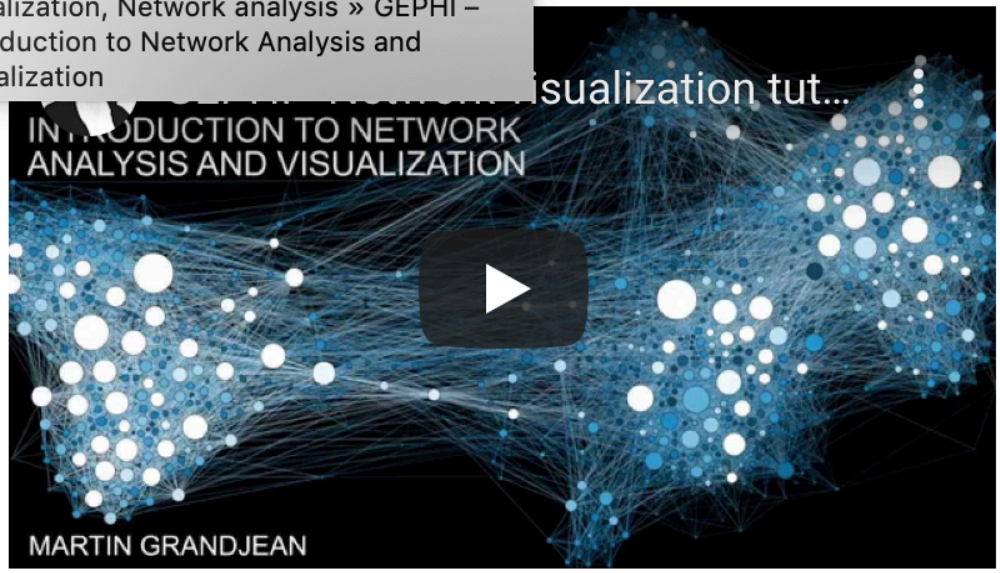
GEPHI – Introduction to Network Analysis and Visualization

COMMENTS 199 SHARE +

Network Analysis and visualization appears to be an interesting tool to give the researcher the ability to see its data from a new angle. Because Gephi is an easy access and powerful network analysis tool, we propose a tutorial designed to allow everyone to make his first experiments on two complementary datasets.

Martin Grandjean » Digital humanities, Data visualization, Network analysis » GEPHI – Introduction to Network Analysis and Visualization



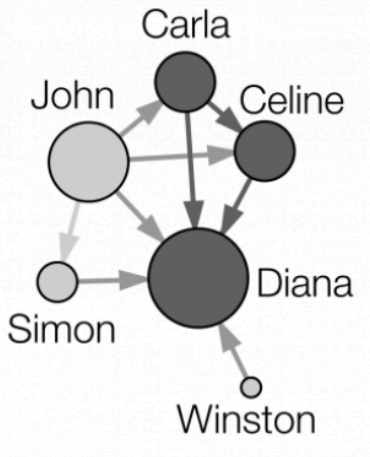
Network Analysis and visualization appears to be an interesting tool to give the researcher the ability to see its data from a new angle. Because Gephi is an easy access and powerful network analysis tool, we propose a tutorial designed to allow everyone to make his first experiments on two complementary datasets.

After a short introduction about the basis of SNA and some examples which shows the potential of this tool and gives some inspiration, this tutorial is divided into 2 main “exercices”: a geographical network of 1000 individuals sending letters all over Europe and a 2-mode network of 100 members of 10 different institutions.

Download the PDF version

1. INTRODUCTION

1.1 A short introduction to Social Network Analysis



Nodes

Id,Label,Attribute
1,John,1
2,Carla,2
3,Simon,1
4,Celine,2
5,Winston,1
6,Diana,2

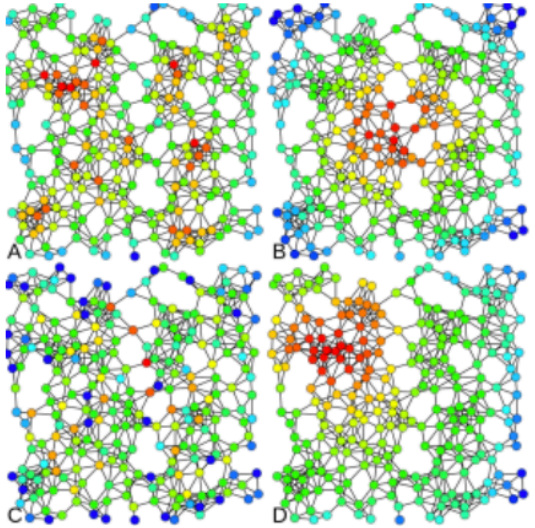
Edges

Source,Target
1,2
1,3
1,4
1,6
2,4
2,6
3,6
4,6
5,6

A network is made of two components : a **list of the actors** composing the network, and a **list of the relations** (the interactions between actors). As part of a mathematical object, actors will then be called *vertices* (***nodes***, in Gephi), and relations will be denoted as *tiles* (***edges***, in Gephi).

Here left, a **very simple directed social graph, with both lists explicited**. Two attributes are attached to the nodes : a **label** (his or her “name”) and a numeric attribute (here, a distinction between boys and girls). In the edge list, “Source” and “Target” entries refer to the nodes’ identifiers (Id).

In our example, the attribute determines the color of the nodes. The size of a node depends on the value of its “degree centrality” (its number of connexions). The **centrality measures** are essential metrics to analyze the position of an actor in a network. They come in many variations, as shown at right (A = *Degree centrality*, number of connexions ; B = *Closeness centrality*, closeness to the entire network ; C = *Betweenness centrality*, bridges nodes ; D = *Eigenvector centrality*, connexion to well-connected nodes).



— 4 types of centrality measures (Claudio Rocchini, Wikimedia)

2. SET UP

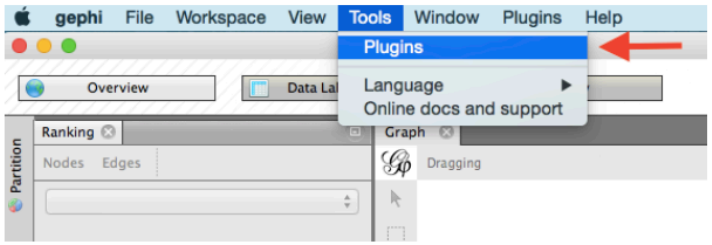
2.1 Downloading and installing the software

The software can be freely downloaded here:

To the Gephi website

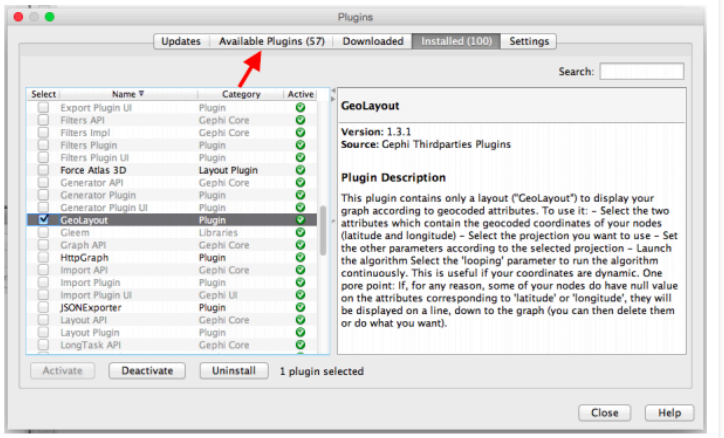
▲ Gephi is working on a previous version of Java. On an Apple computer running a recent version of OS X (10.7 Lion and further), to be able to run Gephi, you'll have to download and install a previous version of Java (Java 6 instead of your Java 7 or 8), find it here. Some compatibility problems may also occur with some Microsoft configurations. You'll find more ressources about this issue on Gephi forums / Gephi Facebook group / other websites (see in particular here for Mac, and here for Windows). After a few attempts, do not hesitate to let a comment here!

2.2 A few plugins



— Plugins

In order to go beyond the basic functionalities of the software, we will work with three additional plugins: **GeoLayout**, **NoverlapLayout** and **Multimode Networks Transformation**. You'll find the **Plugins** in the **Tools** menu. Refresh the list and select the requested plugins. You'll have to restart Gephi shortly after the download (plugins appear only after a restart).



— Plugins

2.3 About the datasets

We will use **two datasets** (different data to explore different features) :

Dataset 1

1.000 nodes / 14.116 edges (1-mode, directed)

Set 1 EDGES **Set 1 NODES**

Dataset 2

110 nodes / 142 edges (2-mode, undirected)

Set 2 EDGES **Set 2 NODES**

	A	B
1	Id	Label
2	376088951	name1
3	17647430	name2
4	32416061	name3
5	550180187	name4
6	28110685	name5
7	14845783	name7
8	63293	name8
9	22606966	name9
10	193763394	name10
11	191004748	name11
12	61502712	name13
13	2230301	name16
14	17870064	name17
15	123516974	name18
16	10161492	name19
17	16856080	name20
18	94154580	name21
19	20059362	name22
20	20646457	name23
21	109562464	name24

	A	B	C	D
1	Source	Target	Type	Weight
2	376088951	17647430	Directed	1
3	376088951	32416061	Directed	1
4	376088951	550180187	Directed	2
5	376088951	28110685	Directed	3
6	376088951	870137221	Directed	3
7	550180187	376088951	Directed	2
8	550180187	17870064	Directed	2
9	550180187	320140078	Directed	1
10	550180187	4308031	Directed	1
11	550180187	34881762	Directed	1
12	550180187	17150632	Directed	2
13	550180187	34621309	Directed	3
14	550180187	57601933	Directed	3
15	550180187	277980462	Directed	2
16	550180187	133671478	Directed	3
17	550180187	65947781	Directed	3
18	28110685	376088951	Directed	3
19	28110685	14845783	Directed	1
20	28110685	63293	Directed	2
21	28110685	109562464	Directed	3
22	28110685	320140078	Directed	1

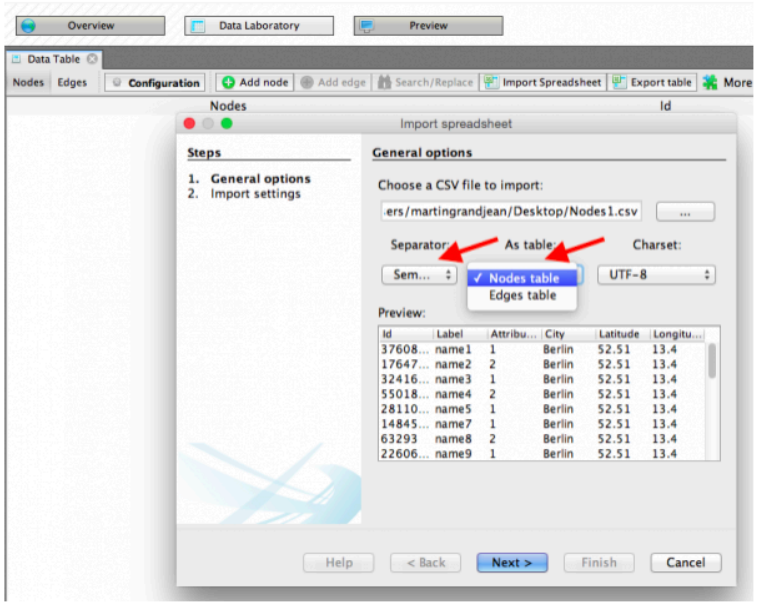
— Dataset1 (screenshot)

Depending on your browser, you may have to “save as” the files on your desktop.

