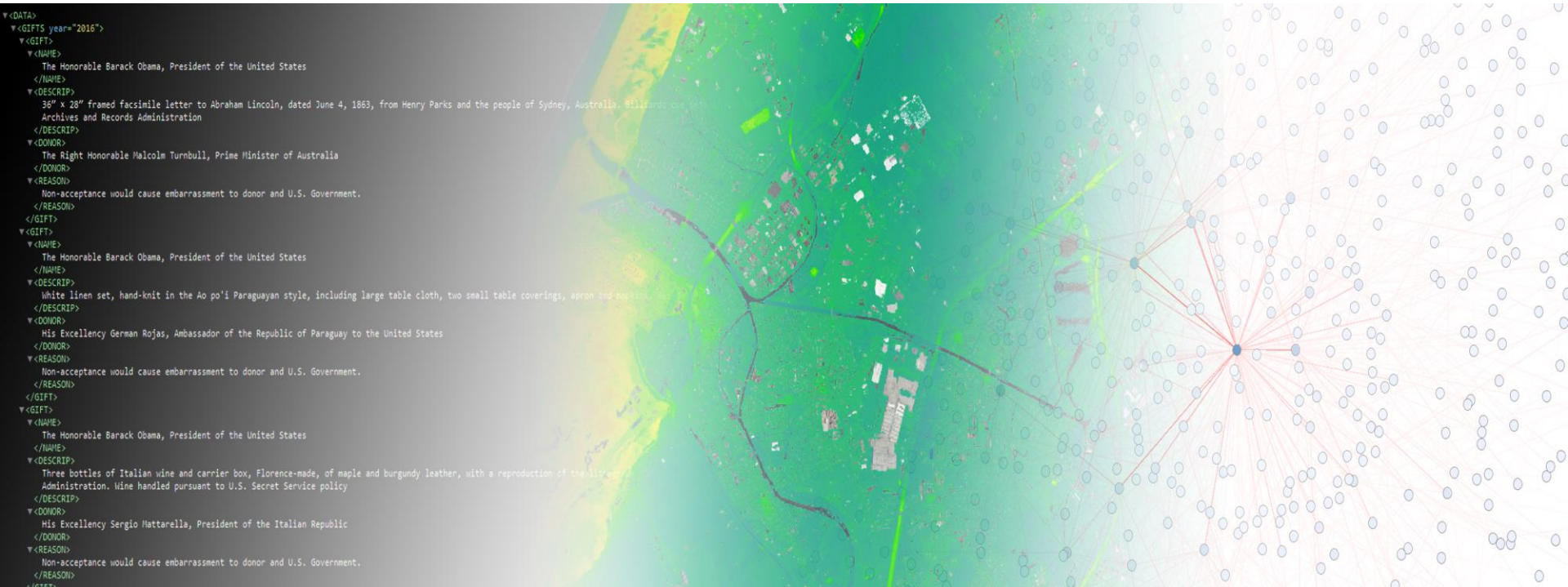
Information Visualization and the Humanities

Network Visualization

Martin Kroon

19-12-’22



Last time

- Data structure
  - The anatomy of XML
  - Relational databases
- Graph Theory
  - What is a graph?
- Graphs and Networks
  - Network is a specific type of graph: which one?

Today

- How do we visualize them?

Visualizing a network
(from last week)

- Often represented with circles and lines

- Size/colour of nodes and edges to indicate "importance"

Figure 1.4 Two graphical methods for showing the same set of relationships between entities.

But how do we get to something like this?

Today

- How do we visualize them?

- I asked you to install Gephi
  - Today we will be looking at the very basics
  - Try and do everything as I tell it!

Open Gephi now

- Ignore any pop-ups or wizards or welcome screens for now

- Download the Star Wars dataset from infovis.lucdh.nl
  - then unpack the zip somewhere!

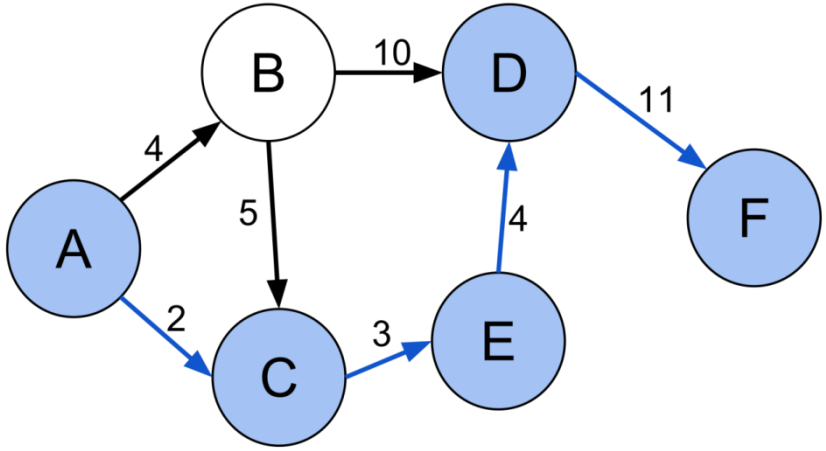
What is Gephi?

- Network analysis and visualization software
- Open-source (so, good for FAIR science)
  - Developed by non-profit consortium
- Widely used in DH, academia in general and even journalism

WARNING

does not support Ctrl-Z!
There is no undoing the done!

