How to build semantic networks

This "How to" will guide you through the steps to make a network that can be visualized using Gephi and analyzed using MATLAB. If you haven't already, please download and install Gephi.

Constructing a network matrix from Plato's *Meno*

1. **Download word embedding model from the link below (file size: 809 MB)**

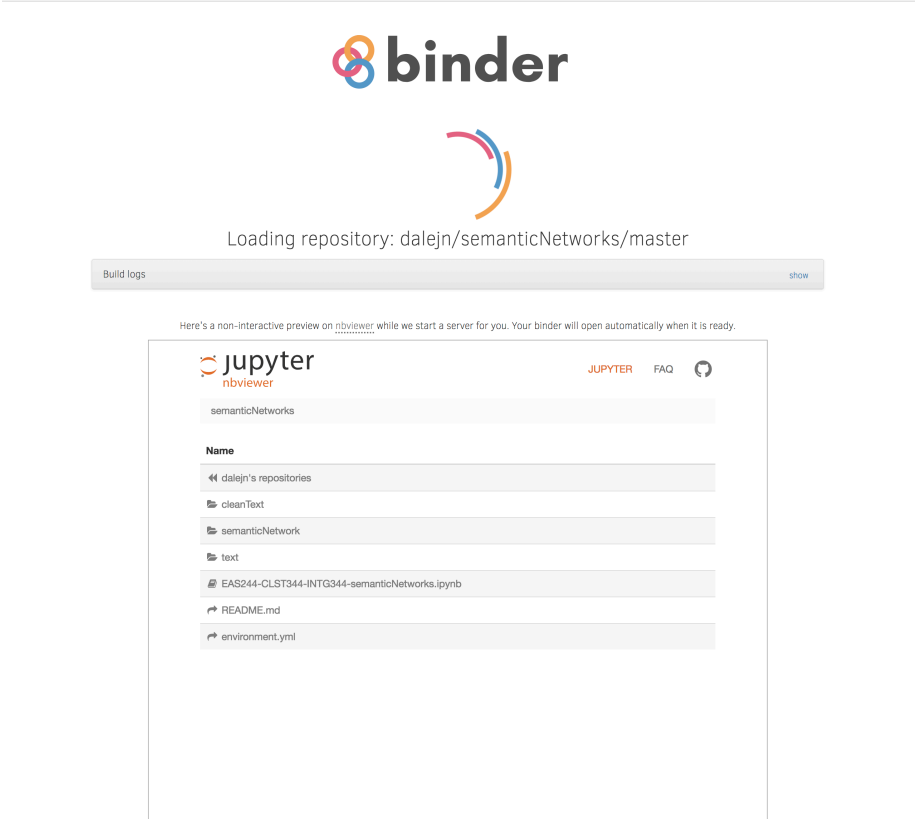
   https://www.dropbox.com/s/je4slqywoh8wnr1/enwiki_5_ner.txt?dl=1

   This is a neural network model of word vector embeddings with 296,630 word vectors representing words' semantic meaning trained on 2,252,637,050 words from Wikipedia. If interested in more details, see last section.

2. **Start programming environment**

   https://mybinder.org/v2/gh/dalejn/semanticNetworks/master

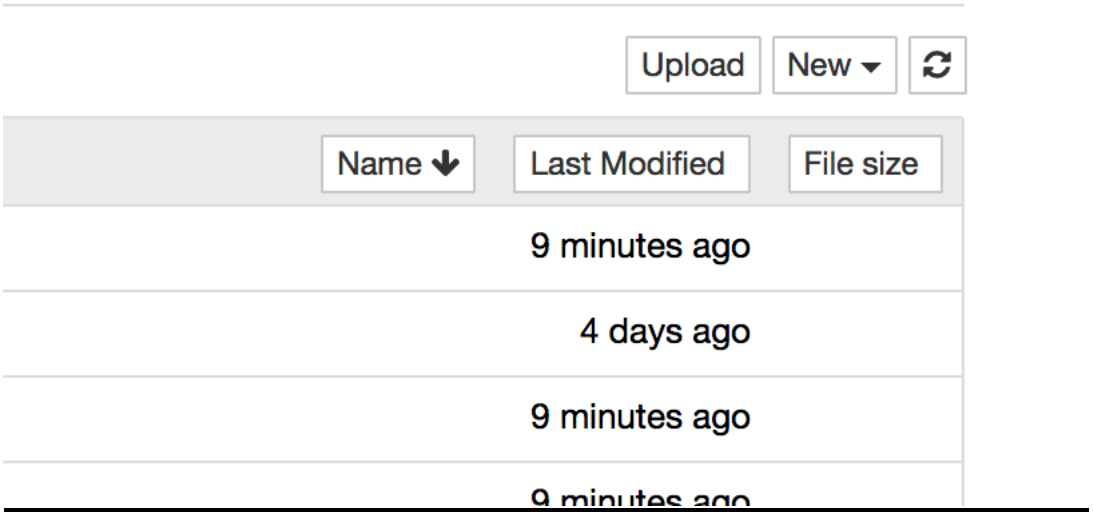
   A. You should see the loading page below. Depending on how many students are trying to access the server, it may take a few minutes or longer to start the programming environment. If it is still loading after 10 minutes, please refresh the page.



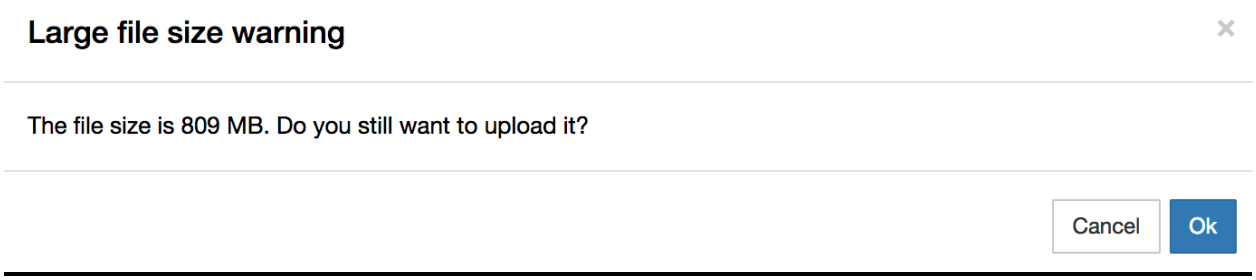
B. When the page is loaded, you will see a screen like the one below. Note, that the environment will automatically shut down after 10 minutes of inactivity (if you leave your window open, this will be counted as "activity").