3. PART 1: MAPPING LETTERS OVER EUROPE

3.1 Importing the data into GEPHI

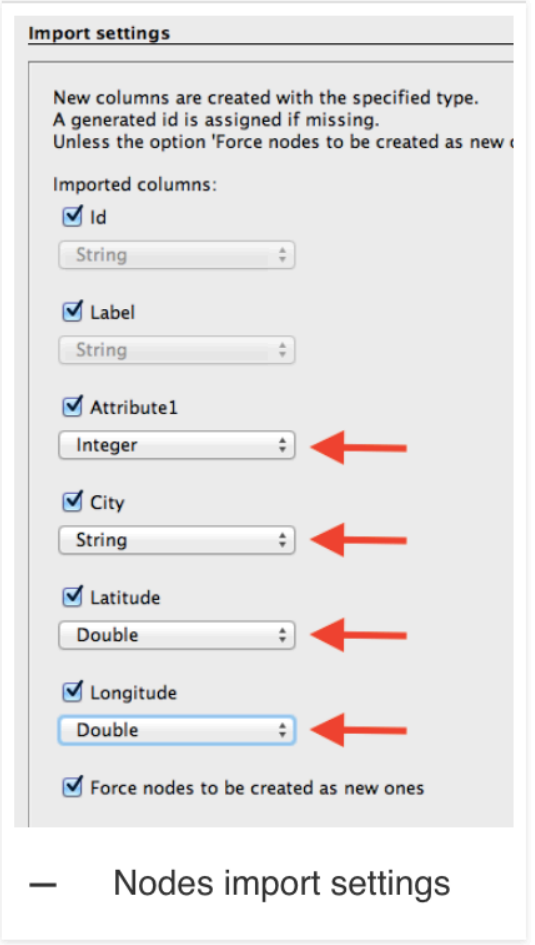
Run the software on your computer and create a “new project” in the start window. In the **Data Laboratory**, click on “Import Spreadsheet” to open the import window and import your first file.



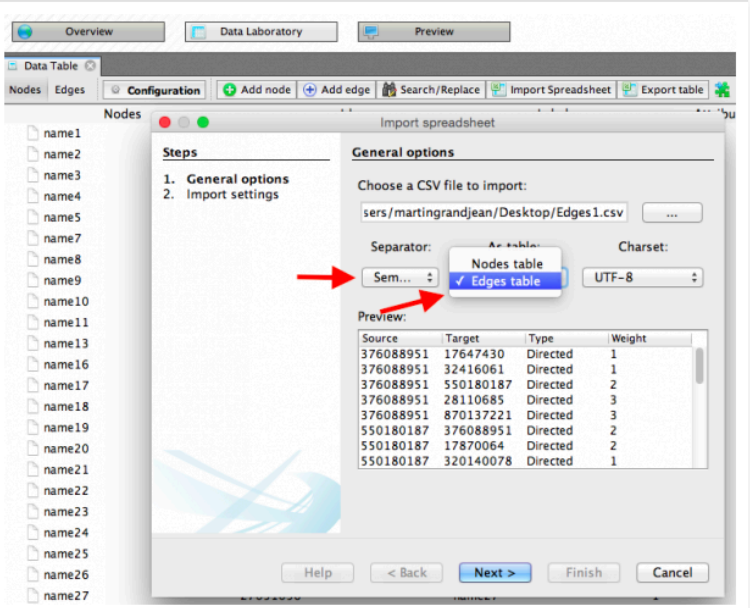
— Nodes import

Nodes 1

Specify that the separation between your columns is expressed by a **semicolon** and do not forget to inform Gephi that the file you import is containing **nodes**. Then press “next” and fill the import settings form as proposed. The “import settings” step is very important: Gephi will recognize some of the columns because of their header, but you’ll always have to check that the software will be able to understand the nature of your data. In our example, be sure to inform Gephi that **our latitudes and longitudes are a “double” variable** (not an “integer”).



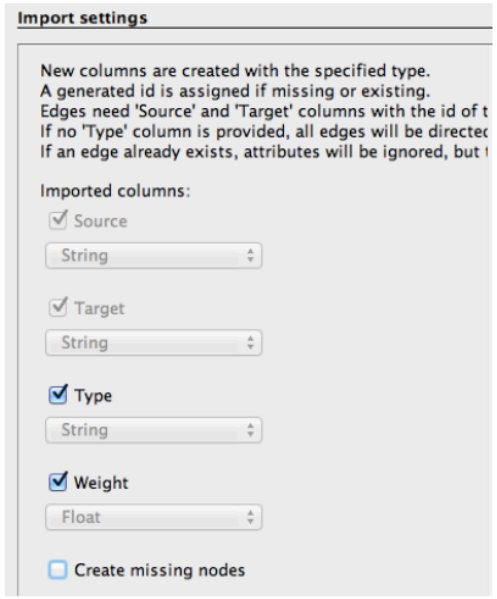
— Nodes import settings



— Edges import

Edges 1

Follow the same procedure, but with the “edges” file downloaded before and fill the forms in the following manner: specify the **semicolon** and inform Gephi that you’re importing the **edges**. Fill in the last fields and uncheck “create missing nodes”, because you’ve already imported them.



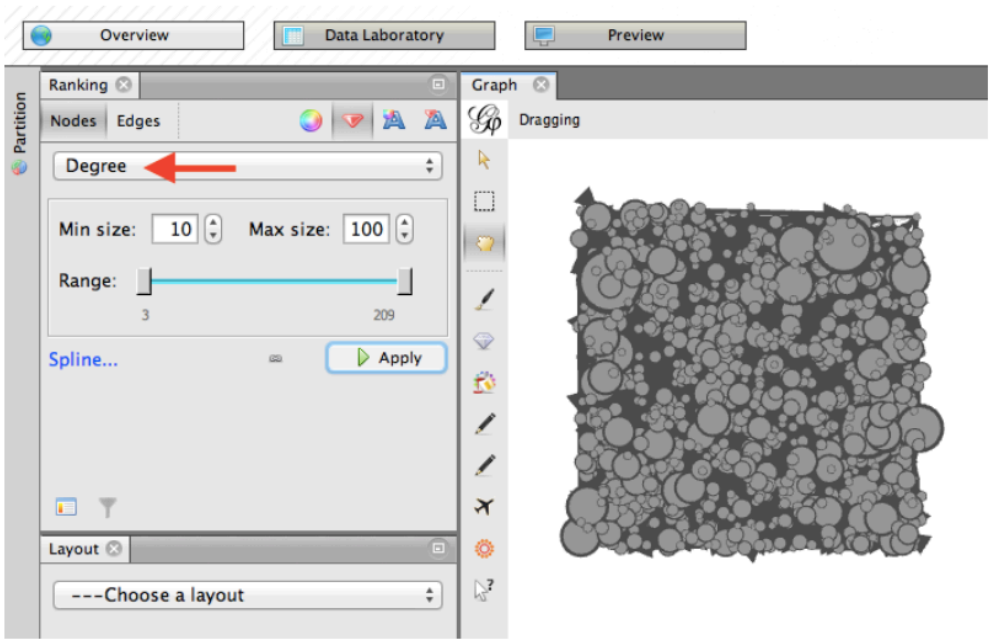
— Edges import settings

3.2 One-mode graph visualization

The action now takes place on the **Overview** panel. The software produces an overview of the graph, spatialized randomly and completely unreadable.

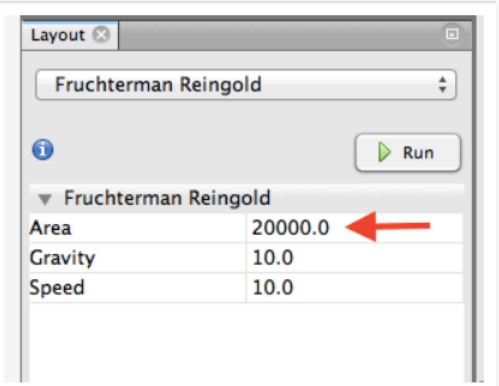
Nodes' size

Let's give nodes a size proportional to their degree (number of connexions). In the **Ranking** panel of the left column (top), select "**Nodes**" and the "**red diamond**", then select "**Degree**" in the rolling menu and enter the minimal and maximal value (we propose 10-100). You'll see that the distribution of degree within your corpus is between 3 and 209: at least one node is connected to more than 200 others (and the least connected node is connected to 3 of them). Be aware that if you want a visually correct result, you'll have to use the "Spline" blue link to edit the shape of the spline: linearly double the radius of a node is more than double the area because of the power function.



