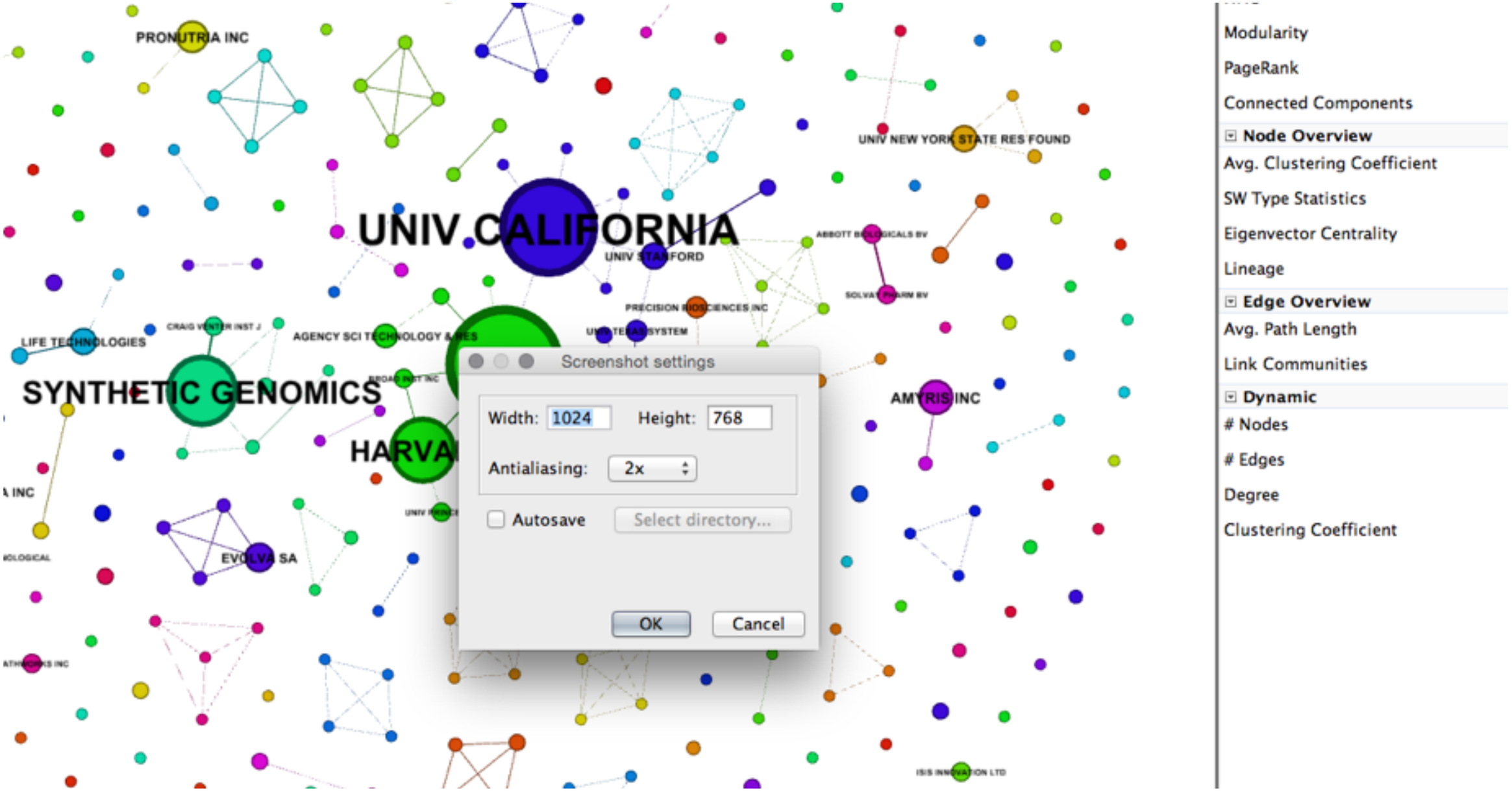
Click the inverted triangle for a new menu

Then select the Labels, Check the Node Box and move over to Size. Change to Node Size. Adjust long labels such as MIT in Data Laboratory, Nodes, Labels. Delete the NA node.