Graphs as csvs
(from last week)



$G = (V, E)$,

where $V = \{A, B, C, D, E, F\}$

and $E = \{(A, B, 4), (A, C, 2), (B, C, 5), (B, D, 10), (C, E, 3), (D, F, 11), (E, D, 4)\}$

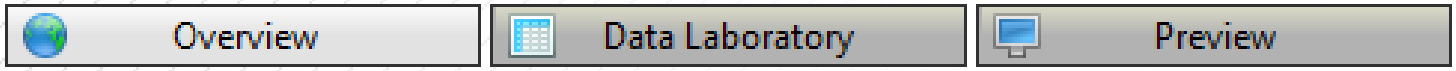
Two csvs:
- one with nodes and its attributes
- one with edges and its attributes

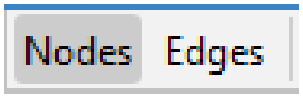
Node_label
A
B
C
D
E
F

Tail	Head	Weight
A	B	4
A	C	2
B	C	5
B	D	10
C	E	3
D	F	11
E	D	4

Loading a csv in Gephi

- First, start a new, empty Workspace/Project
  - under File

- Under Data Laboratory, or Data Table (if you have it open) 

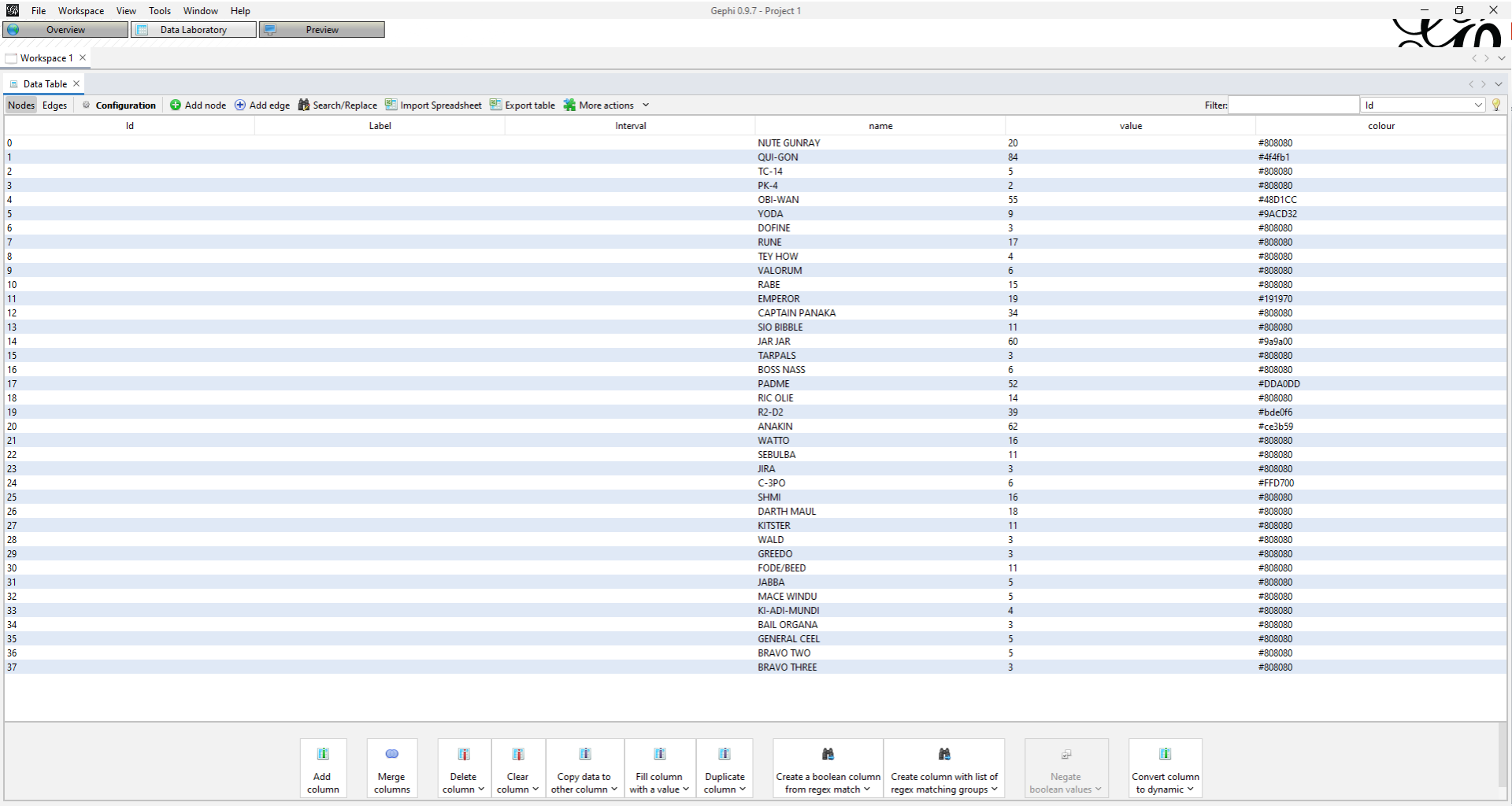
  - There are two tabs: Nodes and Edges: 

  - We need to load both

- Let’s first do Nodes

Loading Nodes

- Click Import Spreadsheet and choose your file
- A wizard pops up:
  - In the first step, you tell Gephi how to read your csv
  - In the second step, you tell Gephi which columns to import and how to interpret time (if present in the data)
  - Click Finish
- An import report pops up:
  - Define the Graph Type (when loading nodes only, this is maybe not very relevant, but to be sure, already choose the correct type)
  - Check either New workspace or Append to existing workspace
- Should look something like this:



Id	Label	Interval	name	value	colour
0			NUTE GUNRAY	20	#808080
1			QUI-GON	84	#4f4fb1
2			TC-14	5	#808080
3			PK-4	2	#808080
4			OBI-WAN	55	#48D1CC
5			YODA	9	#9ACD32
6			DOFINE	3	#808080
7			RUNE	17	#808080
8			TEY HOW	4	#808080
9			VALORUM	6	#808080
10			RABE	15	#808080
11			EMPEROR	19	#191970
12			CAPTAIN PANAKA	34	#808080
13			SIO BIBBLE	11	#808080
14			JAR JAR	60	#9a9a00
15			TARPALS	3	#808080
16			BOSS NASS	6	#808080
17			PADME	52	#DDA0DD
18			RIC OLIE	14	#808080
19			R2-D2	39	#bde0f6
20			ANAKIN	62	#ce3b59
21			WATTO	16	#808080
22			SEBULBA	11	#808080
23			JIRA	3	#808080
24			C-3PO	6	#FFD700
25			SHMI	16	#808080
26			DARTH MAUL	18	#808080
27			KITSTER	11	#808080
28			WALD	3	#808080
29			GREEDO	3	#808080
30			FODE/BEED	11	#808080
31			JABBA	5	#808080
32			MACE WINDU	5	#808080
33			KI-ADI-MUNDI	4	#808080
34			BAIL ORGANA	3	#808080
35			GENERAL CEEL	5	#808080
36			BRAVO TWO	5	#808080
37			BRAVO THREE	3	#808080

Loading Edges

- Basically the same thing as with Nodes

- BUT:
  - Choose Append to existing workspace
  - Otherwise it will load it as a new graph, and what you want is to merge both streams of information into one network: both nodes and edges

Overview

- Let’s get back to Overview
  - Appearance and Layout on the left
  - Statistics on the right
  - Graph overview in the middle

  - We’ll look at all of these things in a bit

- Here you can edit your graph/network

- But first, should look something like this:

File Workspace View Tools Window Help

Gephi 0.9.7 - Project 1

Overview | Data Laboratory | Preview

Workspace 1

Appearance

Nodes Edges

Unique Partition Ranking

#c0c0c0

Apply

Layout

---Choose a layout

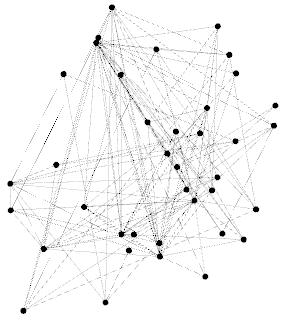
Run

<No Properties>

Presets... Reset

Graph | Data Table

Dragging (Configure)