3. <u>Upload the model downloaded from step 1 to the environment</u>

A. Click on "Upload," go to where you downloaded the file from Step 1 and select it.

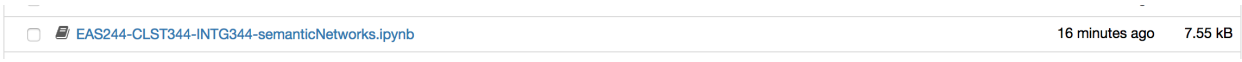


B. Press Ok after selecting the model to upload.

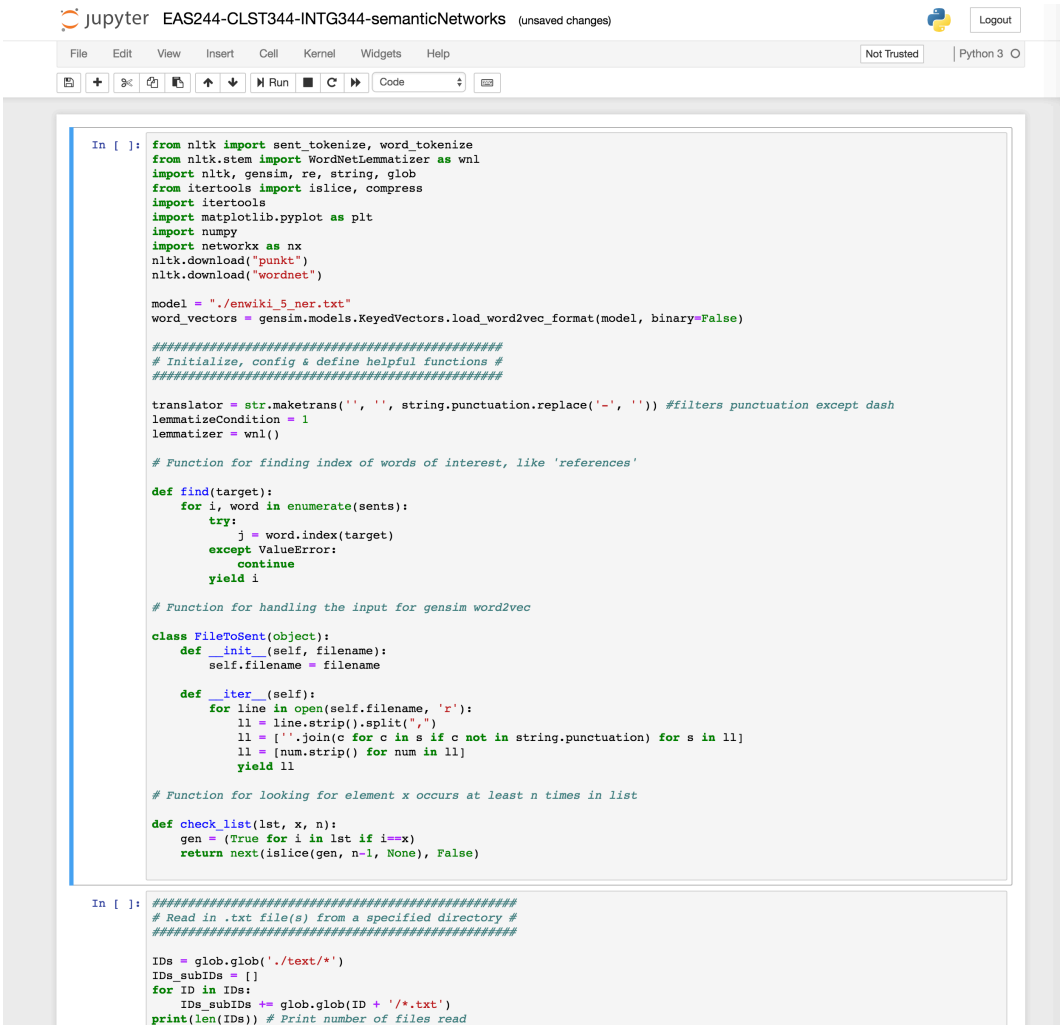


C. Click on the blue Upload button and you should see the percent progress of the upload. Depending on how many students are accessing this environment, this upload can take a few minutes or longer.

4. **<u>Once the upload has completed, click on the text link "EAS244-CLST344-INTG344-semanticNetworks.ipynb" as seen below to open a programming notebook.</u>**

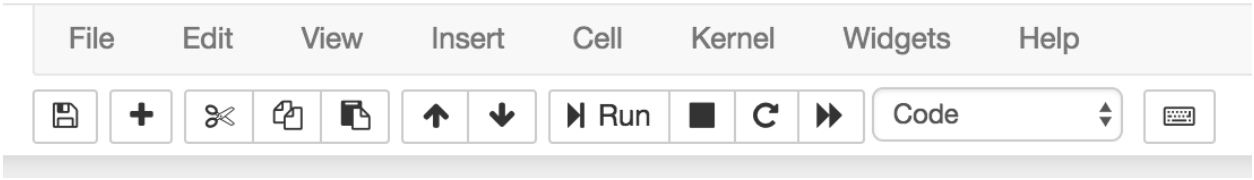


This will open a new tab/page where you should see the below:

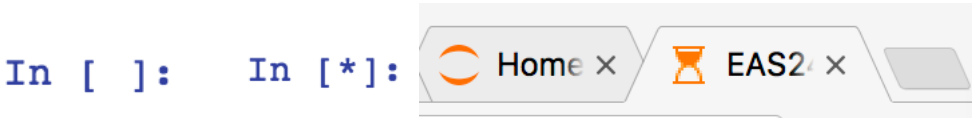


5. **Run the code**

   A. 1st box: loads the relevant packages, modules, and functions. Click on the first box, which should then become highlighted along the borders. At the top of the screen in the toolbar, click the Run button.