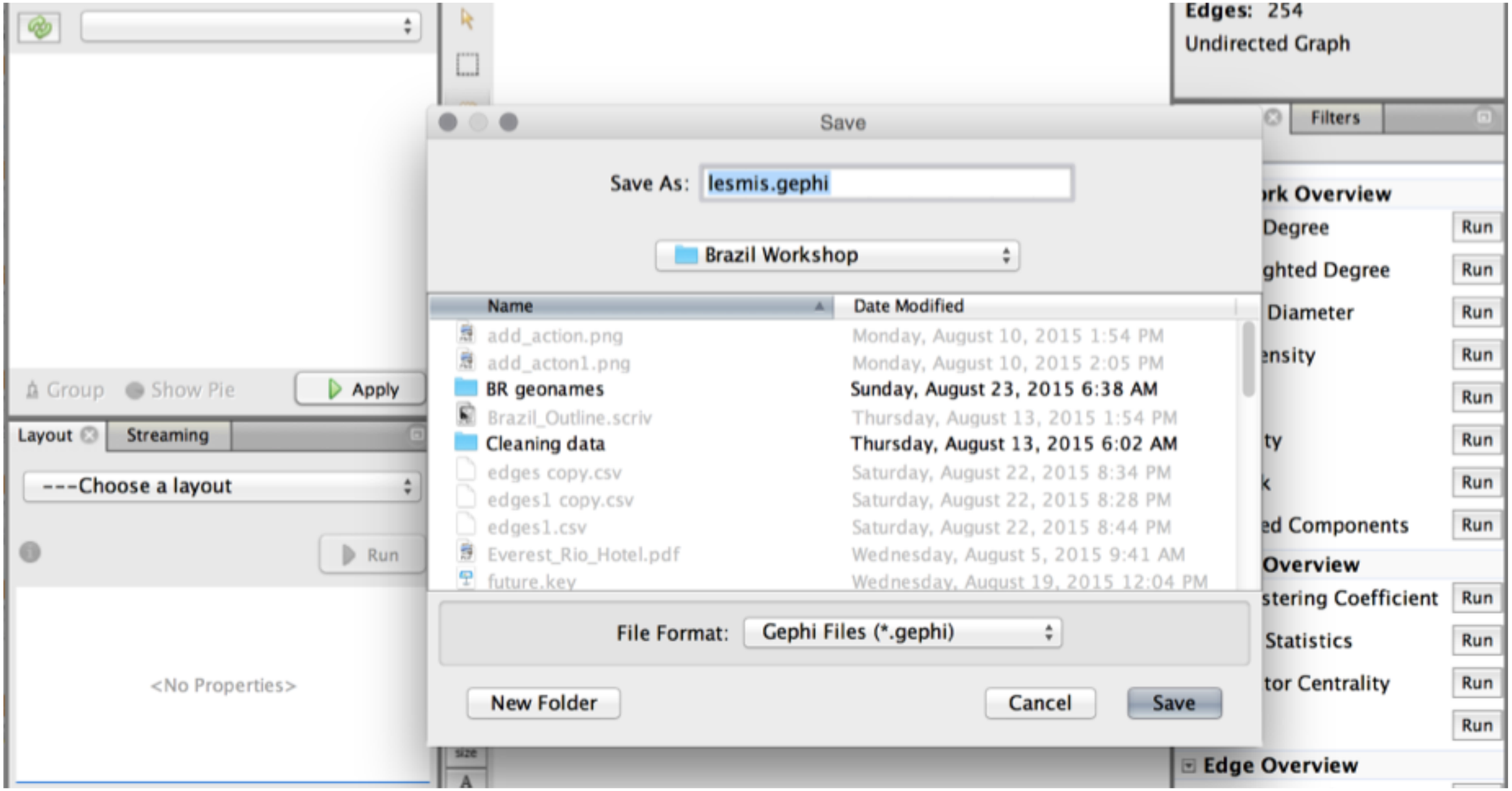


Move the sliders to adjust the sizes

Note the small camera sign, on the bottom menu. This can be used to take a screen shot (useful when not seeking publication quality graphics)



