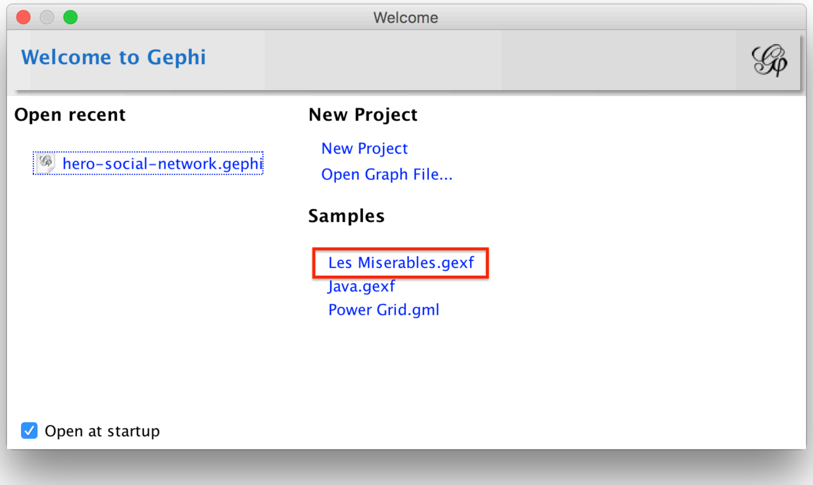
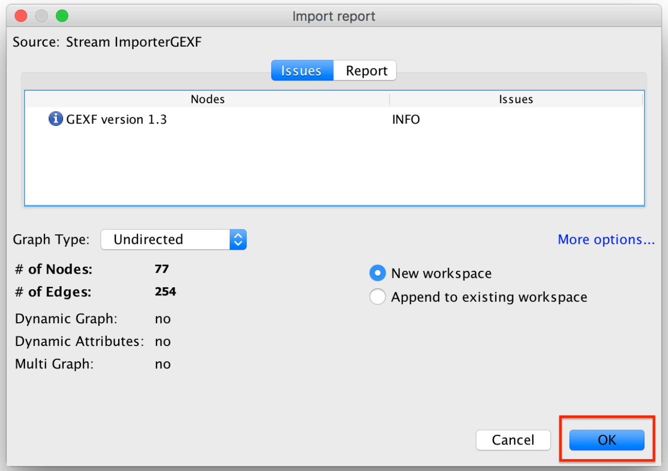
Analyzing and visualizing network data with Gephi

Pablo Barberá

In today's seminar, we will learn the basics of applied social network analysis using the Gephi software. If you haven't done so already, make sure you install the most recent version of Gephi (0.9.2) from the website (www.gephi.org).

After you open Gephi, you should get a welcome screen. (If you don't see it, go to "Window > Welcome" in the drop-down menus.) For now, we will open one of the datasets that come installed with Gephi to help us become familiar with the software. Click on "**Les Miserables.gexf**". In the window that opens after that, click on **OK**. This file contains information about the network of characters in Les Miserables. Each node is a character and the edges represent the number of chapters in which each pair of characters appear together.





