Gephi for Analysis
Part II: Advanced Tools

NEH Digital Culture 2020 Workshop
Advanced Gephi use

Now that you have the basics, we can try some of the other tools Gephi offers.

In general, Gephi has many, many features, so we can’t cover all of them, and not all of them are as useful for humanities research anyway, but here are the next things to try once you feel comfortable.
For this set of exercises, we need some plugins.

Go to Tools in the top menu, then Plugins, then choose the Available Plugins tab. Find the “Circular Layout” option and select “Install.”

The plugins are installed and you are asked to reboot Gephi. Click “OK.”
Metrics

● A more complex way to see important participants in the network is Betweenness Centrality.

● First, calculate metrics using the Statistics pane at the right of the interface
  ○ First, “Run” the “Avg. Path Length” parameter
Node size

- Mathematically, betweenness centrality is about whether a node is on the shortest path between two other nodes.
- Pragmatically, it’s which nodes connect more or fewer other nodes.
- Conceptually, for a question about an information ecosystem, the higher the Betweenness Centrality, the more likely that video is to be shown to people watching the other videos.
  - We could then say that it is likely to have a greater influence.
Clustering

The next thing we can do with Gephi is clustering.

This tells you how much more densely connected the nodes within a community are, compared to how connected they would be in a random network, and lets you find groupings in the data.
Clustering

In the Statistics panel on the right, run Modularity. The default settings are fine.

Then under Appearance on the left, choose Nodes, then Partition, then Modularity Class from the dropdown. It will default to a set of colors, which can be changed with Palette in the lower right corner of the panel.
Clustering

That will give you something like this, but they’re all mixed together.

In Layout, run Force Atlas to put the related ones together. It will look like this. Don’t panic!

Let it run for a bit until you see some of the colors peeling off, and then hit stop. It won’t stop on its own.
Clustering

From your Force Atlas cluster, use the Label Adjust layout to get it moved around to see what’s in it.

You can see that it’s largely sorted by color, though there are some nodes that are probably borderline between two clusters.
Clustering

So what does it mean? Let’s go back to what clustering does: tells you how much more densely connected the nodes within a community are, compared to how connected they would be in a random network.

These color groups are more connected to each other than they would be by chance. Your next step might be to look at the titles in the cluster, or watch the videos, to see what they’re about.
Radial layout

But we can make the groups even easier to see with a radial layout.

Starting from data that has already been partitioned by Modularity, under Layout, choose Radial Axis Layout.

Then we need to do a few steps. First, group nodes by Degree. This will put the ones with the most connections together.
Radial layout

But we can’t see much yet. The next step is to group the nodes again by Modularity Class and Order them by Degree.
Radial layout

Then we have some choices. If we check “Draw Spar/Axis as Spiral,” we can see links inside communities.
Radial layout

If we uncheck “Draw Spar/Axis as Spiral,” it’s easier to see links between communities.
What other features matter?

We’ve been looking at connections so far, because that’s what Gephi is designed to do and best at.

But the YouTube data output also includes information for each video about its comment count, likes/dislikes/ratio/favorites, published at (date), and view count.

You can rank by these features, partition by them, or use them as the label for the node.
What other features matter?

Here’s the network emphasizing viewcount: the “There Are Only 2 Genders” video is notable when ranked by viewcount both in color and node size (darker green and bigger).
What other features matter?

The network ranking by dislike ratio both in color and node size (darker green and bigger) looks different. The Inside Hillary’s Oval Office video is (proportionately) most disliked.
What other features matter?

You can also change the label on the data to one of the other attributes using this setting.
What other features matter?

So if I change the label to Channel Title and then Partition by Channel Title, I can see that just one channel is providing the vast majority of videos in this network.
What does it mean?

Much like in Gephi I, these more advanced techniques don’t tell you something all by themselves. Knowing what the cultural object, moment, or context was for the data is key for these data to have meaning.

Like before, this is a step along the way of analysis, and next steps might be going and watching some of those videos to do textual analysis, looking at the comments on the seed video with textual analysis and/or word frequency, or looking at other sources of information about this incident, like tweets or news articles.