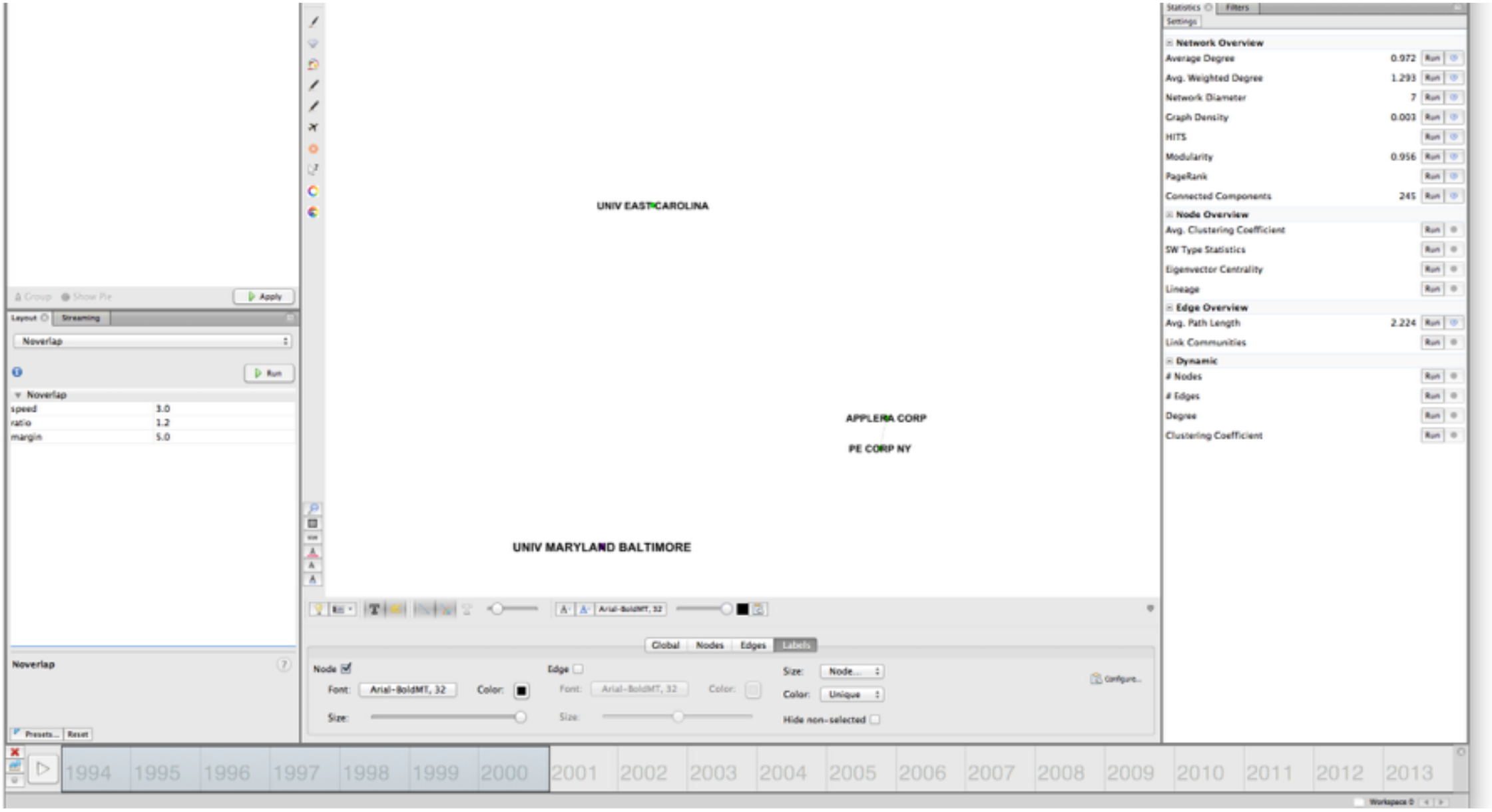
Screenshot settings can be changed



Saving gephi files

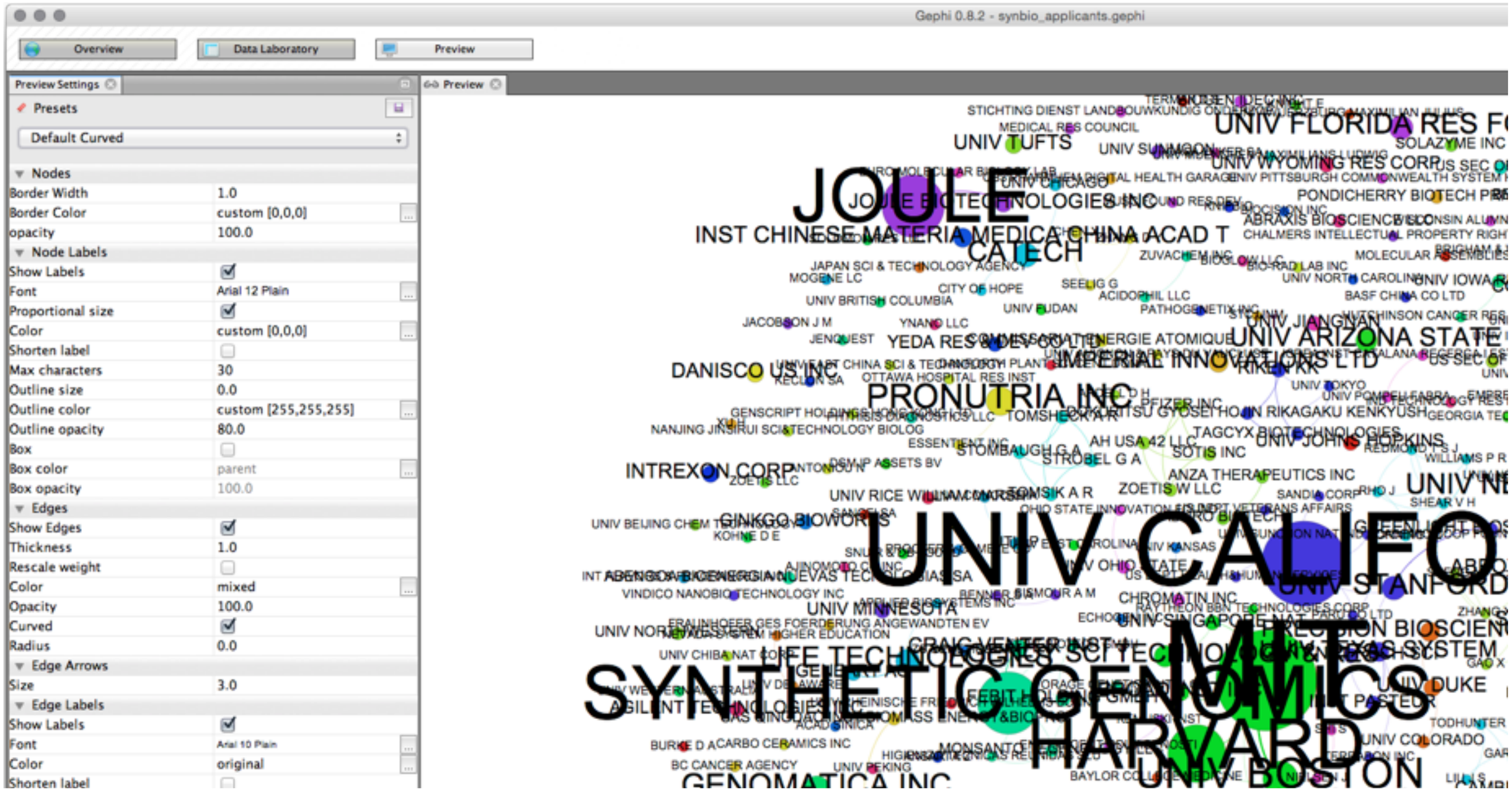
It is important to save files regularly as gephi is unstable.

Using the Timeline



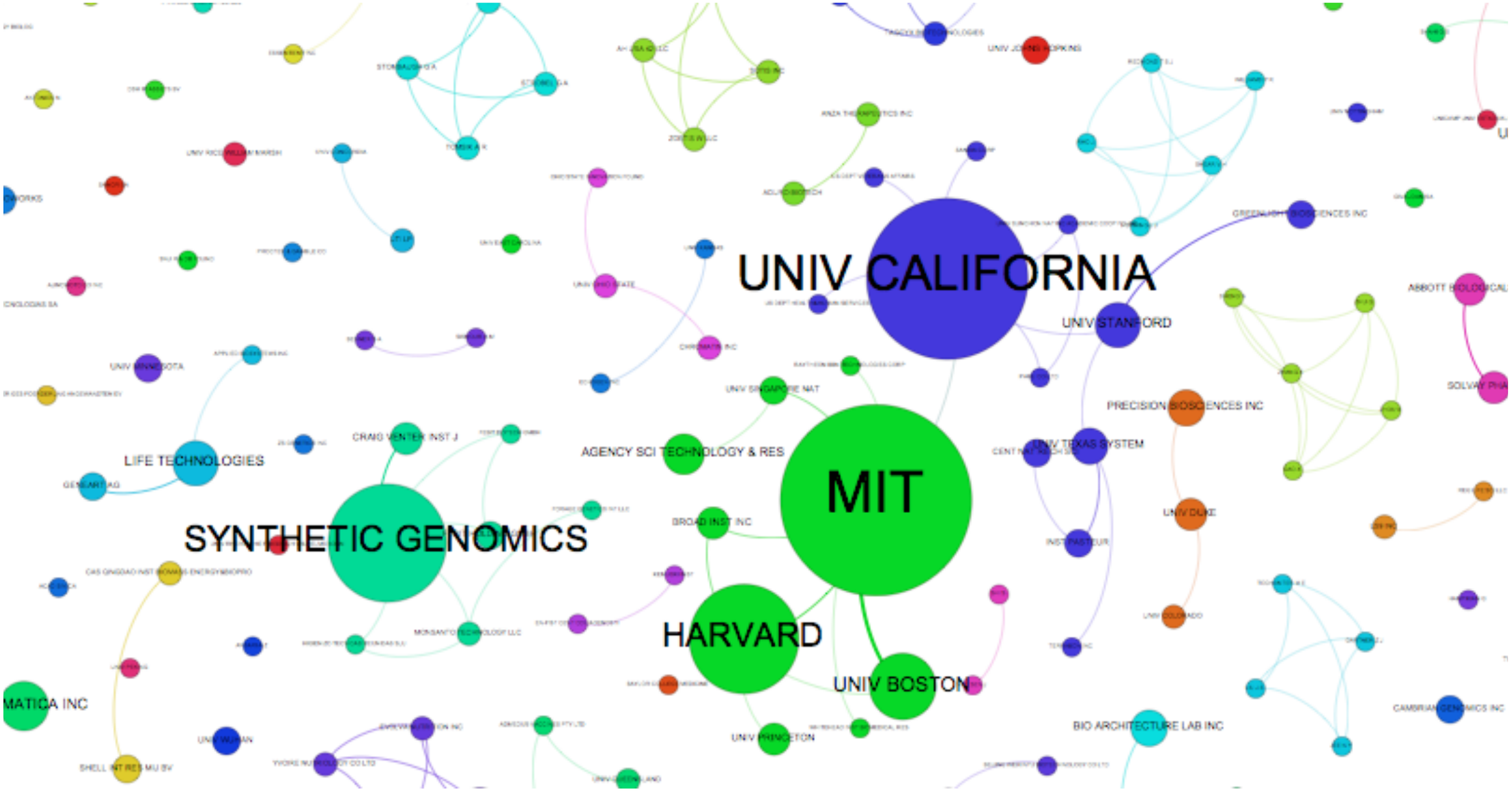
Select Enable Timeline at the bottom of the screen

Then select a few years (1994-2000) and play around. Note that the size of the nodes does not seem to adjust with time. Note that important to lay out the final graph first.