   You should see the left-most text change to the right. Your tab in the web browser will also show an hourglass. This means the code in this box is running. This can take a few minutes or more to finish.



   B. 2nd box: reads in your text and cleans it. When the code is done running, the asterisk above will be replaced by a number and the hourglass will be replaced by a notebook. Now, click on the second box of code, which should become highlighted along the borders. Run this code by clicking the Run button. This will take a few seconds at most.

   C. 3rd box: constructs the semantic network matrix from the shortest N cosine-distances (the strongest N associations). Select the third box. Depending on the number of nodes in the graph (i.e. number of unique words in the text), you may want to change the total number of edges (i.e. connections between the words) in the constructed semantic network. I've set this number to be 19 times the number of nodes, but this can be changed depending on how dense you want the resulting semantic network to be.

```
# The number of connections we want: either as a factor of the number of words or a set number
num_top_conns = len(my_words) * 19
```

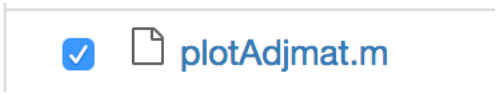
   You will see a list of words outputted.

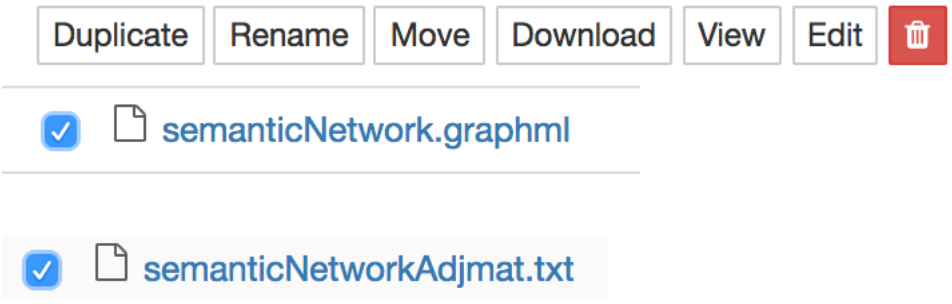
   D. 4th box: constructs a co-occurrence network based on 5-gram sliding window (represents a count of words pairs that co-occur with each other in 5-word chunks across the entire text)

6. **Download the relevant outputted files**

   A. You can close the notebook now. Return to the directory screen and click the "semanticNetwork" link.

B. Check a box (<u>one at a time</u>) to download all the files in this folder. You must select the files one at a time, or else the Download button doesn't appear.



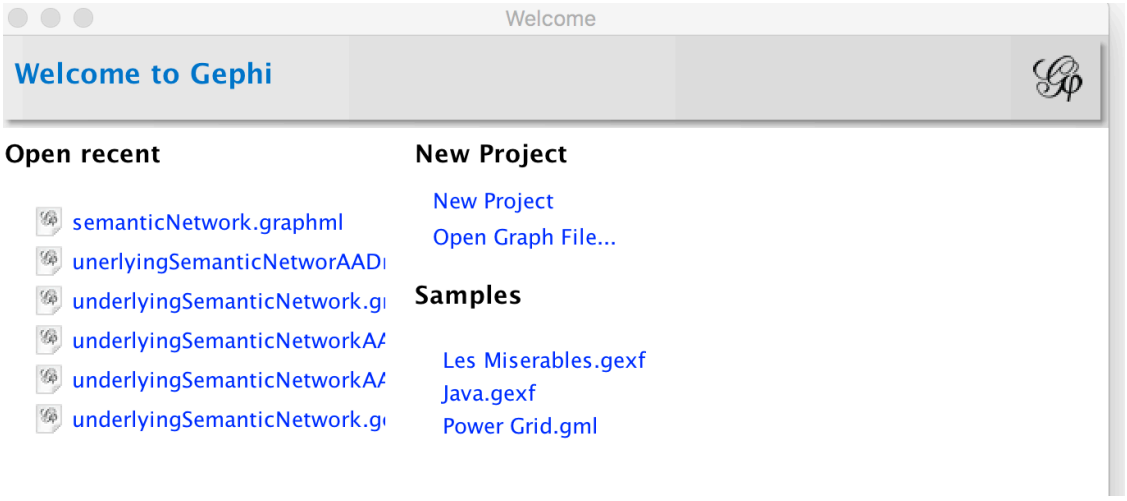


C. Press the Quit at the top right.

7. <u>Visualize the network in Gephi</u>

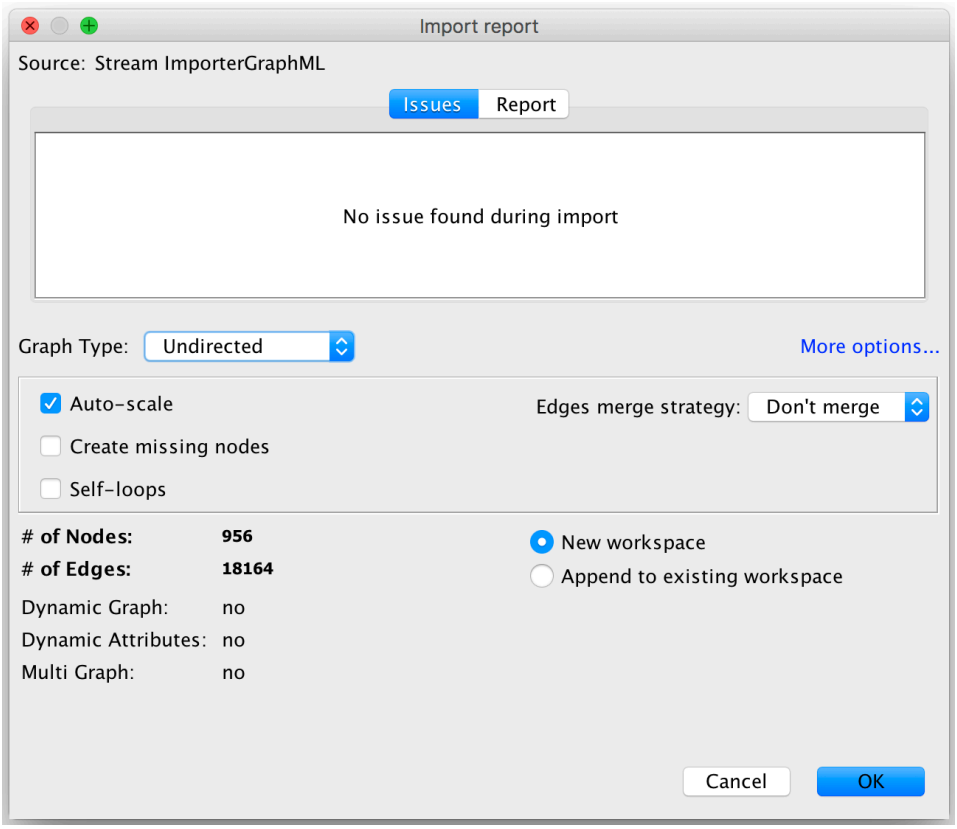
A. Open Gephi. If you haven't already downloaded and installed it, [please do so here](here).