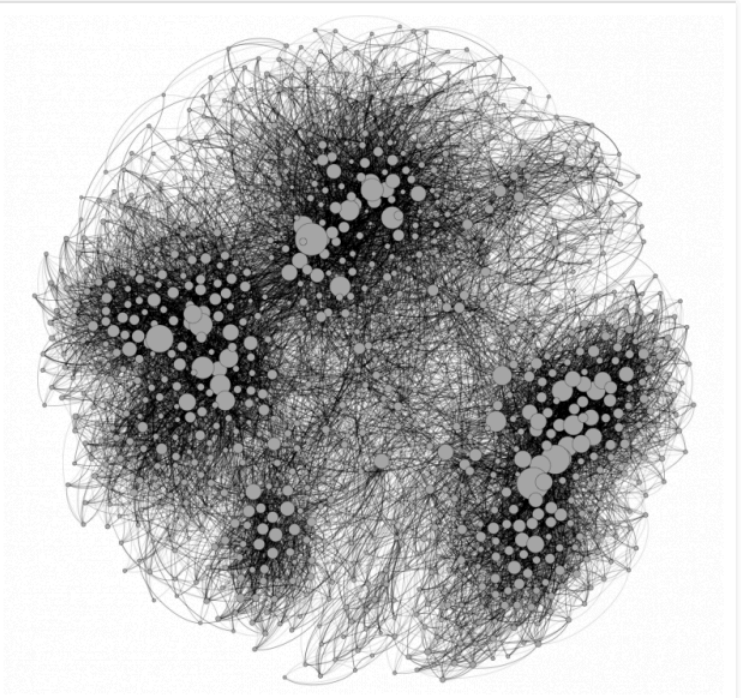
— Size settings

Spatialization

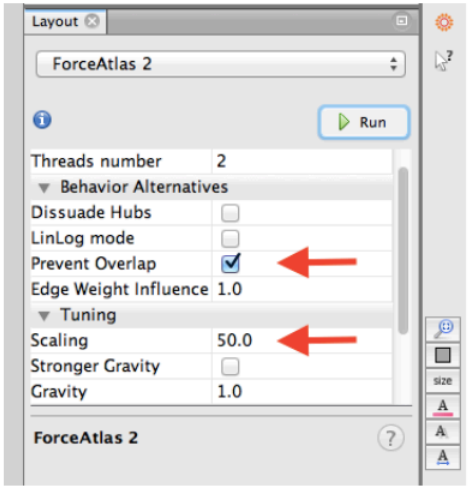


— Fruchterman Reingold

That's the main part! Let's begin with a spatialization that gives more space to the graph, but maintain it in a decided area: **Fruchterman Reingold**, with the same values as in this model (20.000 – 10 – 10). This visualization disposes nodes in a gravitational way (attraction-repulsion, in fact, as magnets). You're already able to distinguish communities (more densely connected parts of the network). Let the function run until the graph is stabilized. Use the **little blue magnifying glass** (bottom left of the graph panel) to re-center the zoom.

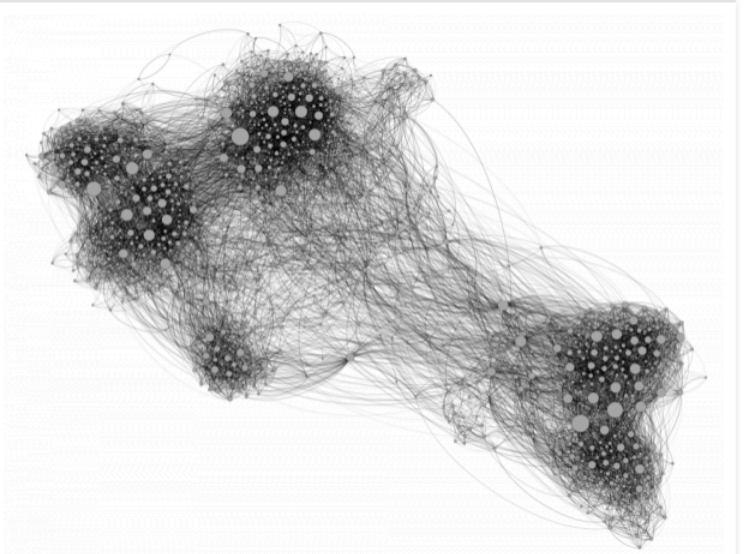


— Fruchterman Reingold



— Force Atlas 2

Then, we propose to use the **Force Atlas 2** (another layout algorithm) to disperse groups and give space around larger nodes. Be careful, the parameters you enter significantly alter the final appearance (proposition: Check "**prevent overlap**" and change "**Scaling**" to **50**). Let the function run until the graph is mostly stabilized. We can apply Force Atlas 2 directly without applying Fruchterman Reingold before, but as the "random layout" from the begining is a … *random* layout, it's better to untangle the network before sumitting it to a strong force-algorithm.

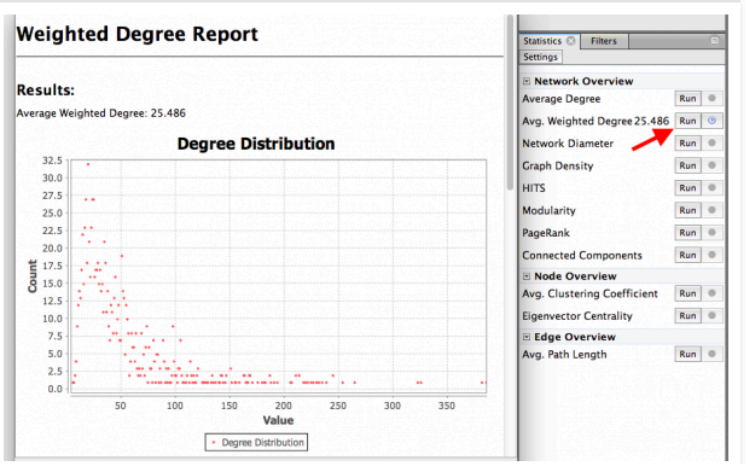


— Force Atlas 2

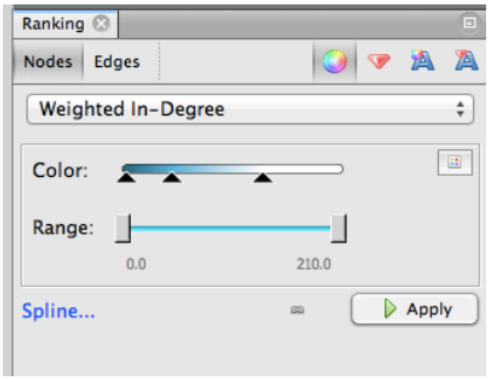
3.3 Final rendering and centrality measures

Weighted Degree

Let's add some more information to our graph by giving the nodes new attributes, influencing their color. In the **Data laboratory**, select the Edges Table, and sort them according to their wheight. Some edges have a wheight of 3, some 2 and some 1. That means that we have to take these differences into account by calculating the weighted degree of the nodes. You also observe that this graph is directed: the edges have a source and a target, a direction shown by a little arrow on the **Overview** display. So, the degree we'll have to calculate has to distinguish the in- and out- connexions. In the **Statistics panel**, click on "**Average Weighted Degree**" to calculate these values for every nodes. You get a report showing the distribution of theses measures.

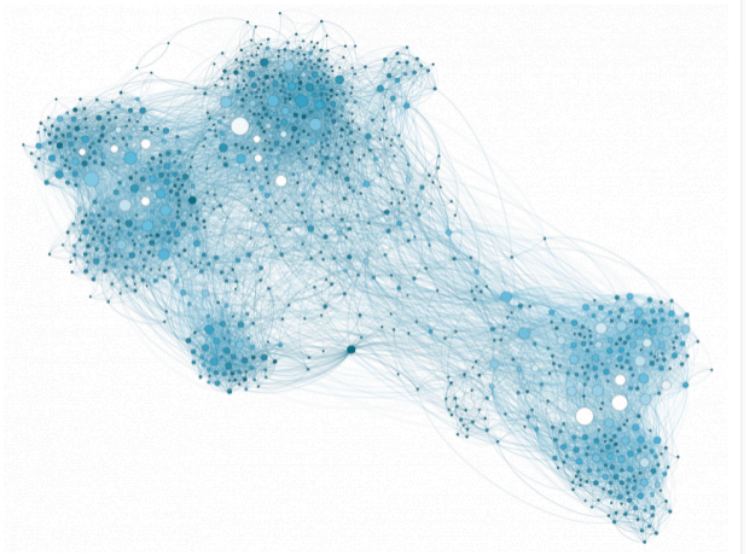


— Weighted degree distribution



— Nodes' color

Now that theses values are calculated, new attributes are available in the **ranking panel**. Select the "**color**" icon, and chose "**weighted in-degree**" to color nodes according to the number of incoming edges. Little visual tip : use a dark color for small values and a light color for the highly connected nodes, in order to make the little nodes visible on the final graph (the well connected nodes are generally more visible).



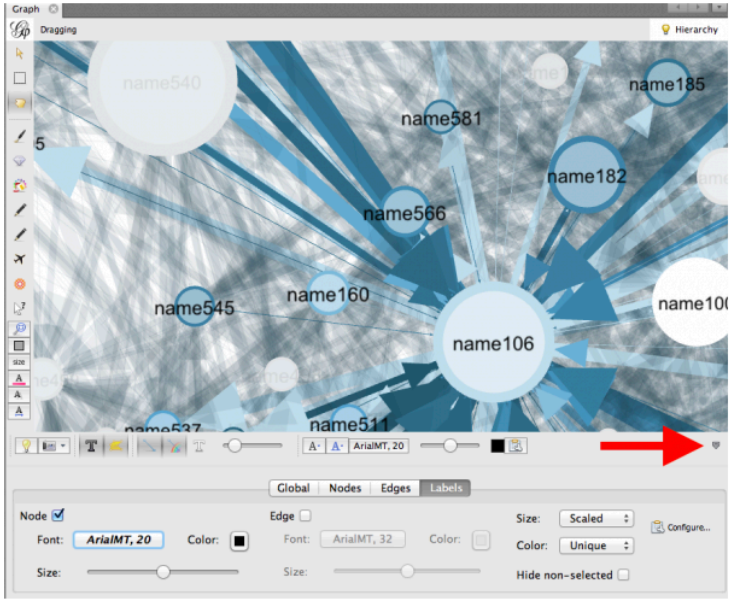
— Weighted In-Degree

Result: the biggest nodes (=with a high degree) are not always those with the biggest weighted in-degree : if we consider an edge like a letter written between 2 people, those who are writing a lot are not

Result: the biggest nodes (=with a high degree) are not always those with the biggest weighted in-degree : if we consider an edge like a letter written between 2 people, those who are writing a lot are not necessary those who are receiving a lot. It's interesting to give different attributes to nodes size and color, to compare them. Of course, you can export this data to conduct a full statistical analysis, scatter plots, etc. (the measures you make are automatically added to your nodes table). Note that if you used the "spline" to adjust nodes'size before, this setting is still used by default here and should be modified (without interfering with you previous choice for the size).

