Gephi: Making your relational data pretty

Jesse Fagan
LINKS Center or Social Network Analysis
aftersox@gmail.com

Data and other links
Attribute data: https://goo.gl/vFVzg5
Edgetlist data: https://goo.gl/zGvFl6

Lots of network data!: https://icon.colorado.edu/#!/
My network visualization slides: https://goo.gl/0biHfK
Gephi Book: https://goo.gl/XbgvKo

Figure 2: Ingredient complement network. Two ingredients share an edge if they occur together more than would be expected by chance and if their pointwise mutual information exceeds a threshold.


Gephi is open source software. You can find the source here: https://github.com/gephi/gephi

Gephi is designed for visualizing very large, complex networks. It’s best if used on a computer with a good graphics card. The network rendering uses OpenGL and the GPU. The layout algorithms also take advantage of multi-core CPUs. This is an introduction to Gephi, we won’t be covering some advanced features (such as directly connecting to databases, longitudinal
Navigating Gephi

There are three sections to Gephi. The Overview, Data Laboratory, and Preview. There are also lots of little windows and tabs and it can all be moved, minimized, and closed. This can result in some situations in which you mess everything up and can’t find the windows you need any longer. If that happens under the Window menu there is an option for Reset Windows which will set all the tabs and such back to the defaults.

**Overview** This is where you will be spending most of your time. This is where all the work is done translating the data into a visualization.

The Appearance window controls the color, size of the nodes, labels, and edges. The Edit tab will let you see the attributes of the nodes, and change them. The Layout controls the position of the nodes. The graph is a visualization / exploration of the network. Context gives you a big current overview of what you’re working on. The Statistics tab runs some basic network analyses. The Filters tab lets you filter out nodes and edges based on a range of different criteria.
The **Data Laboratory** is where you can see a spreadsheet of all the data in your project. It is clumsy and limited. There is a limit on the number of columns it will display. Click the **light bulb** icon on the top right part of the window to choose which columns you want to see.

This tab can be helpful if you notice a potential error in your visualization. For example, you may notice that one edge in the visualization is enormous because you coded 999 as a missing weight while everything else is 1-5. You can make edits in this tab.

You can also select a node and have it highlighted in the Overview tab. And vice versa, you can select a node in the Overview and have it highlighted in the Data Laboratory.
The **Preview** tab is where you prepare your final image for export. It takes information from the **Overview** tab then produces a high quality render of the results.

Gephi also has Workspaces. If you are working on a project with multiple relationships, say a co-citation network and a co-authorship network, you can keep them in the same project but separate them to different workspaces.

**Exploring a Network**

- Zoom with mouse wheel, right click to move around
- Selecting nodes
  - Using ctrl + mouse wheel
  - Rectangle selection
  - Moving nodes around
- Center on the graph
- Gephi screenshots
- Edit / view node attributes using the ? tool
- Manually sizing and coloring nodes
- Manually adding new nodes and ties

**Filtering the network**
- Filtering by attribute (subnetworks)
- Filtering edges (edge weight)
- Sub-filters (removing isolates)
- By topology (degree range, k-core, ego-networks)
- Sending the filtered network to a new workspace

**Sizes and Colors**

- **Nodes**
  - Color (overall, by attribute, nominal data, interval data)
    - Different palettes, editing colors and gradients
    - Splines
  - Size
    - Minimum and maximum sizes
    - Using splines
- **Edges**
  - Size controlled at the bottom.
  - Color by attribute
  - Color by the node colors

**Statistics and Analyses**

Gephi is made for visualization, not for analyses. So things are bit sparse here. There are some plugins that add some analyses too.

- All forms of degree
- Centrality: betweenness, eigenvector, PageRank
- Clustering coefficient
- Components and modularity

**Layouts**

There are a wide variety of layouts available for network visualizations. My favorite is Force Atlas 2, which is only available in Gephi (with a rudimentary version in R). Some of these layout options are helper layouts, not made to be used on their own.

- **Using Force Atlas 2**
  - Most of my tweaking is based on scaling and gravity
  - Prevent overlap if nodes are too crowded
  - LinLog mode for very dense, large networks, then re-tune the scaling and gravity
• Once you get a setting you like, save it as a preset
• Tweaking with expansion, contraction, rotation, and label adjust

Preview and Export
In this window you create a high quality render of the network and export it for sharing. After every change, remember to hit Refresh.

• Zoom, background, preview ratio
• Presets
  ○ Using a default setting
  ○ Saving your own settings
• Nodes
  ○ Border width, color, opacity
• Node labels
  ○ Size, font
  ○ Proportional size
  ○ Outlines, boxes
• Edges
  ○ Thickness, rescale weight
  ○ Opacity
  ○ Curved
  ○ Arrows and labels

Investigating and Changing Data

• Highlight in the Overview / Data Laboratory
• Copying, filling, merging columns
• Showing and hiding columns
• Filtering tables
• Export data

Reading in Data
Data is read using the Data Laboratory tab. Click “Import Spreadsheet”. Gephi only support CSV file format.

The network data must be structured as an edge list. The titles of the first two columns need to be “source” and “target”. These are special column names that Gephi looks for and they will be used to construct the network ties. The word “weight” is also a keyword that Gephi looks for. You can have all sorts of other edge attributes, but make sure you assign them the correct data type.

For nominal data use String, and for interval data use Double. For the most part, you can
entirely ignore the long list of other data types.

If you have node attribute data you should import the edgelist first. The node attribute data needs to have a column called “Id”. The values of the ID’s need to match the names of the nodes in the edgelist. Make sure you have the box “Force nodes to be created as new ones” unchecked, otherwise it will import all your node data as brand new nodes.

- Force nodes to be created as new ones