Arial Bold, 32

Context

Nodes: 38

Edges: 135

Undirected Graph

Filters | Statistics

Settings

Network Overview

Average Degree Run

Avg. Weighted Degree Run

Network Diameter Run

Graph Density Run

HITS Run

PageRank Run

Connected Components Run

Leiden algorithm Run

Community Detection

Modularity Run

Statistical Inference Run

Node Overview

Avg. Clustering Coefficient Run

Eigenvector Centrality Run

Edge Overview

Avg. Path Length Run

Dynamic

Nodes Run

Edges Run

Degree Run

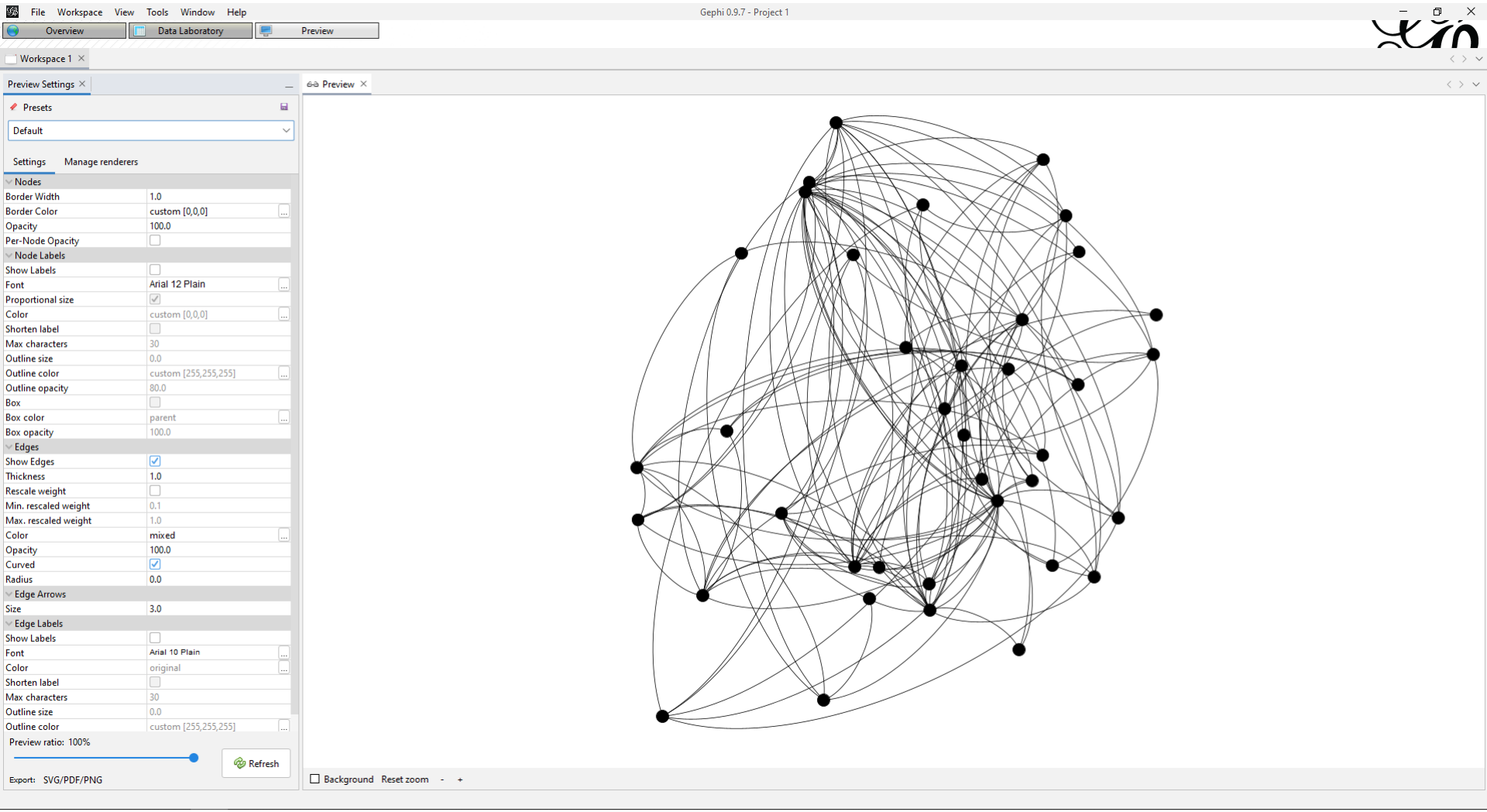
Clustering Coefficient Run

Preview

- Let’s have a look at Preview

- Here you can render your final network visualization and export it
  - Button in the bottom left: Export: SVG/PDF/PNG

- After you click Refresh, should look something like:


File Workspace View Tools Window Help
Gephi 0.9.7 - Project 1
Overview
Data Laboratory
Preview
Workspace 1
Preview Settings
Presets
Default
Settings
Manage renderers
Nodes
Border Width 1.0
Border Color custom [0,0,0]
Opacity 100.0
Per-Node Opacity
Node Labels
Show Labels
Font Arial 12 Plain
Proportional size
Color custom [0,0,0]
Shorten label
Max characters 30
Outline size 0.0
Outline color custom [255,255,255]
Outline opacity 80.0
Box
Box color parent
Box opacity 100.0
Edges
Show Edges
Thickness 1.0
Rescale weight
Min. rescaled weight 0.1
Max. rescaled weight 1.0
Color mixed
Opacity 100.0
Curved
Radius 0.0
Edge Arrows
Size 3.0
Edge Labels
Show Labels
Font Arial 10 Plain
Color original
Shorten label
Max characters 30
Outline size 0.0
Outline color custom [255,255,255]
Preview ratio: 100%
Refresh
Export: SVG/PDF/PNG
Preview
Background
Reset zoom
-
+

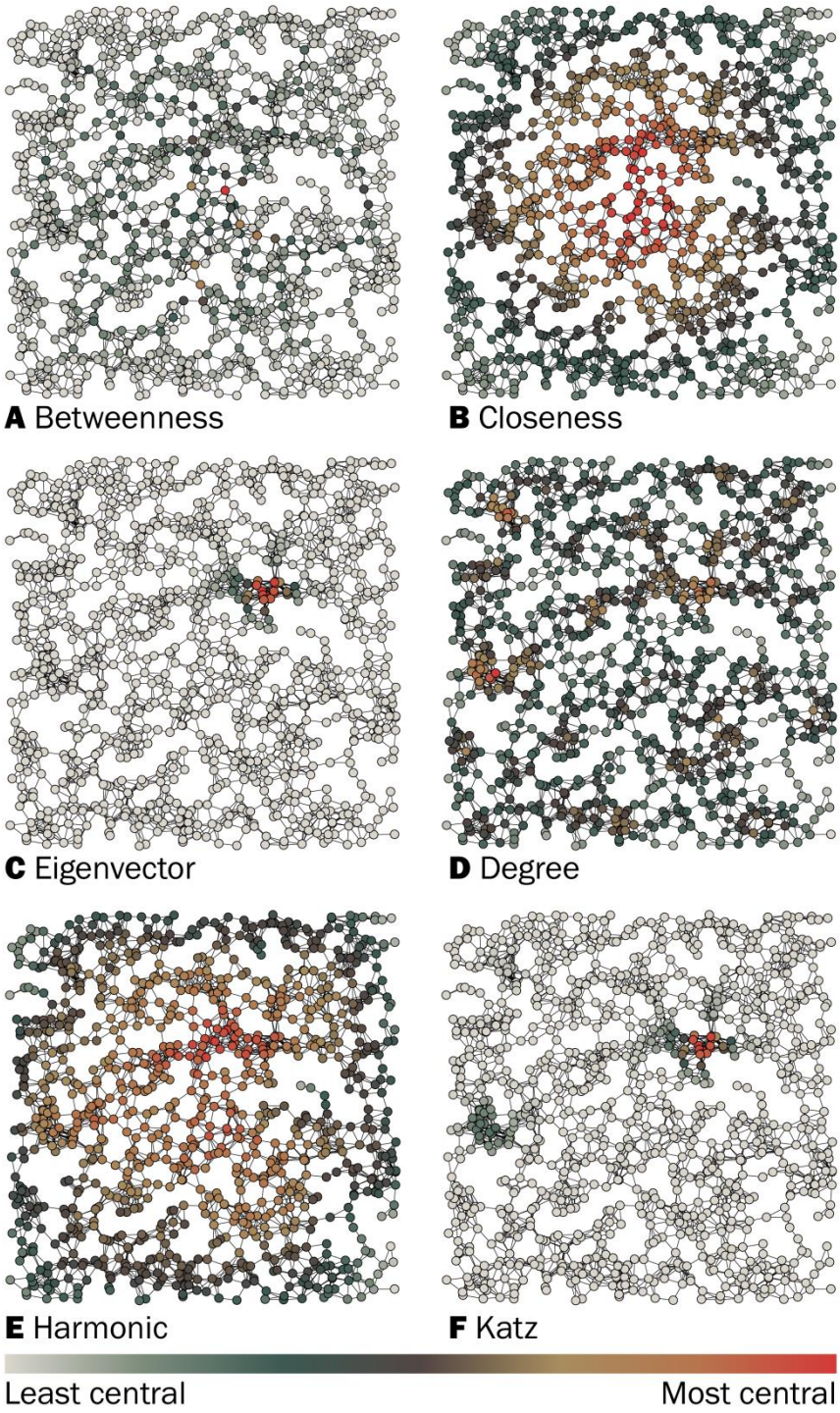
Colouring nodes

- Looks quite cool already, but let's give those nodes some colour!
- Directly left of the graph in Overview, find the Painter button in this bar:

- Go ahead and click any node once, multiple times or hold the button!
  - Change the colour above the graph

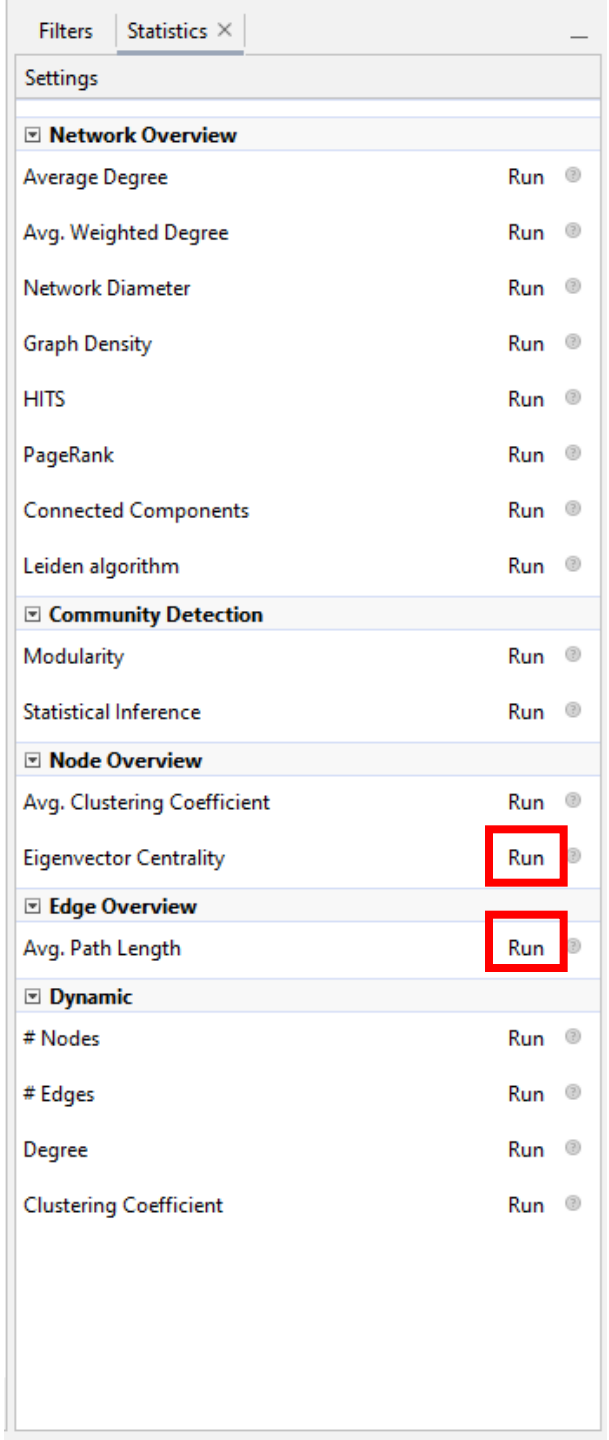
Centrality

- Last week we talked about centrality
  - The “importance” of a node within a network
  - Different ways of measuring this
    - Betweenness (based on number of shortest paths running through node)
    - Closeness (based on length of all shortest paths from node to all other nodes)
    - Eigenvector (high eigenvector score means that the node is connected to many nodes who themselves have high scores)
    - Degree (number of edges incident on node)



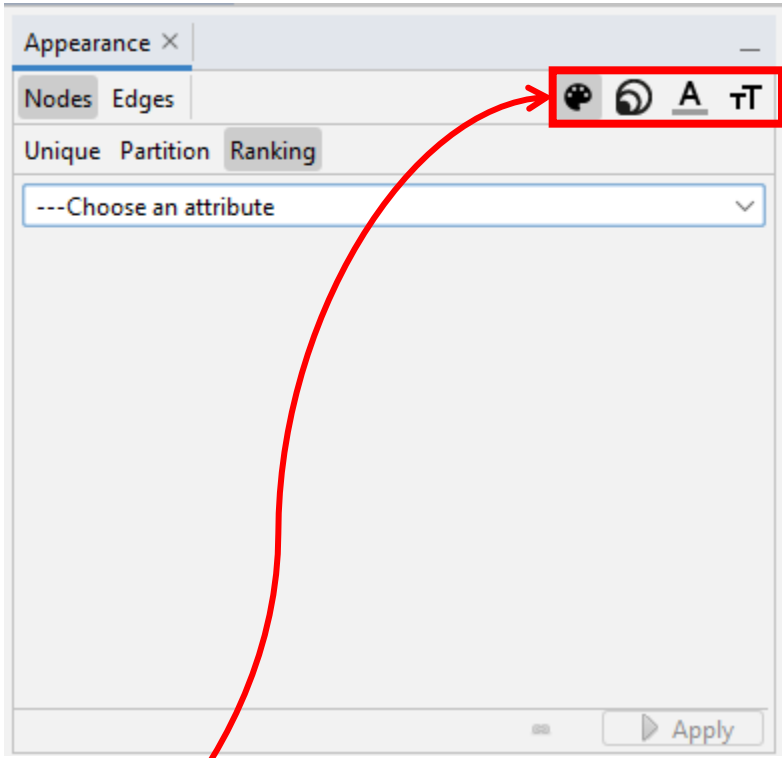
Centrality in Gephi

- Gephi can calculate centralities for all nodes:
  - Under Statistics on the right:
    - For Eigenvector Centrality, click Run
    - For Betweenness and Closeness, click Run for Avg. Path Length
    - For Degree, no need to do anything (automatically calculated)
    - You'll get pop-ups with reports.
    - Check Data Laboratory and Data Table: calculated statistics as attributes of nodes now!



Colouring nodes automatically

- On the left under Appearance
  - Nodes > Ranking
    - Choose an attribute, centralities should be under it if you ran them!
    - Choose a gradient
  - Nodes > Partition
    - Give colour based on discrete category
  - Click Apply.
- The same holds for node size, text colour and text size



Modularity

- You can automatically detect communities/clusters in a network, and assign them a discrete label
  - Often done with Modularity
- Modularity is a measure of how well a system was broken into separate components
  - Modular building, programming
  - Language modules
  - etc.
- In networks: how well was this network broken into separate components/communities/clusters/groups?

Modularity in networks

- High modularity means there are dense connections (i.e. many edges) between nodes *within* a community, but sparse connections *between* (nodes in) other communities

- Given a partition of nodes into groups, modularity is the fraction of the edges that fall within the groups, minus the expected fraction if edges were distributed at random
  - Max 1, min -1/2

Modularity in other words

- Say we divide a network's nodes into multiple groups
- How much do the edges between nodes stay *within* a group?
- If the modularity is high, they stay *within* their groups very well/much
- If the modularity is low, there are many edges *between* groups
  - And then your groups make no sense, your division is not very 'modular'

Louvain Method

- Gephi uses the Louvain method to detect communities
  - Algorithm that tries to optimize the modularity of a network
  - Finds a partition of the data such that modularity is as high as possible
    - Local optimum, so slightly different answers almost every time, but very fast and rather good approximation!

- For the assignment, do some research at home into the workings of the algorithm, and explain in your own words to me how it works. It is important to know how algorithms work in order to better understand and interpret them.

- Blondel, Vincent D; Guillaume, Jean-Loup; Lambiotte, Renaud; Lefebvre, Etienne (9 October 2008). "Fast unfolding of communities in large networks". *Journal of Statistical Mechanics: Theory and Experiment*. 2008 (10): P10008. arXiv:0803.0476. Bibcode:2008JSMTE..10..008B. doi:10.1088/1742-5468/2008/10/P10008. S2CID 334423.

Modularity in Gephi

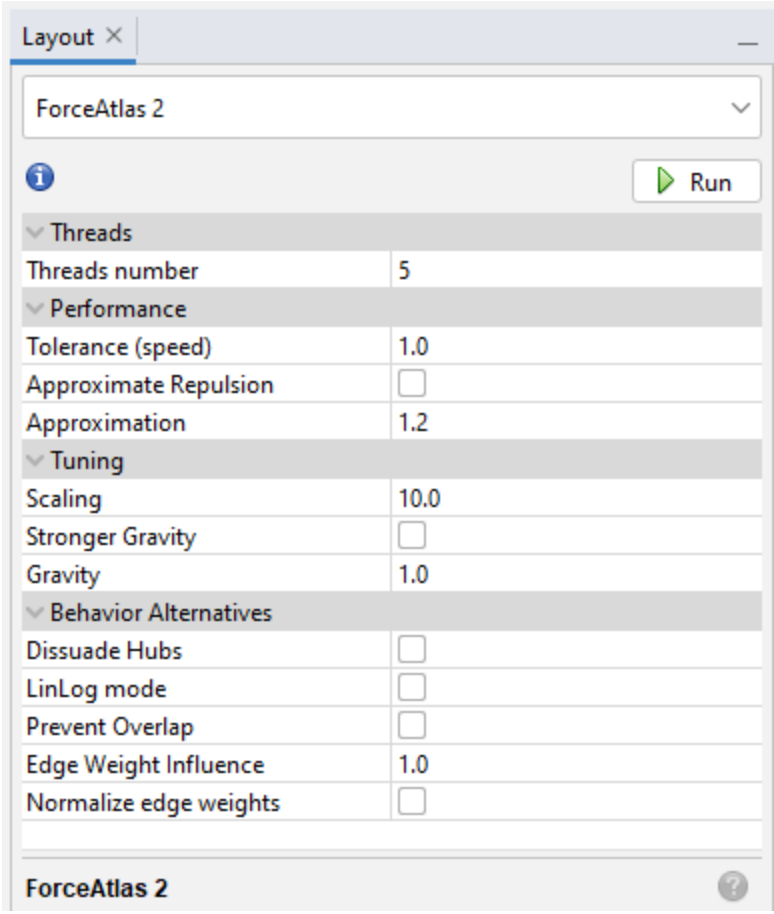
- Under Statistics, click Run for Modularity
  - Randomize improves results
  - Use weights: whether to use edge weight in the calculation. Higher weights are stronger connected edges
  - Resolution to tweak number of communities
- Result: report with number of communities found, and number of nodes per community
- Appearance > Nodes > Partition > Modularity Class