You will see a Welcome screen like below. Click on "Open Graph File…" If you don't see the welcome screen, go to File and Open on your computer's toolbar.

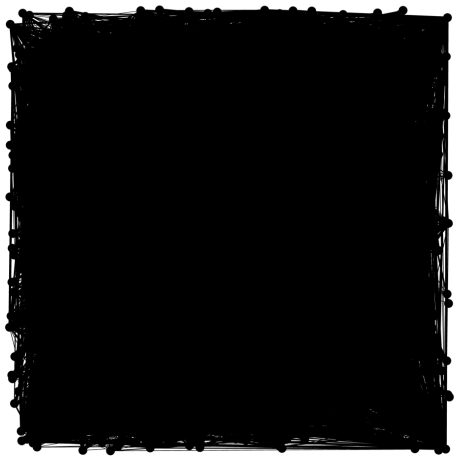


B. Navigate to where you downloaded semanticNetwork.graphml and open this file.

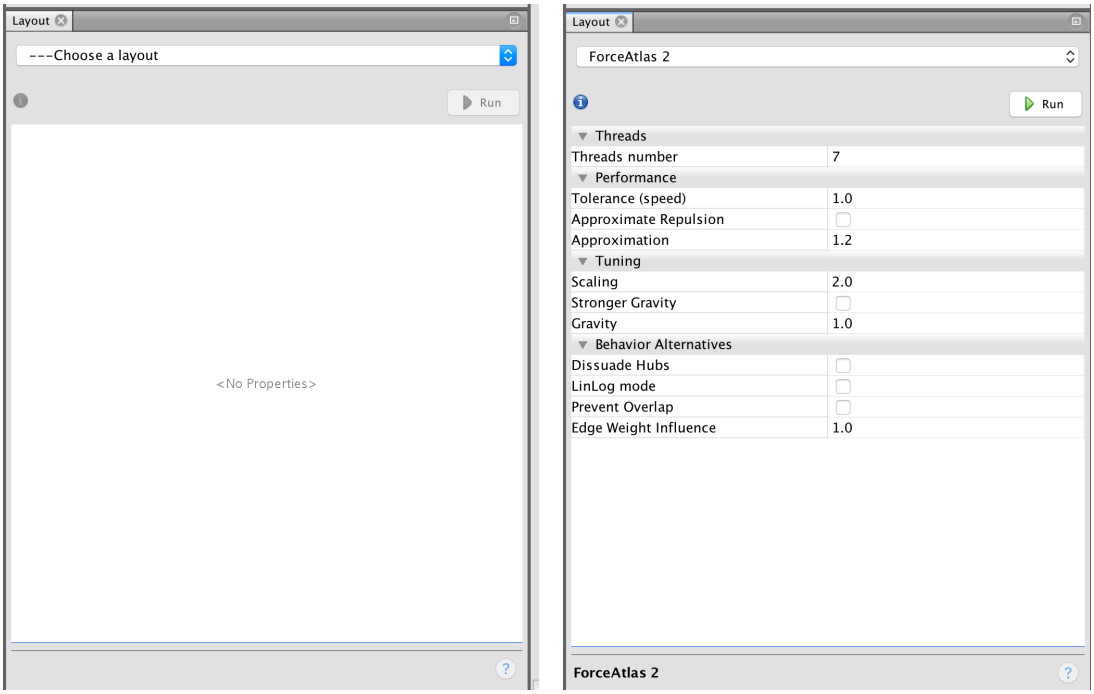
C. A new screen should pop up. Change the settings so they match with the ones below and then press OK.



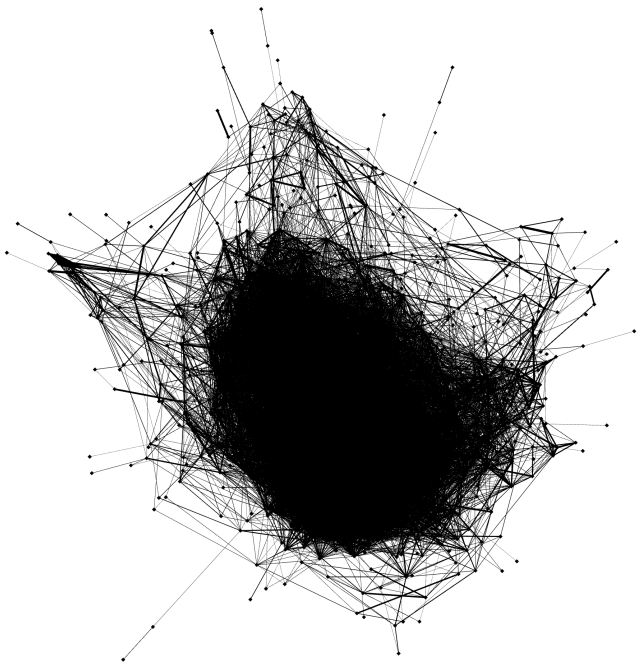
D. The network should be visible. If not visible, you may need to go to Gephi's menu at the top and select Window > Graph. You can scroll your cursor over nodes to see connections. We'll now clean this up to make it more interpretable.