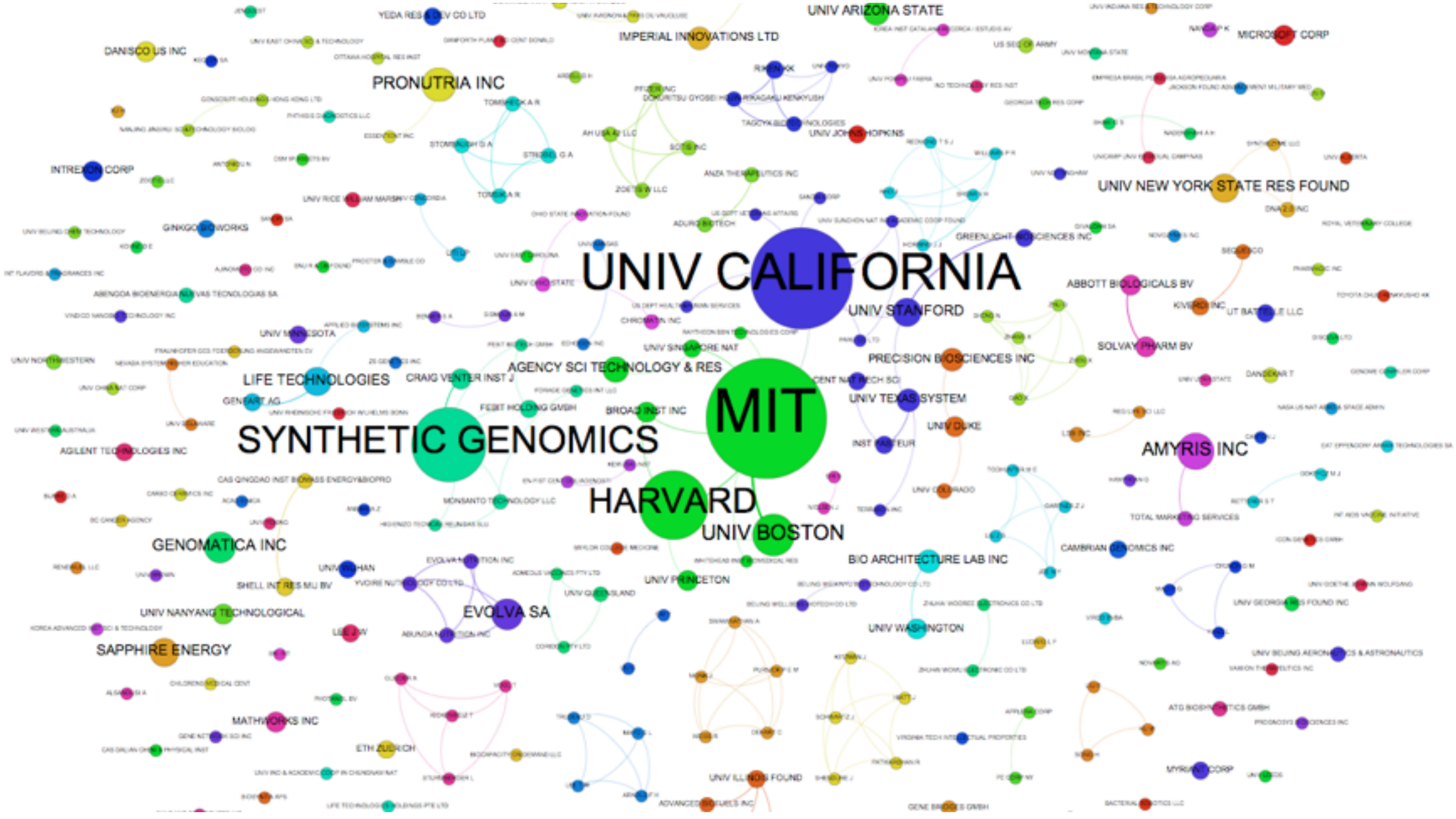
Preview Pane - for publication graphics

Select Default Curved. Then under Font reduce the Font size to 3 (experimentation is normally needed here). Use Refresh to View changes. You may need to edit long labels in Data Laboratory Long (or merge if duplicates found).



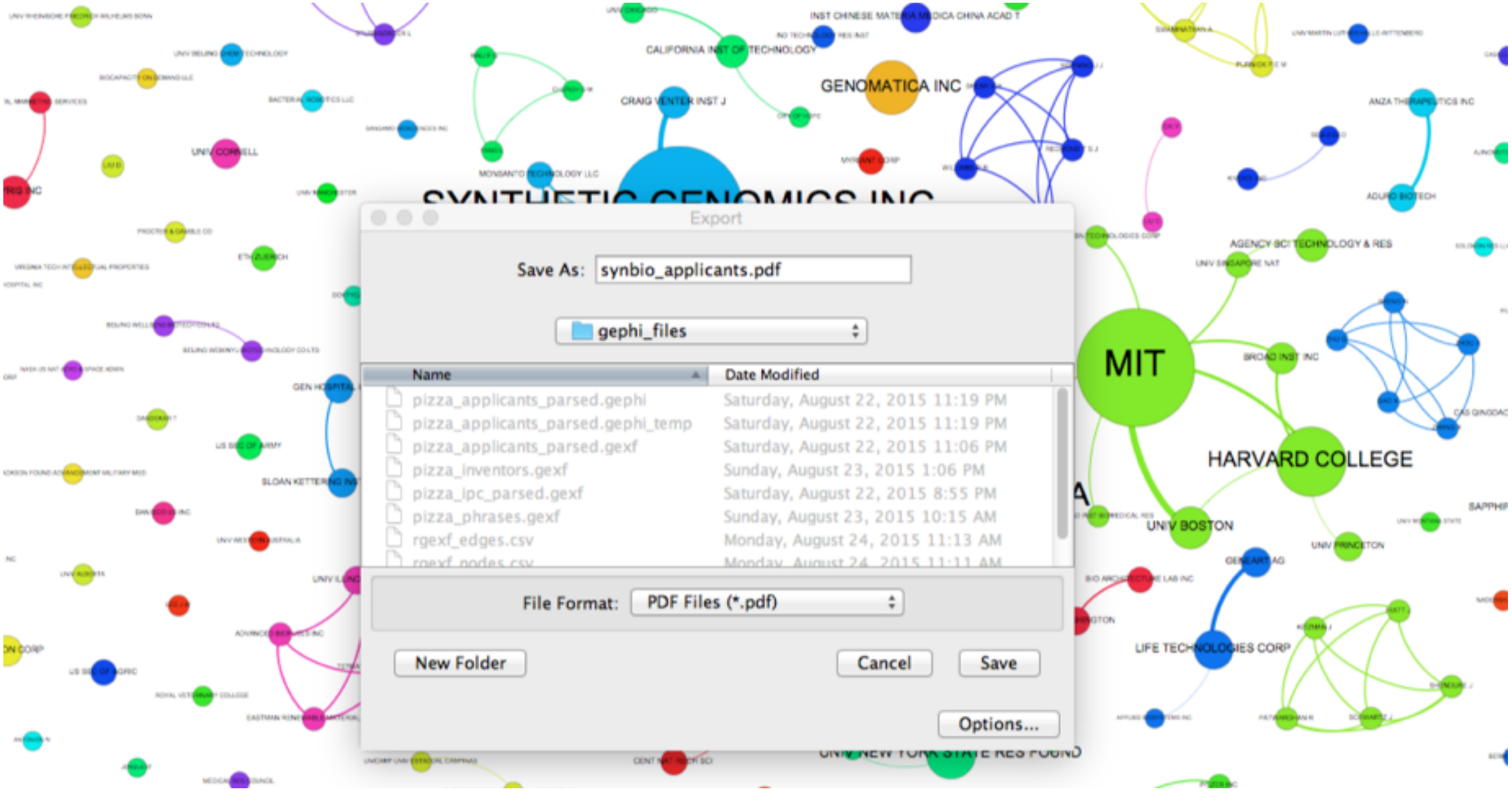
Adjust Overlapping Nodes and Labels

Starting at 12 o'clock move around the graph to identify nodes that are overlapping. Go to overview and move the node the amount necessary. Then in preview use Refresh to check results.



The outcome

No overlapping nodes or labels. Attempt also to avoid edges intersecting with unrelated nodes wherever possible (and make a note where not possible)



Choose Export near Refresh Button

The pdf defaults are good (so leave them as is). You may need to adjust overhanging labels and outline text for publication in Illustrator or GIMP.

This repository Search Pull requests Issues Gist

poldham / **opensource-patent-analytics** Unwatch 1

Branch: **master** **opensource-patent-analytics** / brazil_workshop / **gephi_files** / +

moving

poldham authored 20 seconds ago latest commi

..

pizza_applicants_parsed.gexf	moving
pizza_inventors.gexf	moving
pizza_ipc_parsed.gexf	moving
pizza_phrases.gexf	moving

Assignment

In the brazil_workshop folder there are ready made .gexf files. Experiment with these tonight to create networks for the morning. Include the results in your Tableau Public pizza patent workbook.

Assignment Tips

- Use Welcome - Open Graph File.
- In Data Laboratory - Delete NA at the very beginning.
- There are a lot of nodes. Use Filters then Attributes then Range = records and use the slider or type in value until you have something reasonable (keep pressing Filter). The aim is communication not completeness.
- Edit long labels in Data Lab - Nodes.
- Save your work with .gephi extension before using Preview.
- Tableau - create a new dashboard and look at options.
- Bring to the session tomorrow morning