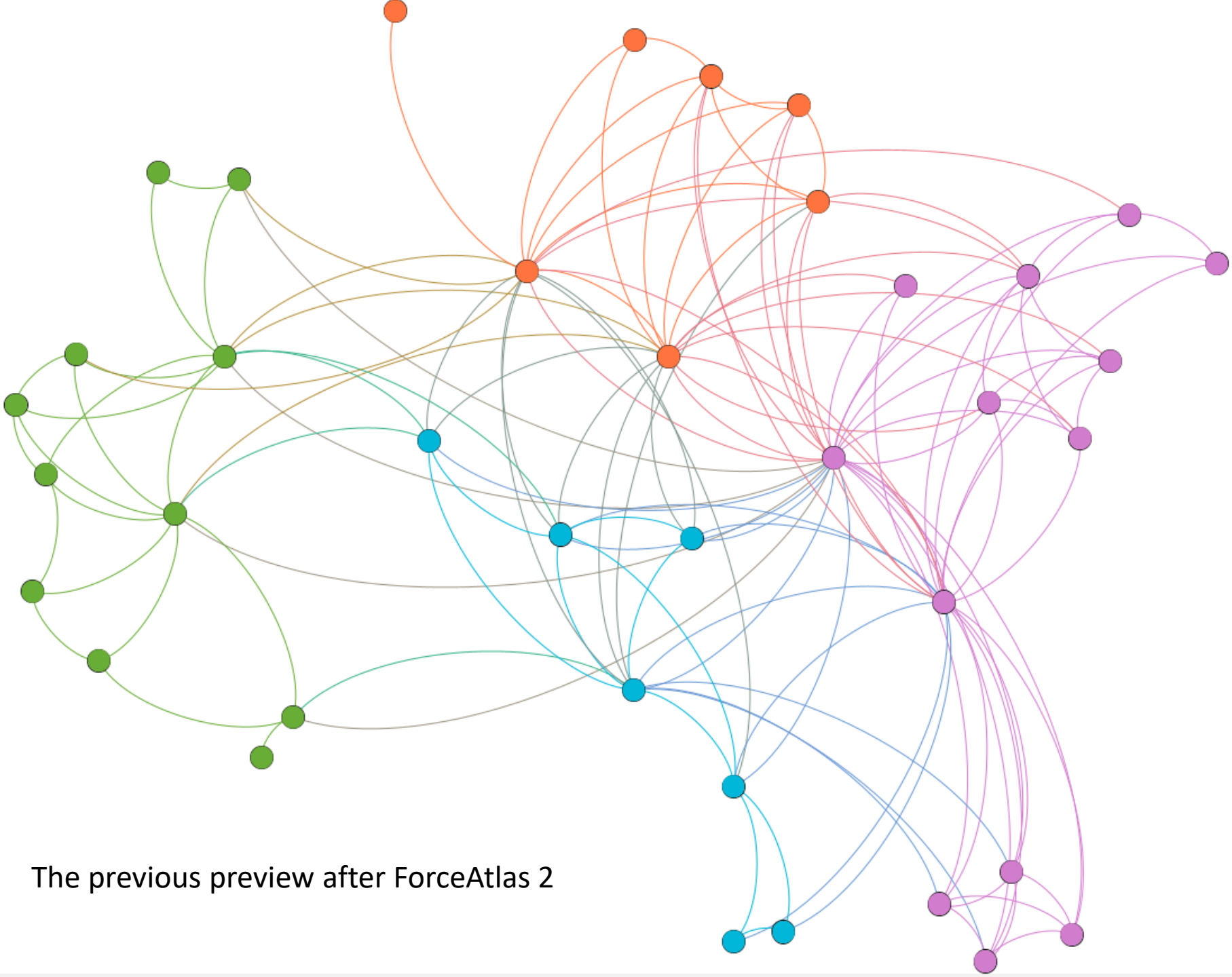
Preview

- Let’s have a look at the preview now
- Terribly cluttered!

Layout

- Under Layout, on the left under Overview, you can choose a layout algorithm
- Different ones. Play around!
  - Also with the parameters, to see what happens
- ForceAtlas 2 is a simple force-directed layout





The previous preview after ForceAtlas 2

Connected components

- Some Layout algorithms push nodes out of the screen
  - Possibly due to multiple connected components, that are pushed away from each other, because there is no edge to ‘pull them back in’
  - A connected component in a (undirected!) graph is a set of nodes such that all nodes are reachable from every other node

  - Statistics > Connected Components
    - Gives number of connected components in network
    - Can be used to colour your nodes, too.

Moving nodes

- You can of course also manually move nodes
- This can be done with this button

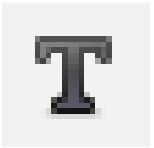
- If a node (or group) is pushed away, you might want to drag them closer to the rest
- Tip: while ForceAtlas 2 is running, you can move nodes around to try and force another layout!

Deleting nodes and edges

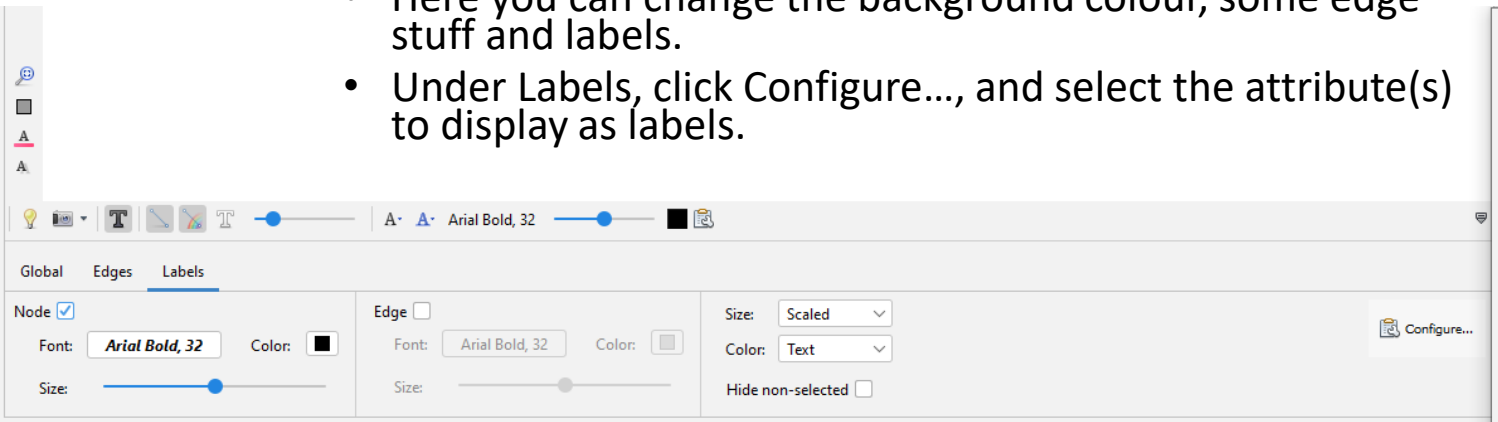
- If one node (or edge) is an outlier or wrong, or just not wanted, you can of course delete them:
  - Right-click on the node in the graph Overview
  - Data Laboratory/Data Table > Click a node > Press the delete key
  - Outliers: for example nodes that are not connected to anything (or perhaps even those that are just VERY far away). You can use Modularity Class or Component ID for this to identify one-node classes and delete them.