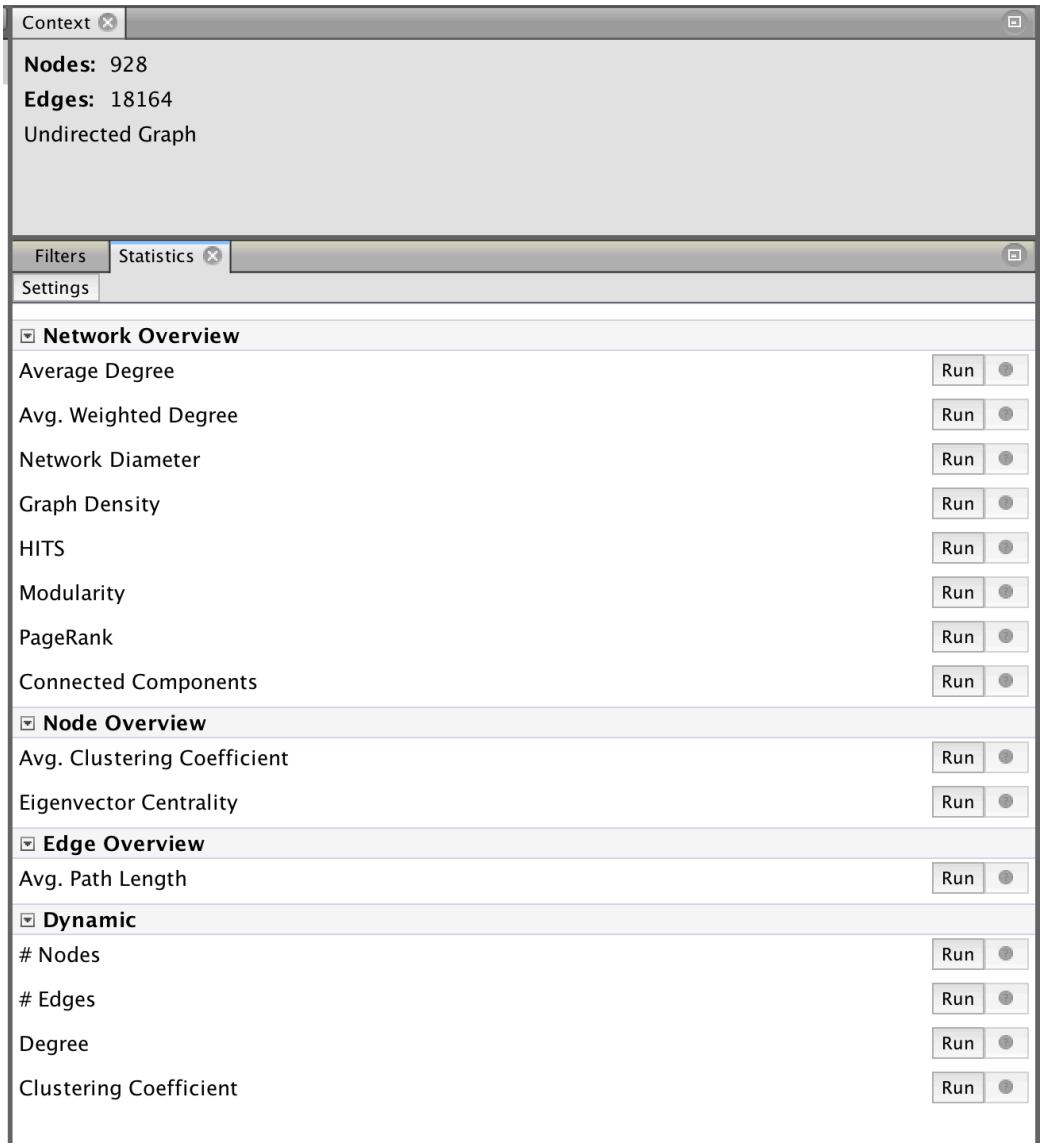
E. On the bottom left, click on "---Choose a layout" and select "ForceAtlas 2." Then, press the Run button, which will then be replaced by a Stop button.

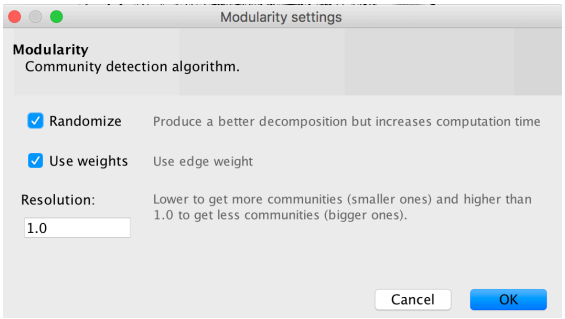


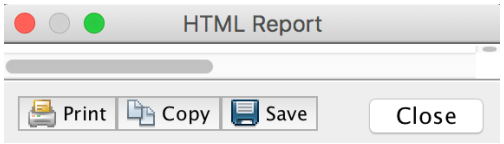
F. Use your mouse wheel to zoom in and out. Hold right click and drag to move the network around. When the network has stabilized, press the Stop button.



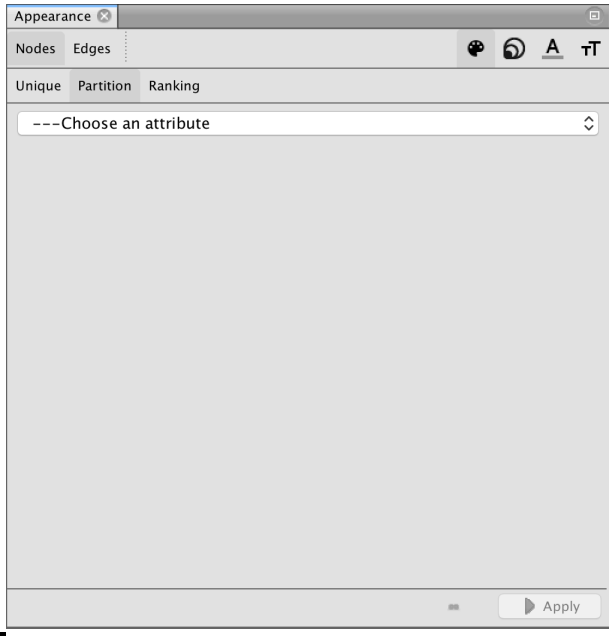
G. On the right-side toolboxes, click on the "Statistics" tab and then click "Run" for the Modularity measure. A box will pop up; click OK. Another pop-up labeled HTML Report will appear; click Close.



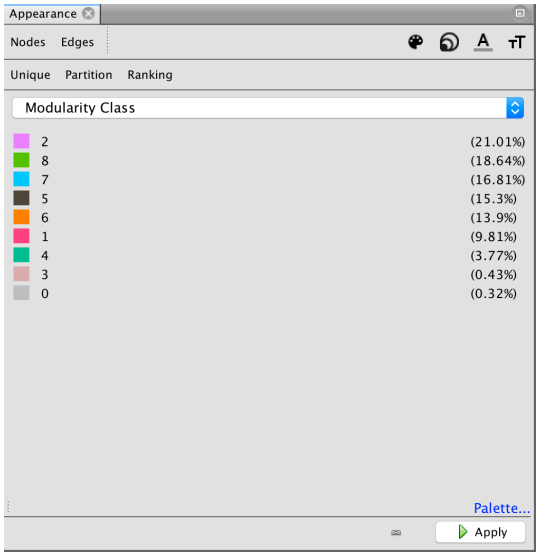




H. At the top-left toolbox labeled Appearance, click on the “Partition” tab.

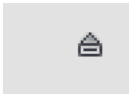


I. Click on “—Chose an attribute” and select “Modularity Class.” Click “Apply.”



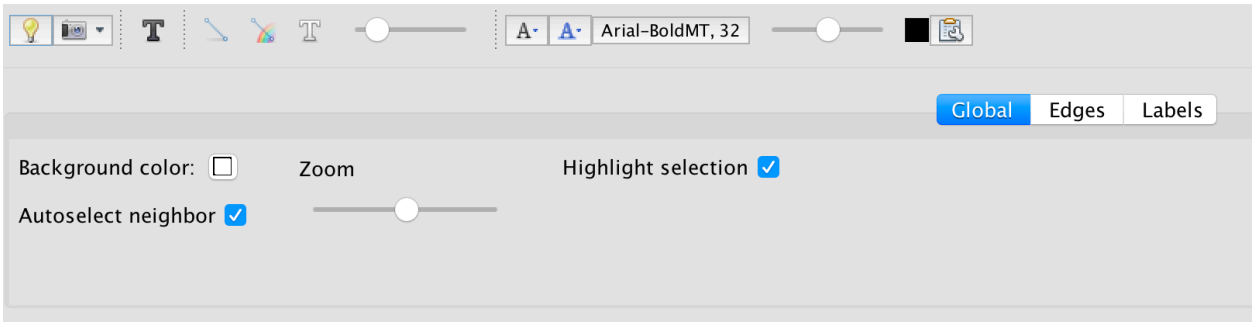
Your graph should now be colored by module.

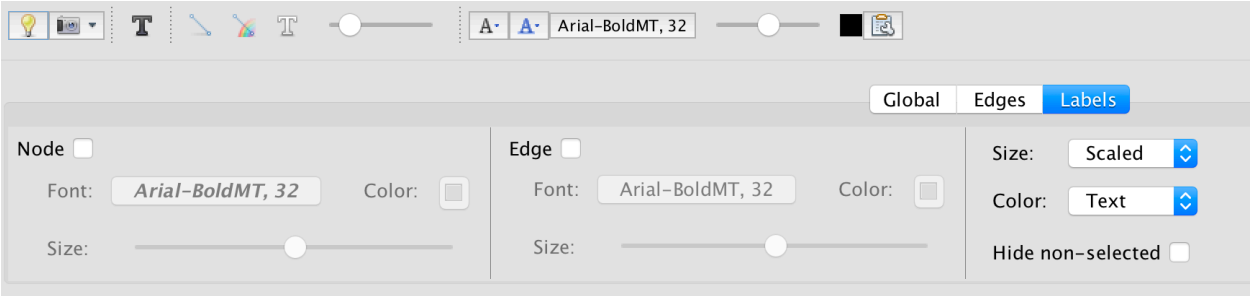
J. Label each node with the unique word it represents. At the bottom, there is a toolbar; click the small button all the way to the right.

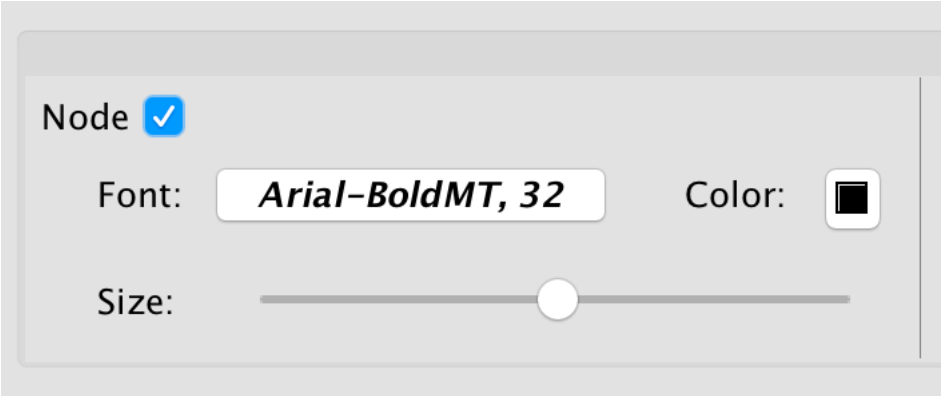


← Small button

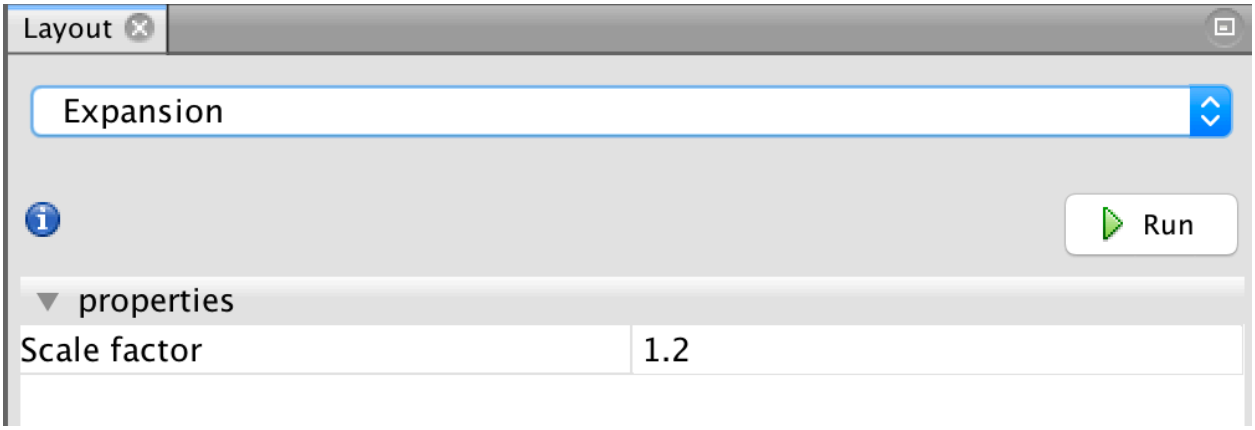
K. Click on the "Labels" tab and click the checkbox for Node.



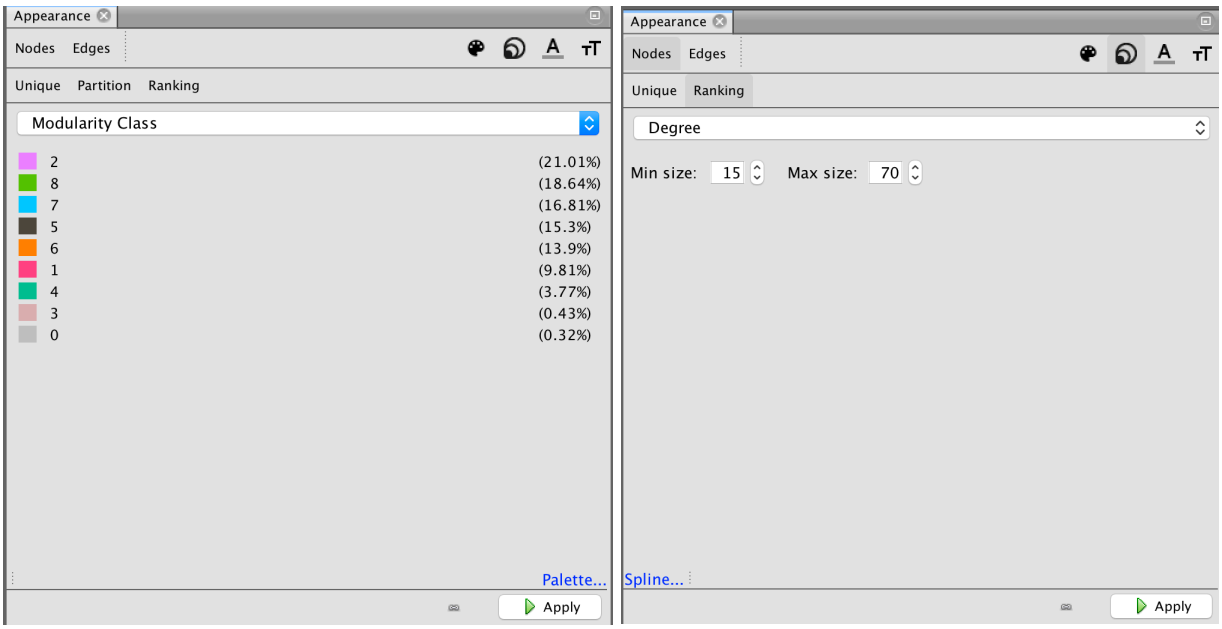




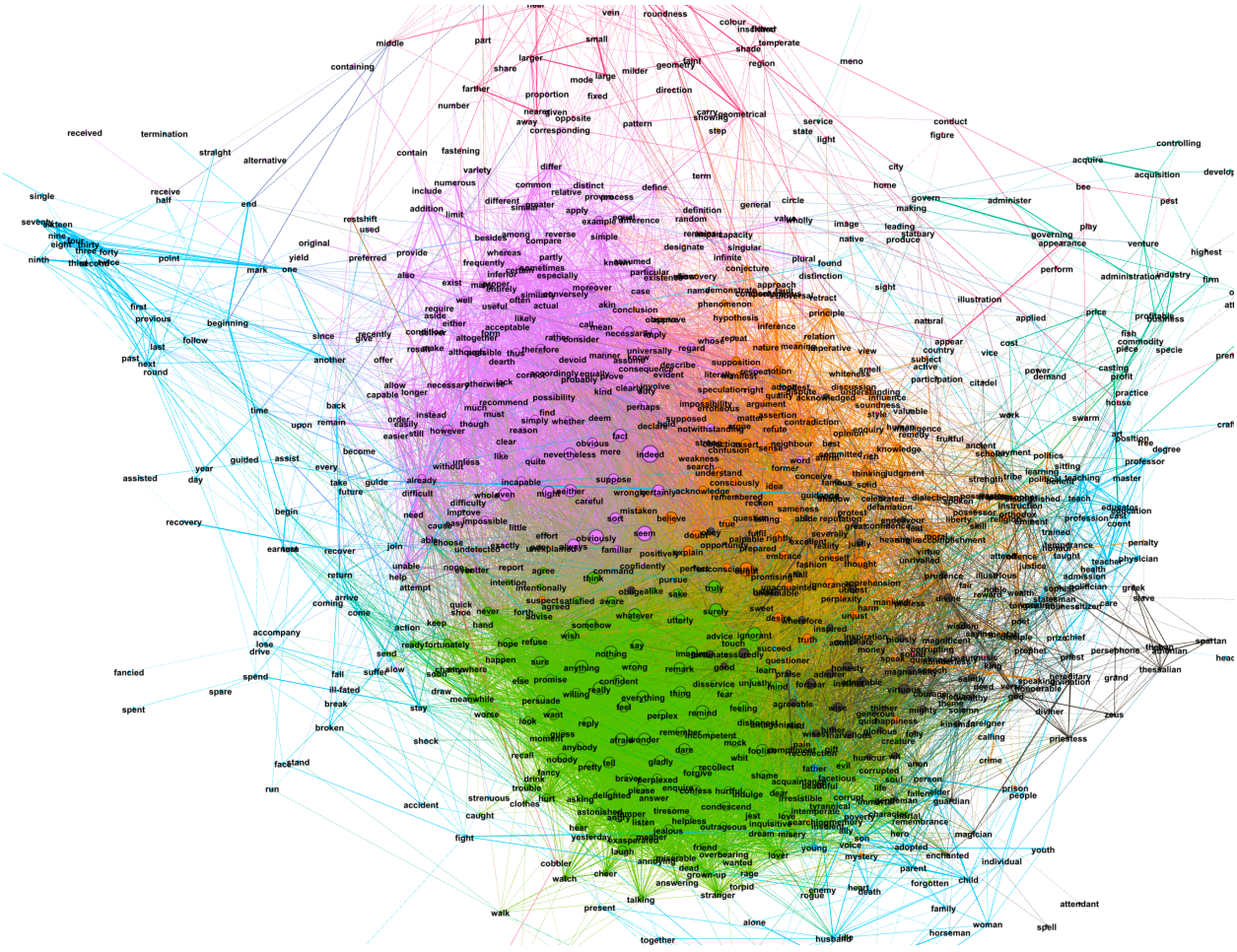
L. Spread out the graph to read the text more easily by going to the "Layout" pane again (the same place where you applied "ForceAtlas 2") and select "Expansion." Click on Run a few times until you're satisfied with the spread of the text labels. Then, zoom out on the graph using your mouse wheel.



M. Finally, navigate back to the "Appearance" pane and click the  button. Click on the "Ranking" button. Then, select "Degree" from the dropdown menu and change the "Min size" and "Max size" to the ones below. Click "Apply".



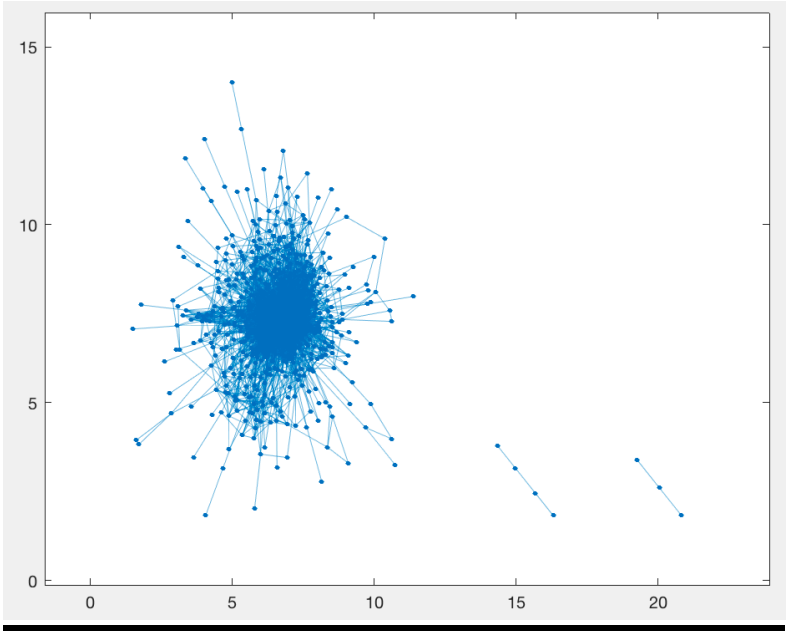
N. Explore your graph! You can scroll over individual nodes to see what words are connected.



O. Repeat step 7 with the coOccurrenceNetwork.gexf. How do the networks differ?

8. <u>Visualize/analyze the network in MATLAB.</u>

A. Open plotAdjmat.m in MATLAB. Change your working directory to where you downloaded "semanticNetworkAdjmat.txt" or copy and paste the full path to that file—e.g.: load("/Users/dalezhou/Downloads/semanticNetworkAdjmat.txt"). Run the code.



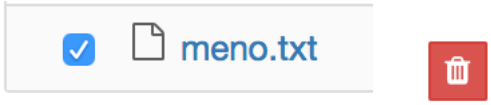
B. Perform any additional analyses of interest.
C. Repeat with "coOccurrenceMatrix.txt."

<u>Constructing a network matrix from your own text</u>

1. <u>Follow the same steps as above, but on Step 3 also do the following:</u>

A. Click on the "text" link to navigate to that directory.

B. Check the box for "meno.txt" and then delete it by clicking the red trashcan button.



C. Click on the “Upload” button and upload your text files (must be in .txt format) and make sure the filename ends with .txt
D. Continue Steps 4 and on as described in the above steps.

More details

1. **Word vector embeddings model**
   Pre-trained model from http://vectors.nlpl.eu/explore/embeddings/en/models/
   Removed all "universal part of speech" tags from the corpus

   =====
   General model information:
   dimensions: 300
   window: 5
   iterations: 5
   vocabulary size: 296630
   id: 3
   ====
   Training corpus:
   ===
   description: English Wikipedia Dump of February 2017
   url: https://dumps.wikimedia.org/
   tool: Wikipedia Extractor
   case preserved: True
   stop words removal: NLTK
   id: 2
   tokens: 2252637050
   lemmatized: True
   tagset: UPoS
   tagger: Stanford Core NLP v. 3.6.0
   NER: True
   public: True
   ====
   Training algorithm:
   name: Continuous Skipgram
   url: https://github.com/RaRe-Technologies/gensim
   tool: Gensim
   version: 2.1
   command: None
   id: 0

Math underlying word2vec: https://arxiv.org/pdf/1411.2738.pdf,
Nice visualization/demonstration of some toy training data: https://ronxin.github.io/wevi/
Google’s original word2vec paper: https://arxiv.org/pdf/1301.3781.pdf