Source: Stream ImporterGEXF

Nodes	Issues
GEXF version 1.3	INFO

Graph Type: Undirected

# of Nodes: 77

# of Edges: 254

Dynamic Graph: no

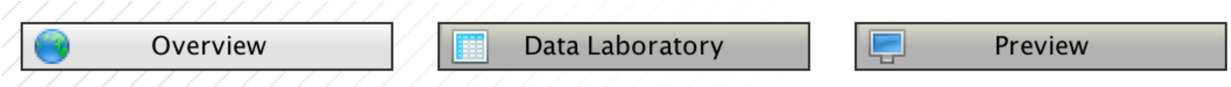
Dynamic Attributes: no

Multi Graph: no

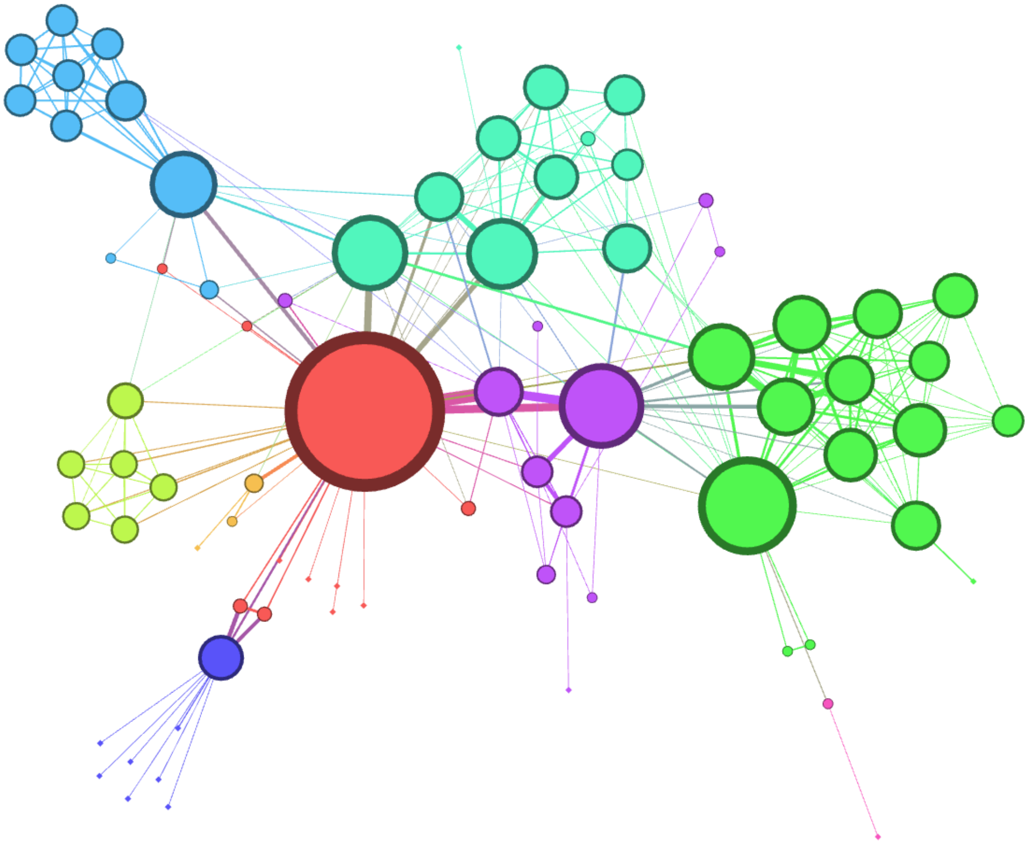
● New workspace

○ Append to existing workspace

Gephi splits the three main tasks in network analysis (data management, data analysis, and visualization) into three different tabs, which you can find at the top of the screen: Overview, Data Laboratory, and Preview.

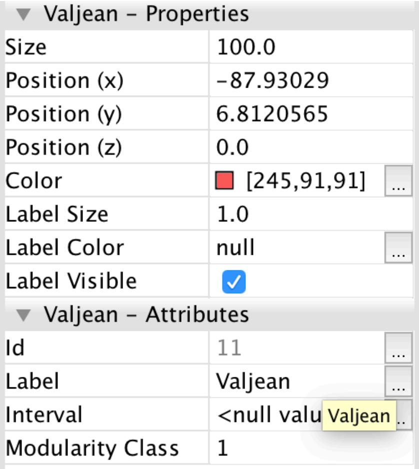


We'll start with Overview, which is the main tab for data analysis. You should be able to see the network, and a range of buttons and menus (many more than what we need for today's class!) to play around with it. The top-right panel tells you the number of nodes and edges in the networks, and the type of network it is. But of course the most important panel here is the Graph panel, in which you should see the following:



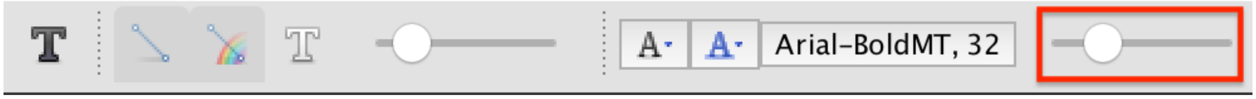
You can zoom in and out with your mouse. If you get lost, you can always return to the main view by clicking the following button:

Just looking at the structure of the graph, we already learn a lot – we see high levels of clustering and the node in red appears to be very central. How can we learn the character that each node represents? One option is to focus on a specific node of interest, which we can do by clicking on and then on the node of interest. This should make the top-left panel change and display all the metadata for that particular node. For example, if you click on the largest node, you should see:

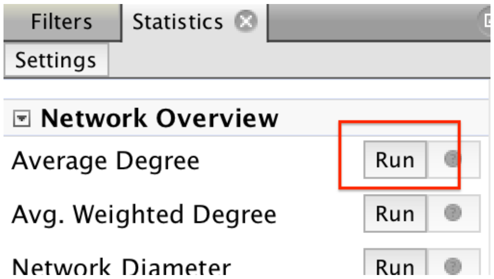


Valjean – Properties	
Size	100.0
Position (x)	-87.93029
Position (y)	6.8120565
Position (z)	0.0
Color	🟥 [245,91,91]
Label Size	1.0
Label Color	null
Label Visible	☑
Valjean – Attributes	
Id	11
Label	Valjean
Interval	<null valu Valjean
Modularity Class	1

This node is Valjean, which is the main protagonist of the book, so no surprises here. You can also find out who each node is by making the figure display all the node names. How to do that? Click on the capital T icon at the bottom of the screen: T You may have to change the size of the labels with the slider all the way to the right (highlighted below), which will make the labels smaller or bigger:

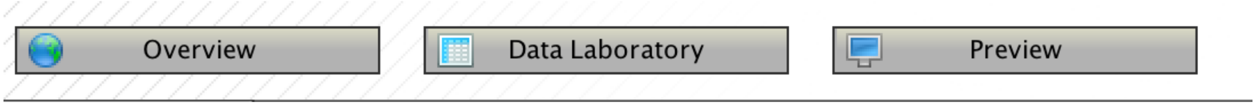


The last panel we will focus on while here is the "Statistics" panel on the right. This panel offers a list of analyses you can run on this network dataset. For now, let's focus on computing centrality. To do so, click on the "Run" button next to "Average Degree".



A new window will pop up, but you can ignore it for now and just click on "Close". Now, do the same with "Avg. Path Length", a little bit further below the list. Click on "OK" and again "Close" on the new window that pops up. What have done here? We have computed degree centrality and betweenness centrality; as we turn to the next tab, we'll see where the results are.

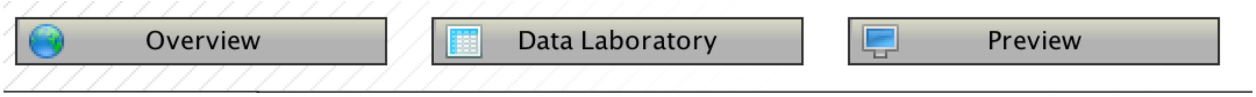
Now it's time to move to the Data Laboratory.



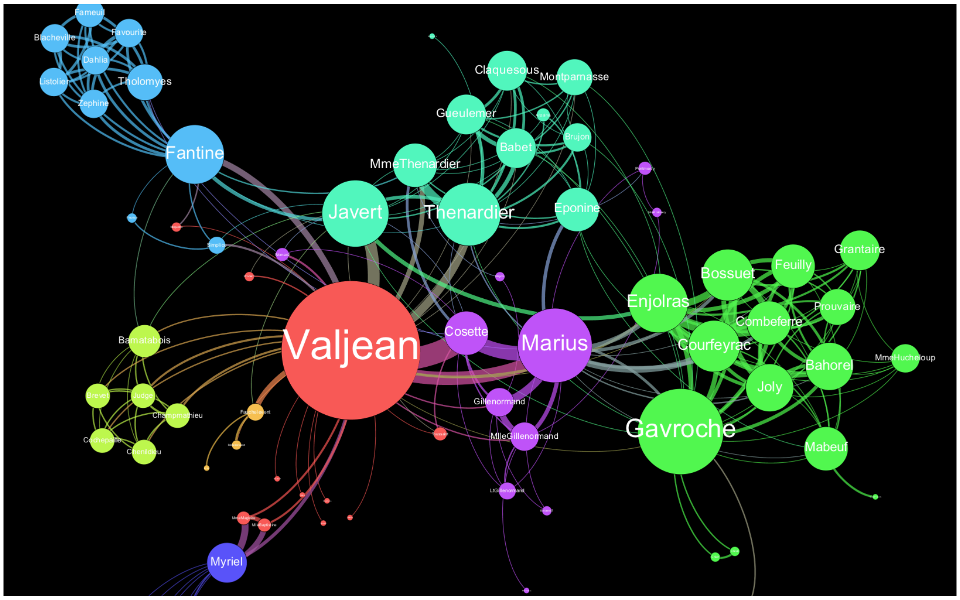
This tab displays the network data in two different ways. If you click on "Nodes" you will see the list of characters in the book, along with any metadata available, including the centrality measures we just computed. Let's sort the nodes based on Degree, for example, by clicking on the Degree column title. Then, do the same but for Betweenness Centrality. Which is the most central character? According to both measures, that would be Valjean, which confirms our previous analysis.