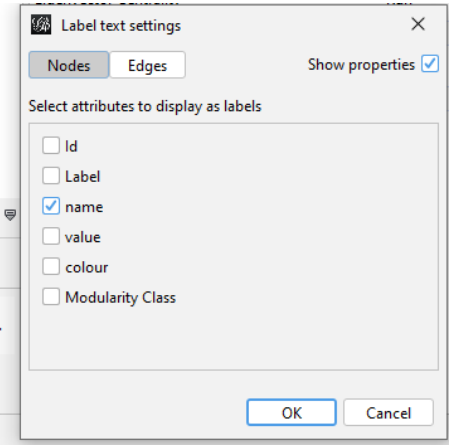
Labelling nodes and edges

- Now, of course our nodes and edges have no labels
  - Below the graph in Overview, click Show Node Labels:



- Still nothing happens?
  - This is because Gephi automatically takes the value of the Label fields in the data as input for labels. Which is empty.
  - We need to tell it to take other data as label input.
  - Below, click More Settings…
    - Here you can change the background colour, some edge stuff and labels.
    - Under Labels, click Configure…, and select the attribute(s) to display as labels.





Edge thickness

- Same problem holds for edge thickness
  - Edge thickness uses the attribute Weight, which is empty.
  - For now, the fact that Gephi uses Weight cannot be changed.
  - But we can copy the contents of our value attribute into the Weights column:
    - Data Laboratory > Edges > Copy data to other column > value > Weight
    - Now, the edge thickness is scaled by its weight.

But wait! Weight!

- Another result actually is that the statistics that we calculated before, did not take into account the edge weights, nor did the Layout algorithm.
  - Having the right values in the column called Weight is important for Modularity and Layout!
  - However, Weight is currently not supported for many other statistics, such as centrality and path lenghts…
    - Tip: you may need to change the Scaling factor in your Layout algorithm when using Weight.

Other tools

- Do check out the following tools, too!
  - Manual node sizing
  - Pencils
  - Shortest path
  - Heat Map

All kinds of other functionalities

- I’m sure there are many functionalities in Gephi that I do not yet know!
- I highly recommend checking out some of their own tutorials: https://gephi.org/users/
- In fact, Gephi supports plugins to deal with functions currently not in Gephi-proper.
  - I also recommend checking those out if there is something you want to do, but Gephi won’t let you.
  - https://gephi.org/plugins/#/

Final slide

- Good luck with the final assignment
  - When I say to give information “in a way that is detailed in such a way that I would in theory be able to roughly replicate the visualization”, this does not mean “write a tutorial”. Write academically, give me the information I need to know in order to reproduce: any parameters or settings you used (what software?), anything that isn’t “default”.
    - E.g. I can find the right button if needed.
  - You are allowed to use different software than Gephi!

- Thank you for this semester, I enjoyed it a lot!

Happy holidays! 

Final assignment

- I. Network (50%)
  - Visualize a network of Star Wars characters
  - Write a short report in which you do some network analysis
- II. Free choice (50%)
  - Self-contained visualization of your choice, any story, any data, any type of visualization
- Deadline: before 16th January 2023 9.00 AM

Final assignment

- I. Network (50%)
  - Visualize a network of Star Wars characters, from any episode or multiple
  - Data is available on the website infovis.lucdh.nl (Gabasova 2016)
    - Interactions and mentions of characters within scenes
  - Report: 500-1500 word report
    - A clear reference to the network data used
    - a discussion of all your design decisions
    - the following measures of your network:
      - a table with the ten highest ranking nodes as measured by Betweenness and Degree; and an interpretation of said measures with respect to the data
      - the number of nodes per Louvain community; and, if possible, an interpretation of the communities
    - An explanation in your own words and references to outside sources where applicable of:
      - What is the difference between Betweenness and Closeness?
      - How does Louvain Modularity work?
    - a discussion of challenges and opportunities (if any).

Final assignment

- II. Free choice (50%)
  - Self-contained visualization of your choice, any story, any data, any type of visualization
    - Can be viewed independently
    - Can be understood independently
  - Find a story that you would want to visualize
    - Suggestions of datasets on the website
  - Not necessary to create a new dataset
  - Report 1500 words max.
    - An explanation of the story you are visualizing
    - The data and design tools you used and choices you made in creating this visualization in a way that is detailed in such a way that it would in theory be able to roughly replicate the visualization (don't forget to reference where appropriate!).
    - Challenges encountered during the project and opportunities you see for this or similar information visualizations
    - A project timetable, including an overview of time spent on the project and optional self-learning

- Deadline: before **16th January 2023 9.00 AM**