Nodes' label

We will come back to these measures and extra features after, but let's try to finalize our artwork for now by giving a label to the nodes. At the bottom right of the graph display, you'll find a **little sign** which allows you to developp a new panel. In **label**, choose "nodes" to add their labels to your nodes and set their font, color and size. If needed, for example if your data don't have any "Label" column, click on "configure" to set the column content you want to get displayed (the "ID" may be used as a label, i.e.).

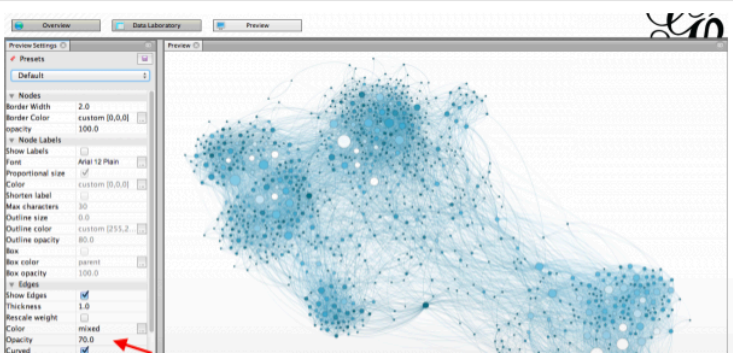


— Label settings

Finalizing the graph

Go to "**Preview**" for trimming the final details. Unlike during previous stages, changing settings in this menu is reversible, and do not affect the structure of the graph.

In the this screenshot, you will find **a suggestion of settings** for a good rendering (like setting the edges opacity to 70% for a better contrast with the nodes). Be aware that due to its large size, the graph may take a few seconds to update after each change (click on "refresh" to apply the changes). About curved edges : As a graphical convention, we use curved edges to show the direction of the edge, always turned clockwise. Non-curved edges are



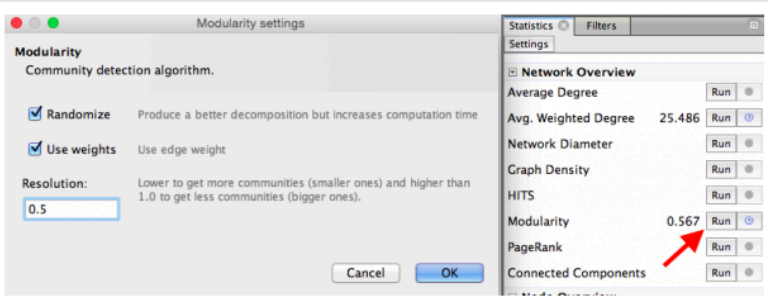
generally non-directed graphs.

— Preview menu

At the bottom of this preview column, you find an export link. Note that exporting in **.png** produces figure with a poor resolution. You may want to opt for **.svg** or **.pdf**, which have the advantage of being modifiable by your own image/drawing software (I recommend the open source program inkscape for manipulating .svg files).

Modularity

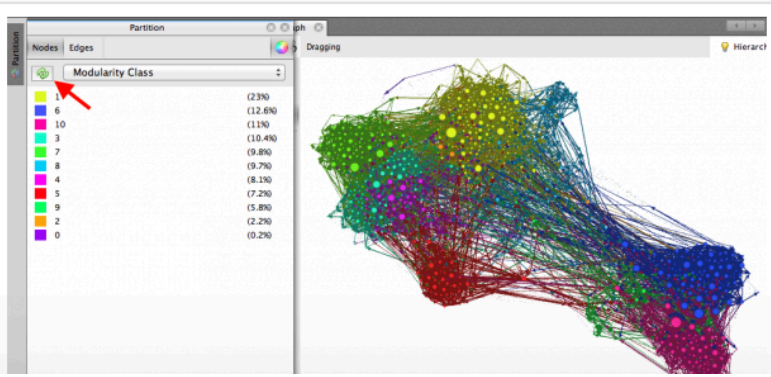
The visualization is only one step, network analysis often needs other mathematical means to provide the researcher with a satisfactory result. Feel free to explore the "**Statistics**" menu, for example by playing with degree measures, density, path length, modularity.



— Modularity settings

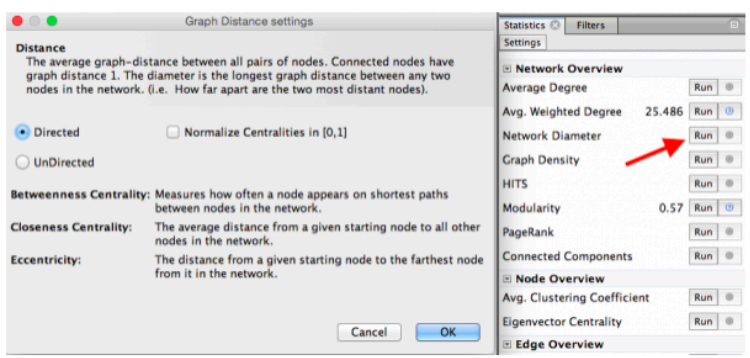
A network contains internal subdivisions called **communities**. There are methods that permit to highlight these communities, which depend on the comparison of the densities of edges within a group, and from the group towards the rest of the network (More here) In the right column of the "overview" page, click on **Statistics/Modularity/Run** to display the modularity window. Choose a resolution (between 0.1 and 2), click OK and close it.

The next step takes place in the **Partition** menu situated in the left column. Select "**Nodes**" and "**Modularity Class**" (rolling menu). You will be then able to modify the colors attributed to the detected communities by clicking on them. Do not hesitate to repeat this operation with many "Resolutions" ! If you decide to do so, you must deselect and reselect "Modularity Class" in the left column, and refresh color

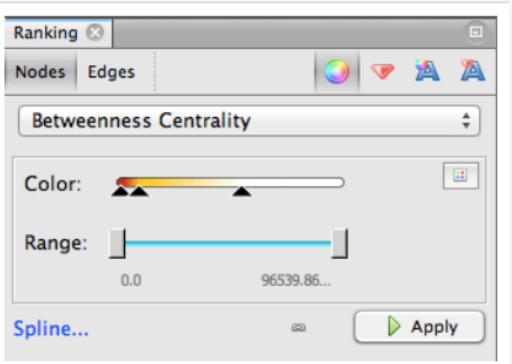


Betweenness centrality

The betweenness centrality measures all the shortest paths between every pairs of nodes of the network and then count how many times a node is on a shortest path between two others. It's a very interesting measure in the case of a network of letters sent and received as it allows the researcher to detect people that occupy an intermediate position between two other people or groups. In the **statistics panel**, click on "**Network Diameter**".

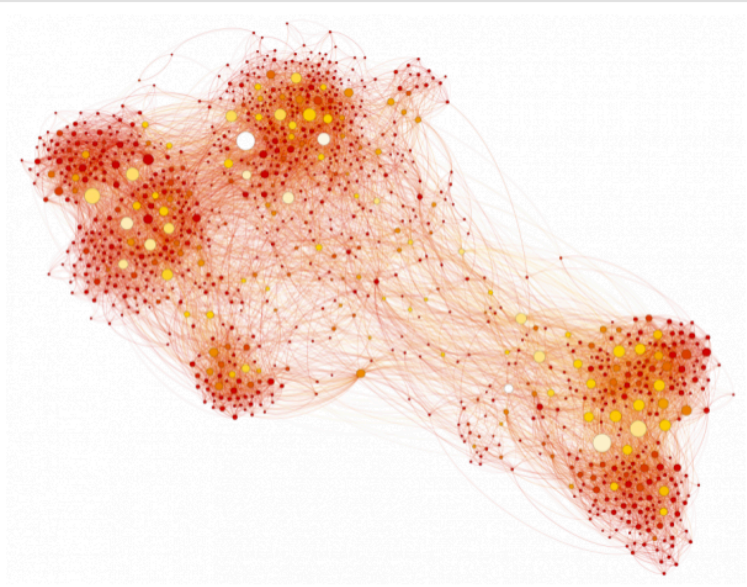


— Network Diameter



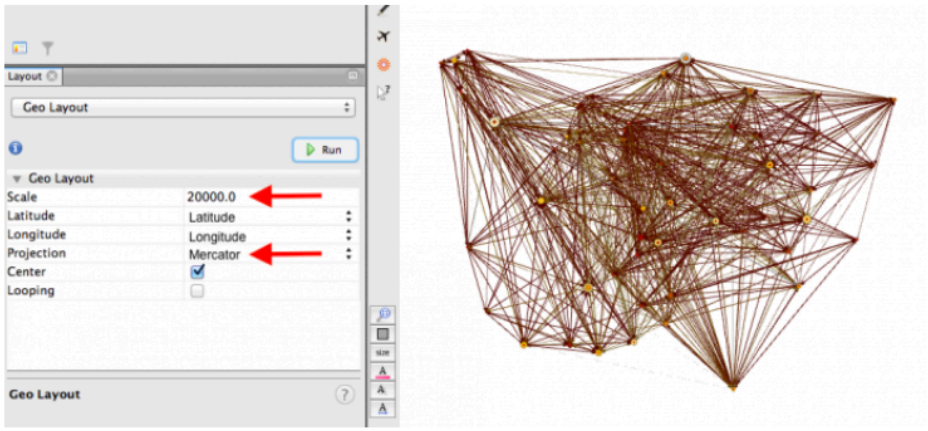
— Nodes' color

Like the Weighted In-Degree before, find a colorful way to highlight nodes that have a high Betweenness centrality. It quickly appear that nodes with a high degree/weighted degree does not always have a high betweenness.

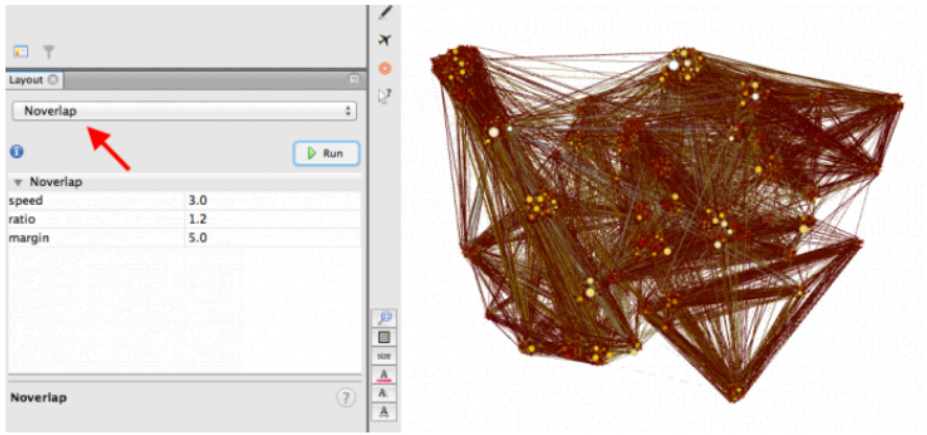


— Betweenness centrality

3.4 Geographical layout

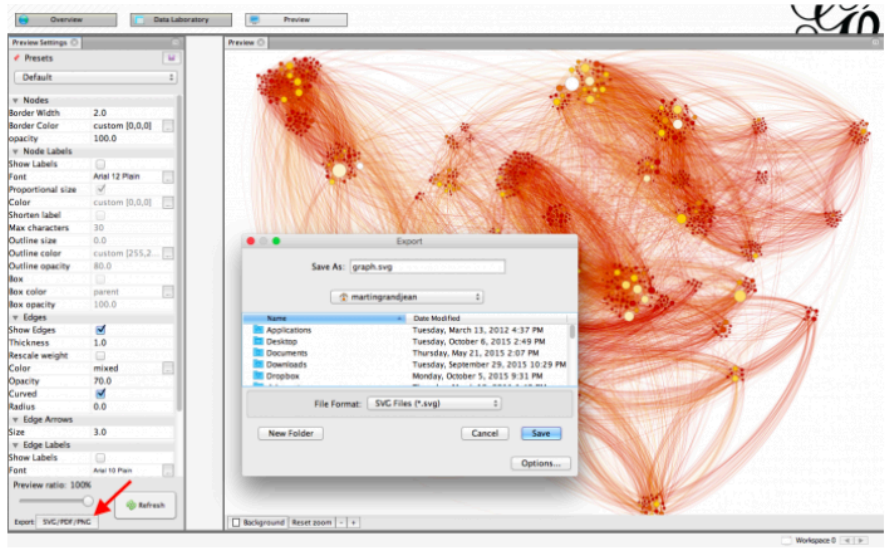


Geo Layout

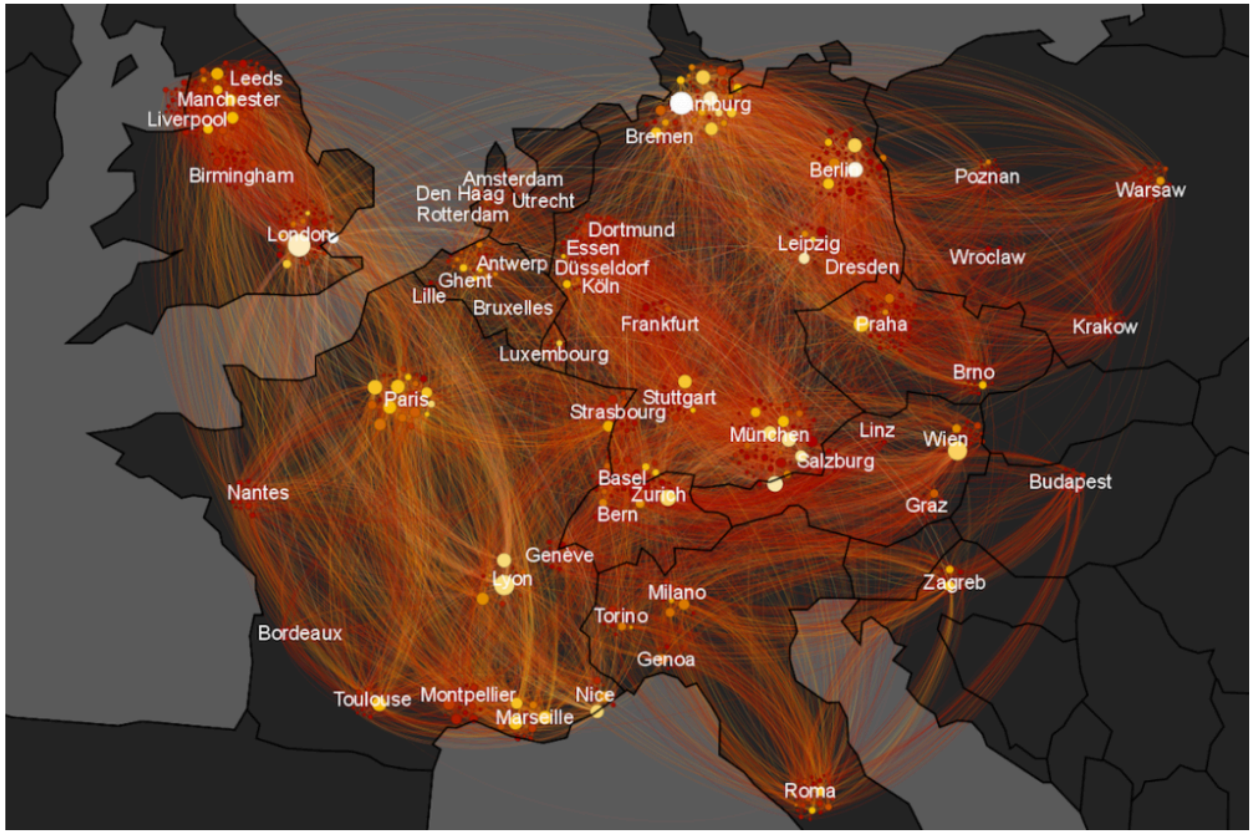


Noverlap

During the import, you've noticed that every node was given a Latitude and a Longitude. The **Geo Layout** plugin will help you display the nodes in a geographical way. In the Layout panel, select Geo Layout and give it a scale of 20.000. Be sure that the plugin understand correctly that "Latitude" as a "Latitude" and "Longitude" as a "Longitude" and set the projection to "**Mercator**" (this projection should be adapted to the map you'll use after). As nodes are now grouped on a geographical coordinate, you'll have to give them some space: use the **Noverlap** layout plugin to avoid them overlapping (a margin of 5.0 is enough with the chosen map scale).



— Preview and export



— Final map

In the **Preview** panel, check the final appearance of your artwork and export it in .svg. You'll then be able to import it on a background map. If you're familiar with Inkscape, download the map provided here (created to fit with the chosen scale and Mercator projection). Open it, and after having imported your network in it, select the city names layer and bring it to the front to make it readable.

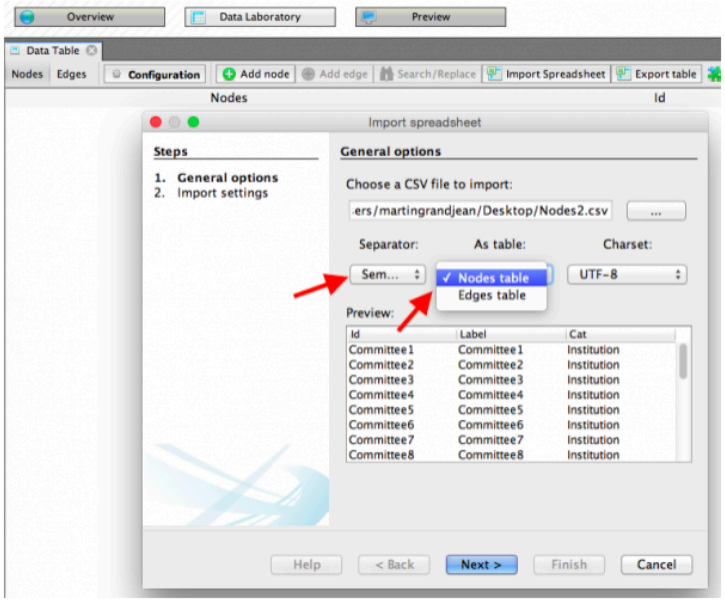
Map background

Feel free to try the same map with modularity, the result shows that communities are strongly related to geographic particularities.

4. PART 2: COMMITTEES AND THEIR MEMBERS

4.1 Importing the data into GEPHI

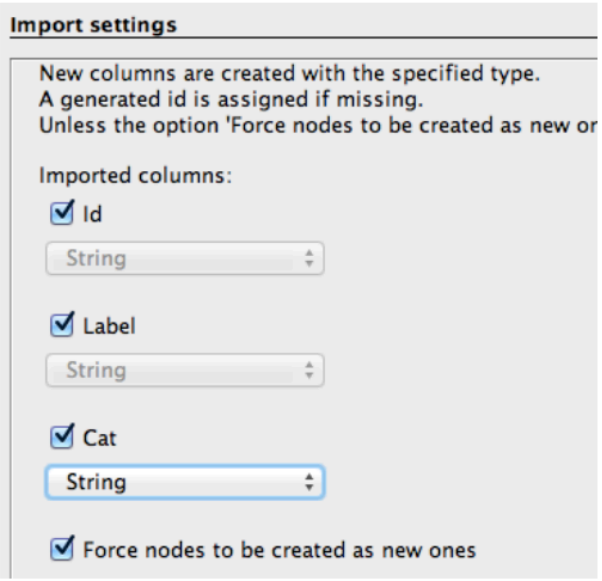
Create a “new project” in the start window. We’ll work on a different type of dataset: a 2-mode network (2 types of nodes, committees and individuals). In the **Data Laboratory**, click on “**Import Spreadsheet**” to open the import window and import your first file.



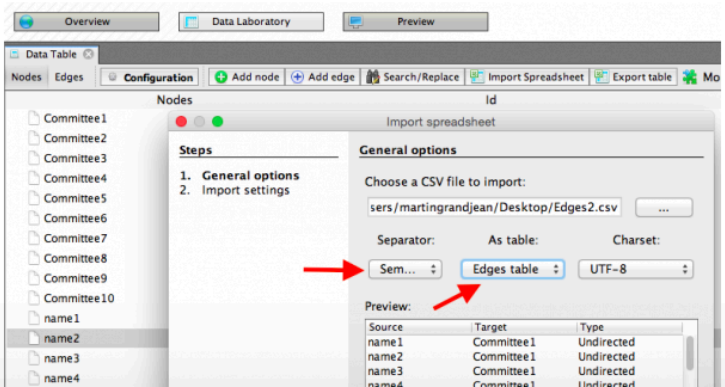
— Nodes import

Nodes 2

Specify that the separation between your columns is expressed by a **semicolon** and do not forget to inform Gephi that the file you import is containing **nodes**. Then press “next” and fill the import settings form as proposed. Inform Gephi that our “Cat” variable is a “String” (this variable will be useful to separate “members” and “committees” in a further step).

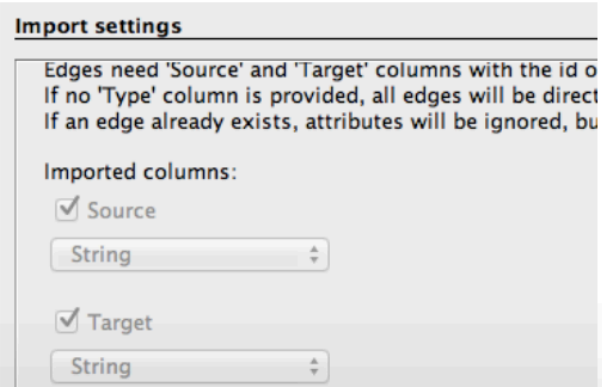


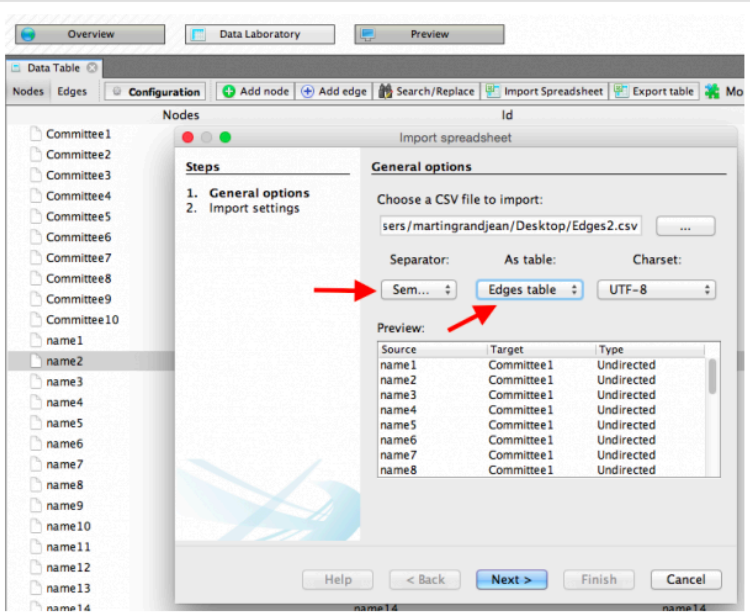
— Nodes import settings



Edges 2

Follow the same procedure, but with the “edges” file downloaded before and fill the forms in the following manner: specify the **semicolon** and inform Gephi that

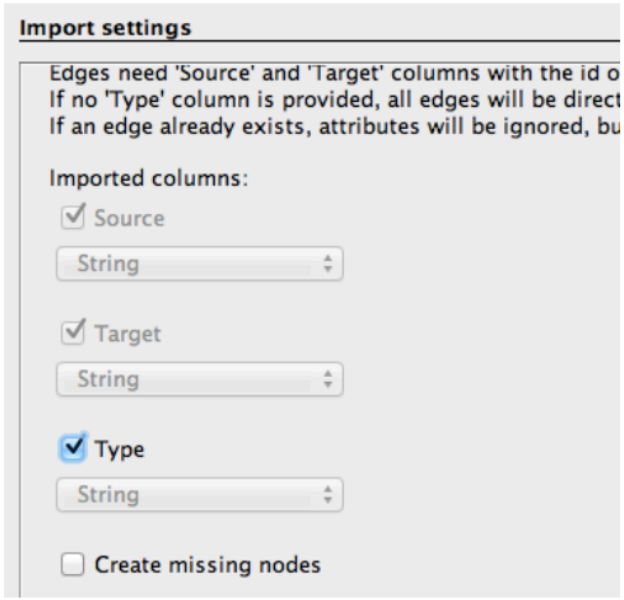




— Edges import

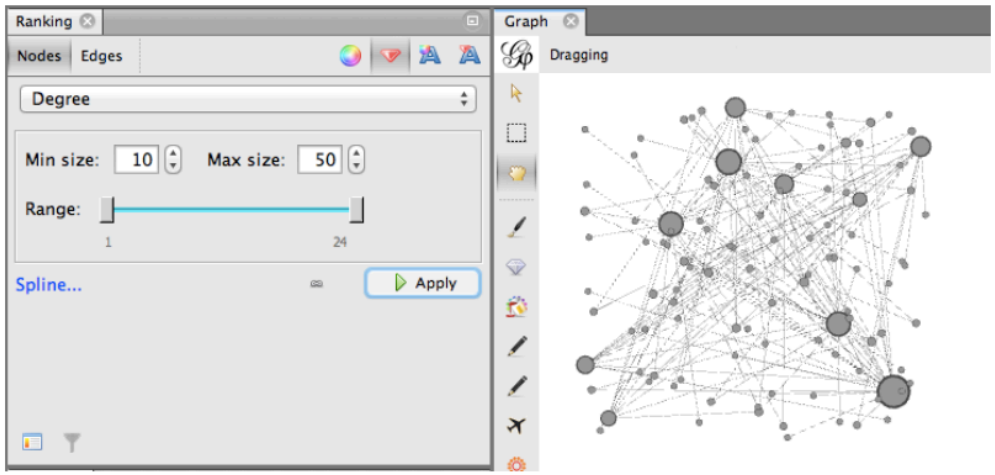
Edges 2

Follow the same procedure, but with the “edges” file downloaded before and fill the forms in the following manner: specify the **semicolon** and inform Gephi that you’re importing the **edges**. Fill in the last fields and uncheck “create missing nodes”, because you’ve already imported them.



— Edges import settings

4.2 Two-mode graph visualization



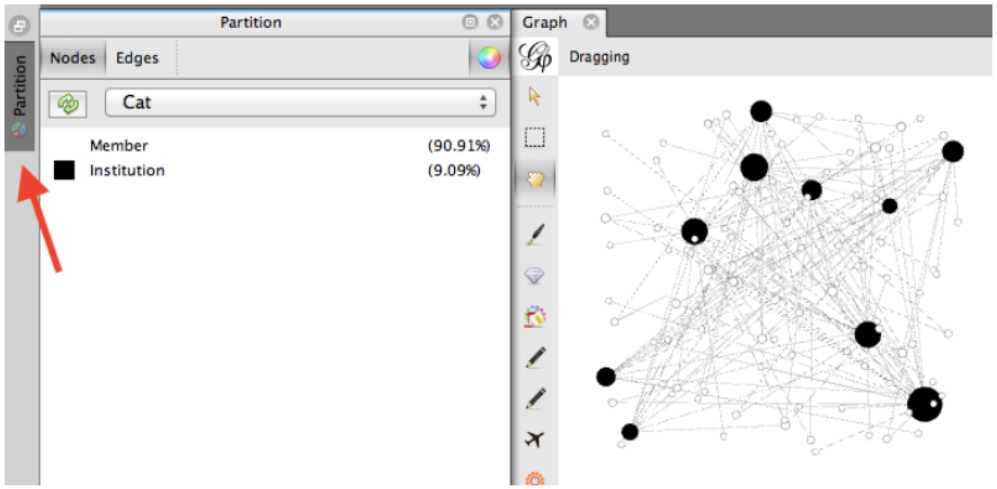
— Nodes’ size

Nodes’ size

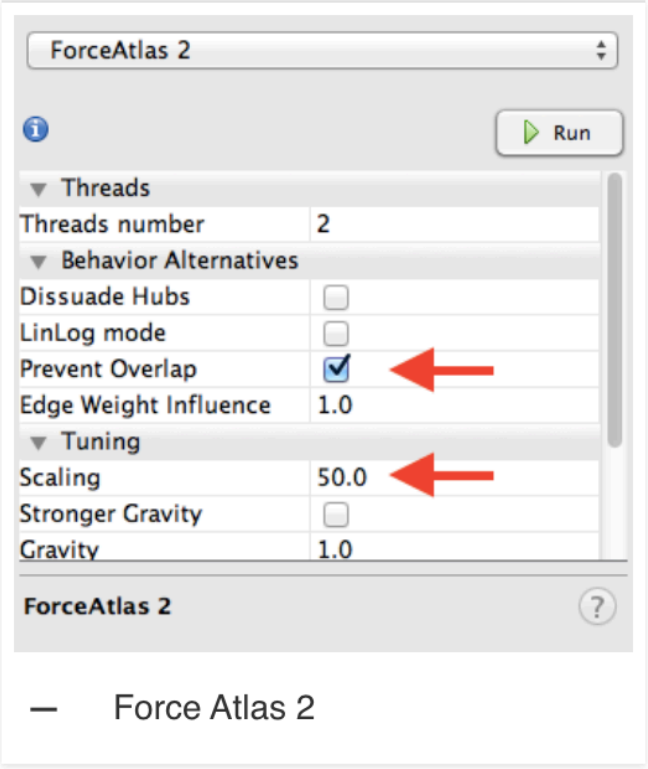
In the **Ranking** panel, give a size to your nodes (here, according to their degree between 10-50). In a 2-mode network, the degree centrality may not be a very interesting value, because of the structural bias brought by the two different categories of nodes: in our case, the “committees” will be naturally much more connected than the “members”. But in this first step, we’re just trying to visually distinguish the 2 categories.

Nodes's color

In the **Partition** panel, refresh the menu to make the nodes' attributes appear (we uploaded only one attribute: "Cat"). Give a very different color to both categories and apply it on your network.



— Nodes' color (partition)

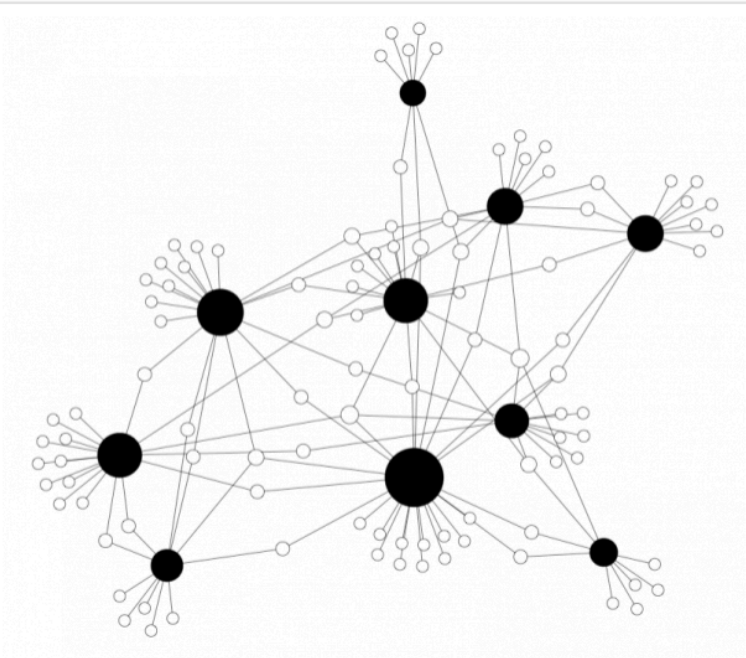


— Force Atlas 2

Set a layout

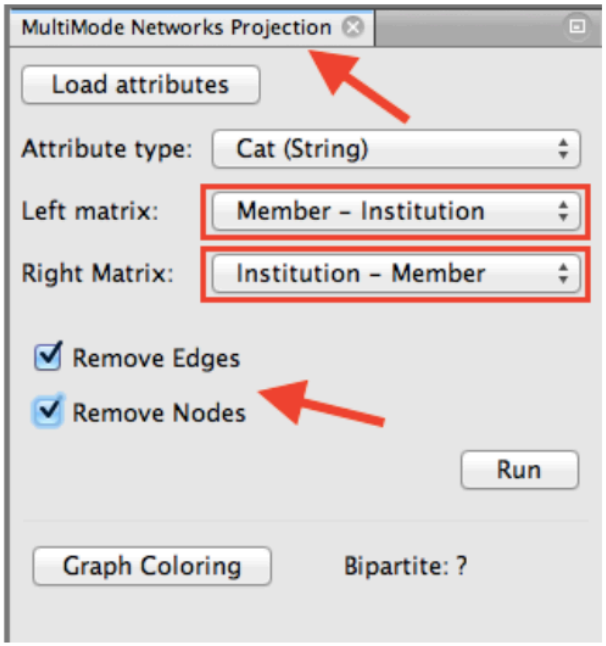
Deploy the network using the **Force Atlas 2** algorithm (Prevent node overlapping and scale it to 50). Your graph is now visually readable and looks very similar to many organizations networks.

For many researchers, this visualization will be already enough to conduct their analysis. Don't forget to display the nodes' label if needed.



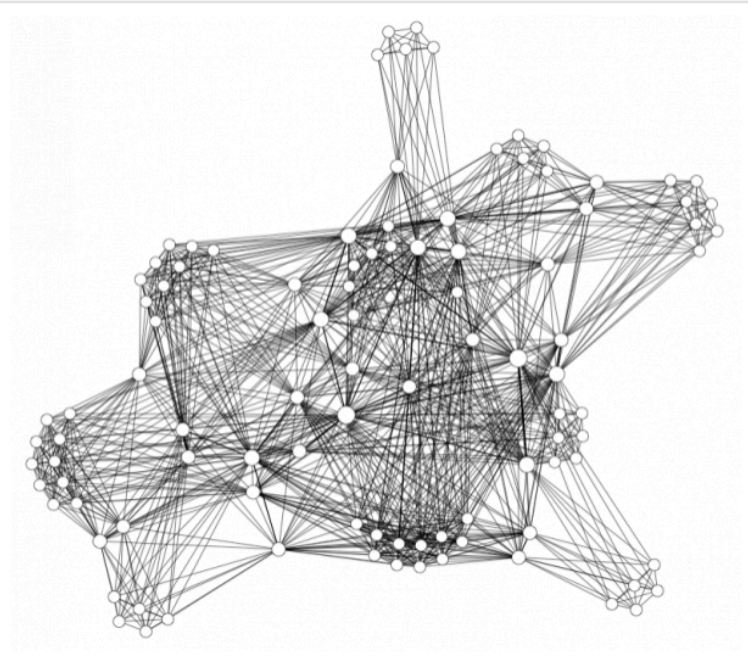
— 2-mode network

4.3 Projection to one-mode graph



— Projection plugin

Use the **MultiMode Networks Projection** Panel (available through the plugin you dowloaded in step 2.2) and “load attributes”. You’ll now “project” the Institutions on the Members: if two members have an edge linking them with the same committee, they’ll now have a direct edge between them (and the committee will be evacuated).

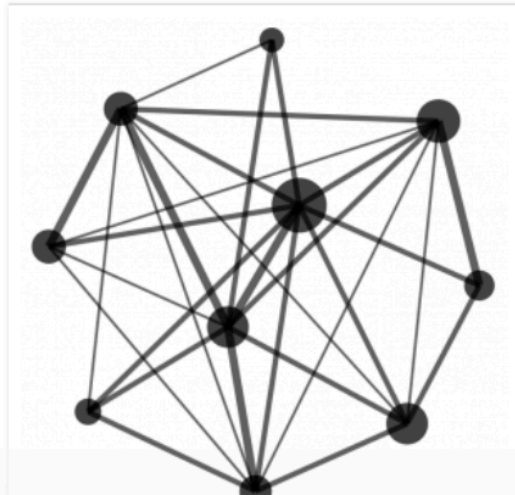


— Projected graph

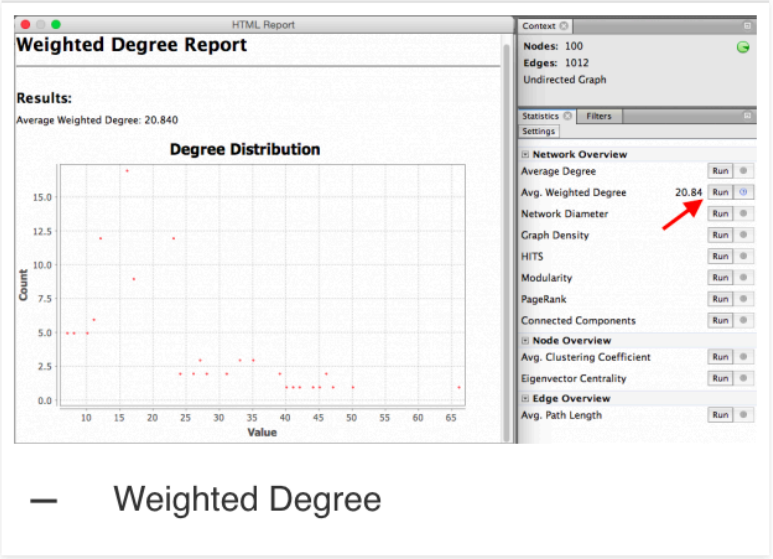
Select the right attribute type (“Cat”), and **set the matrix as proposed here** (Member-Institution / Institution-Member): They must be symmetric with the type of node you want to keep at the beginning and the end.

Check the “Remove Edges” and “Remove Nodes” buttons, in order to clean the graph from the old “Committees” nodes and edges. And finally click on “Run”.

Note that you can also project the Members on the Institutions, with the result presented here on the right (edges are getting larger if many members were connected in the same committees).



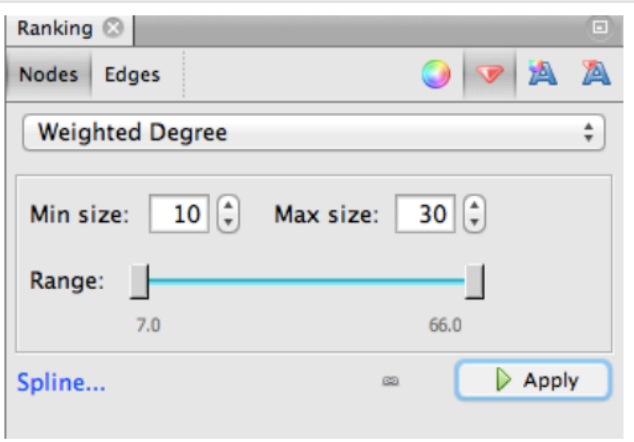
4.4 Centrality measures and layout



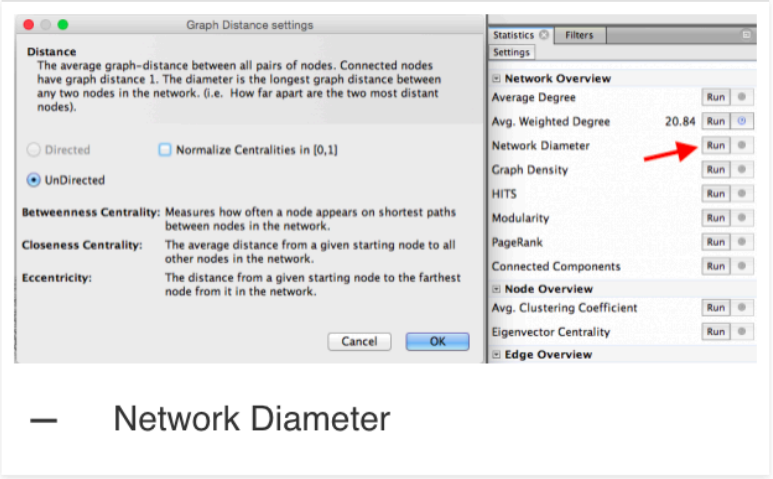
— Weighted Degree

Nodes' size

Calculate the new Degree centrality of the nodes by clicking on "**Avg. Weighted Degree**" (Statistics panel). In the **Ranking** Panel, apply this new measure to the nodes, as proposed here. The new degree may be very different from the degree in the 2-mode original network: a projection add lots of edges (in particular when lots of nodes where connected to a few very central nodes from the other type).



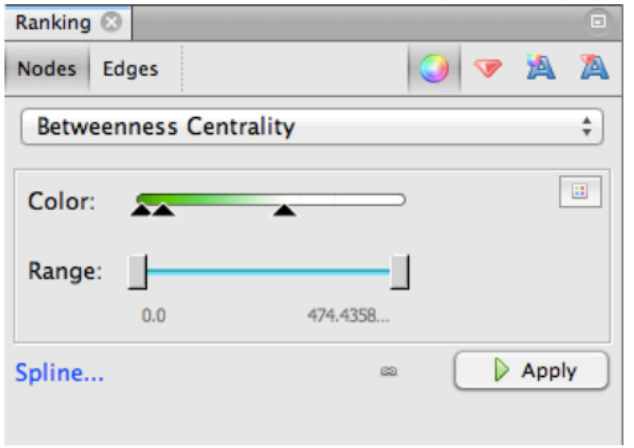
— Nodes' size



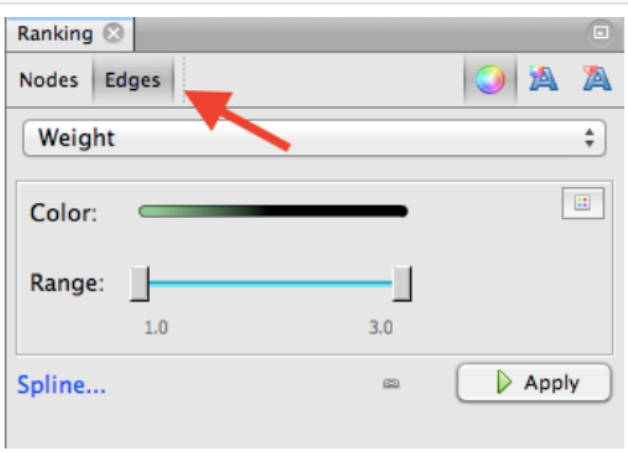
— Network Diameter

Nodes' color

In the statistics panel, click on "**Network Diameter**" to calculate the **Betweenness centrality** of your nodes. Then use this measure to color the nodes. In such a network of people working in different committees/institutions/companies, knowing who's at the intersection of two groups may be very important for HR officers, i.e..



— Nodes' color



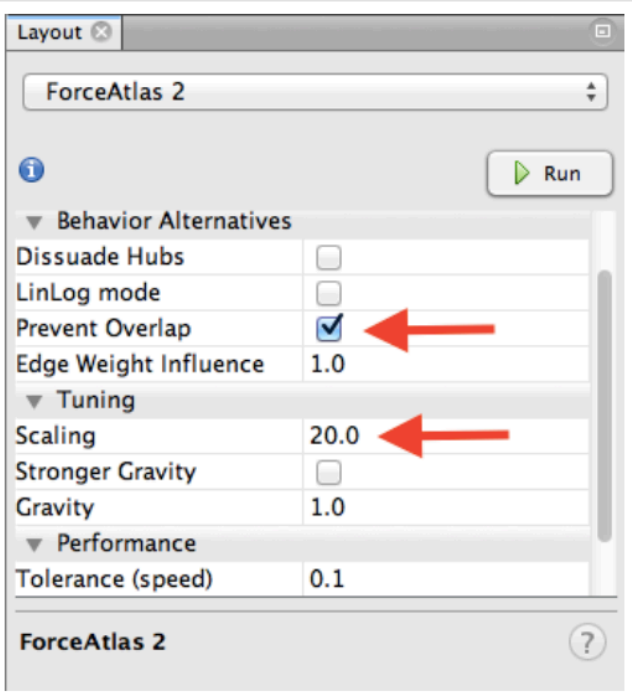
— Edges' color

Edges' color

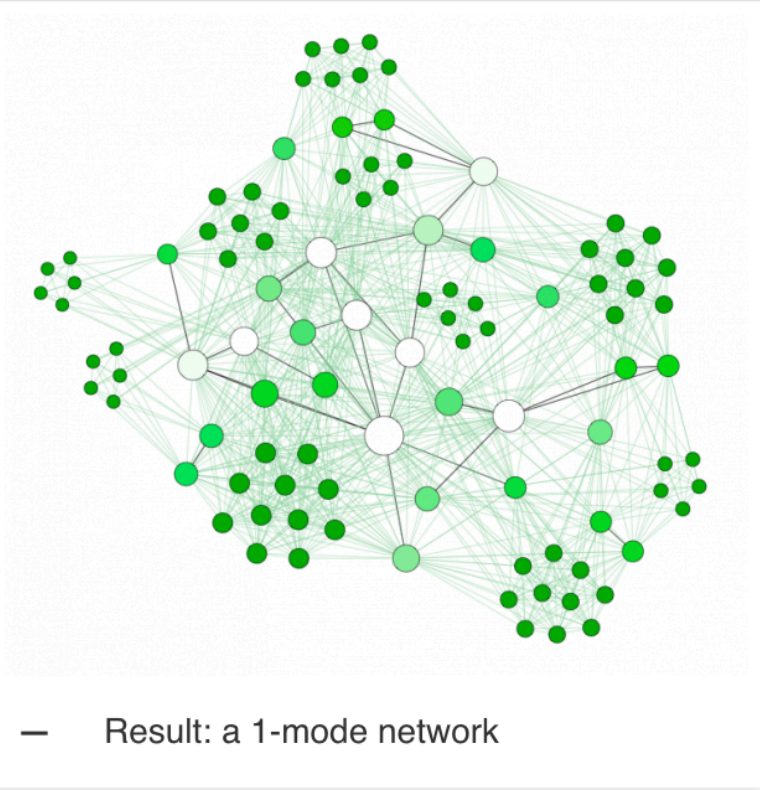
In order to highlight weighted edges, give them a color that will make the stronger edges more visible in your final display (Suggested here: black for all the edges bigger than 1).

Layout

Spatialize the graph once again (it kept the positions of the nodes before the projection from 2-mode to 1-mode), with Force Atlas 2.

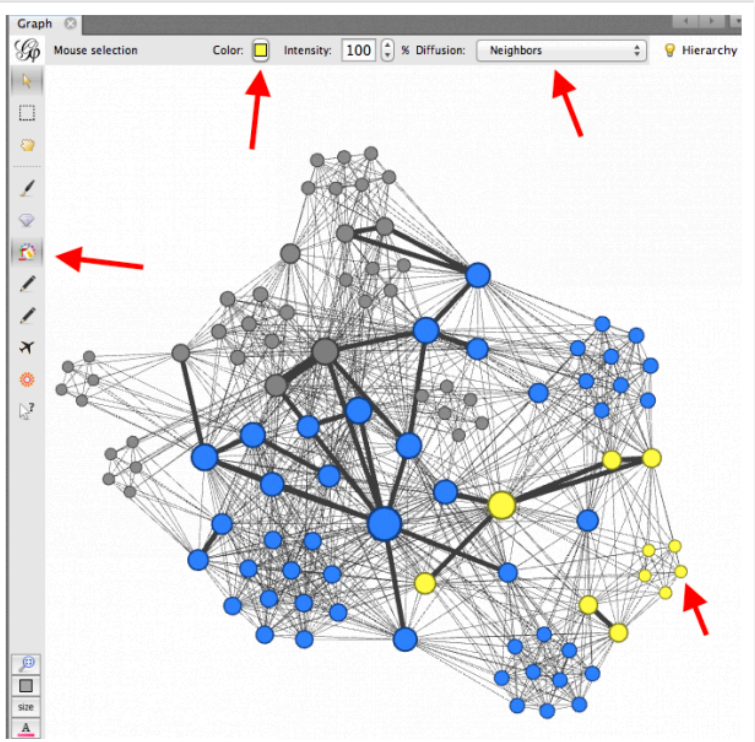


— Force Atlas 2



— Result: a 1-mode network

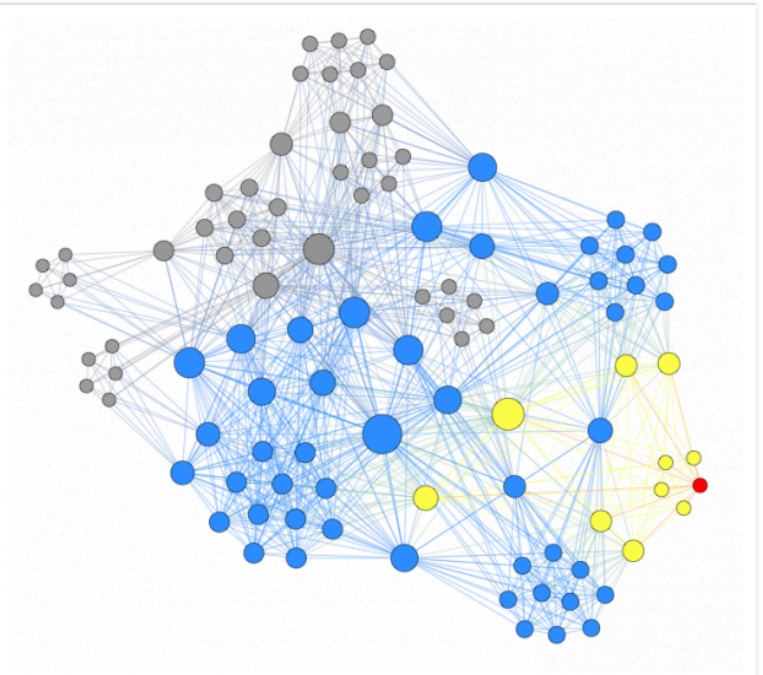
4.5 Neighbors highlighting



— “Paint bucket” tool

This type of network is well suited to a “Linkedin” of analysis: Who’s in my network? Who are the people that I will be able to reach through them (what are their own connections)?

Click on the **little paint bucket**, on the left of the Graph area, and play with the tools on the top of this menu. First paint the “Neighbors of neighbors” (after having given a neutral color to all the nodes), and then the “Neighbors” of a selected node. In our example, the red node, member of only one committee, is directly connected to 10 colleagues, which are themselves connected to 49 other individuals.



— Neighbors and neighbors of neighbors

5. CONCLUSION

Data visualization is a game, let’s play! Please help me to improve this tutorial by dropping a comment below with remarks, suggestions, links to your own results, etc.!