If you go to "Edges" you will see the list of connections in the network. Each row represents a potential combination of two characters (represented here as numeric IDs in the Source and Target columns), as long as the number of chapters where they appear together (the Weight column).

Finally, let's switch to the Preview tab.

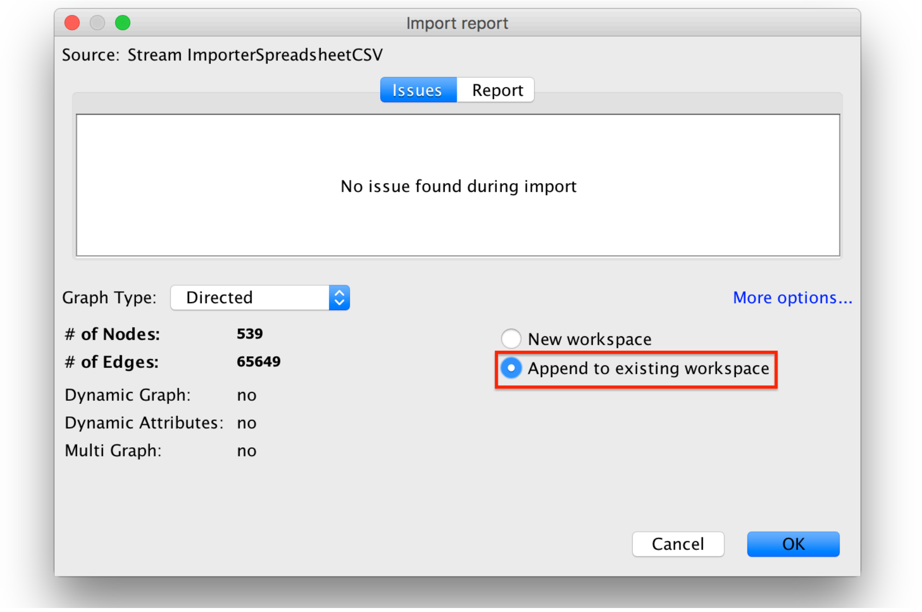


This tab offers a range of options to visualize the network in different ways. To see it displayed in the screen click on "Refresh" at the bottom of the panel. Cool, huh? You can choose a different Preset (for example, Black Background and then changing the Node Label Font to something smaller) to make it look different. Make sure you click Refresh every time you change the options. If you want to export the network visualization to a file in your computer, click on "Export" and follow the options. Here's mine:



To continue practicing our network analysis skills, we'll now open a different dataset that I collected from Twitter. It represents the Twitter network of the Members of the UK House of Commons who have a Twitter account, with each edge indicating whether one Member follows a different Member on Twitter.

To open this dataset, first go to File > New Project in the drop-down menus. You may want to save first your current work (although you can always re-do all the steps above as additional practice!). Once you have cleared all the windows, go to Data Laboratory and click on "Import Spreadsheet". We'll start importing the list of edges, so go ahead and select the file "UK-twitter-edges.csv" in the folder where you downloaded the data. Then click on "Next", then "Finish" with the default options. Important: in the Import report window, make sure you select "Append to existing workspace":



Source: Stream ImporterSpreadsheetCSV

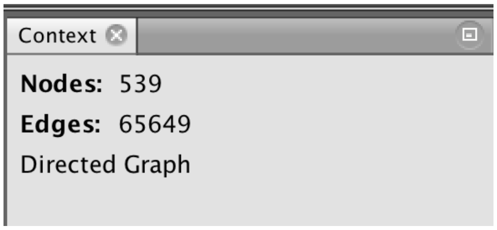
No issue found during import

Graph Type: Directed

# of Nodes: 539
# of Edges: 65649
Dynamic Graph: no
Dynamic Attributes: no
Multi Graph: no

○ New workspace
● Append to existing workspace

Then click "OK". If the import process was correct, you should see the following in the Overview tab:



Nodes: 539
Edges: 65649
Directed Graph

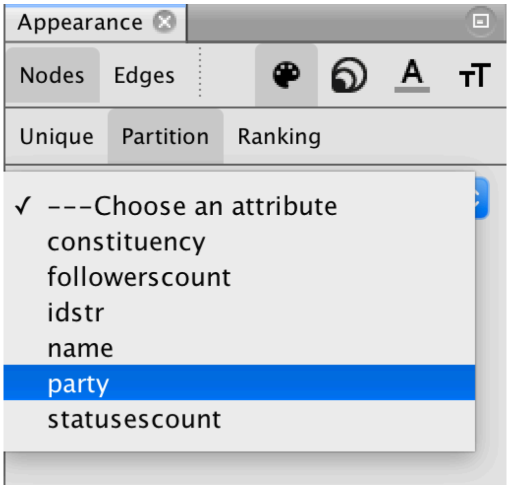
Now, we'll import the node metadata. Go back to Data Laboratory and click again on "Import Spreadsheet". This time, select the file "UK-twitter-nodes.csv" and follow the same instructions, again making sure you select "Append to existing workspace" so that Gephi knows the two files corresponds to the same network.

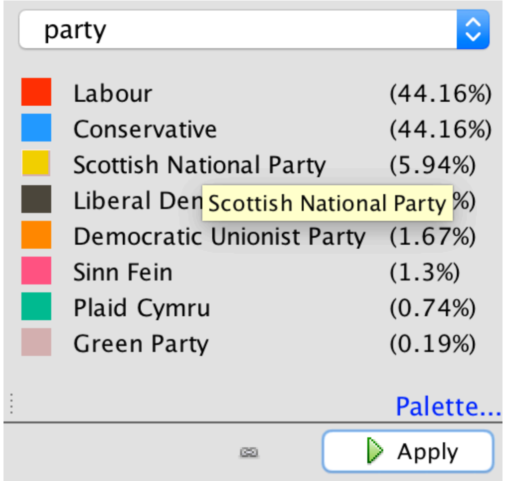
In Data Laboratory, if you go to the Nodes tab, you will see the metadata available for each legislator: Twitter handle, name, constituency, party, Twitter ID, number of followers, and number of tweets sent.

Q1. Using these last two columns, find out who are the legislators that have the most followers and those that have sent the most tweets.

Now let's switch to the Overview panel. This time the network doesn't look as great, at least in the default options. So let's try to fix that.

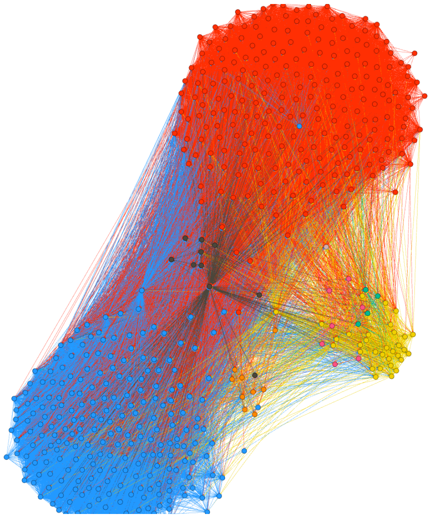
First, let's give each node a color based on the party. To do so, select "Partition" in the top left panel; in "Choose an attribute", select "party". Then, click on each of the colors and drag the pointer to the color you want (e.g. red for Labour, blue for Conservative). Once you've chosen color you like, click on "Apply".




party

🟥	Labour	(44.16%)
🟦	Conservative	(44.16%)
🟨	Scottish National Party	(5.94%)
⬛	Liberal Den	%)
🟧	Democratic Unionist Party	(1.67%)
🟥	Sinn Fein	(1.3%)
🟩	Plaid Cymru	(0.74%)
🟫	Green Party	(0.19%)

Ok, that should have helped a bit, but we still cannot see much in this network. We need to play around with the layout, in the bottom-left panel. A layout algorithm is a set of rules to choose where to place the nodes in the graph so that the visualization is more effective. Feel free to try a few different layouts. For networks like these, ForceAtlas2 tends to be the best one. (Run it for a few seconds and then stop it.) You can then zoom out to see what it looks like; hopefully something like the following:



Q2. What do we learn from looking at the structure of this network?

Q3. Do you notice anything unusual in the group of Labour MPs? Use the tools we used above to identify individual nodes to find out more about what could be happening.

Q4. Compute the measures of degree centrality and betweenness centrality for the nodes in this network, using the steps described above. Among the top 10 legislators with highest betweenness centrality, do you find anything unexpected (i.e. low degree but high betweenness?). Can you guess which node does this legislator correspond to in the network plot?

Q5. Explore the options in the Preview panel and produce a visualization as informative as you can. (If producing the graph takes too long each time, you can reduce the "Preview ratio" to something lower to speed up the process.) Here's mine: