

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)

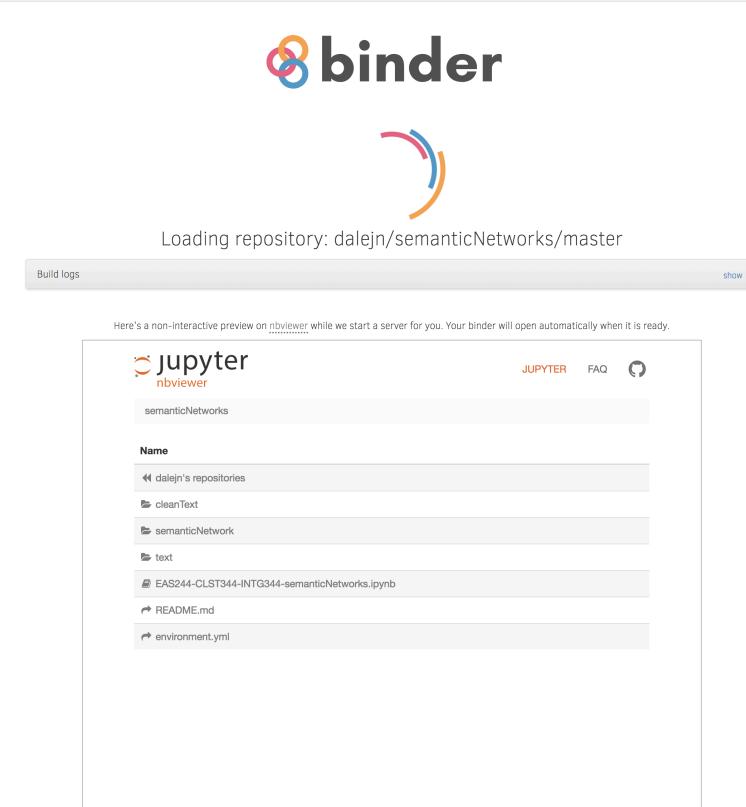
https://www.dropbox.com/s/je4slqywoh8wnr1/enwiki_5_ner.txt?dl=1

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”).

The screenshot shows the Jupyter Notebook dashboard. At the top, there's a logo, the word "pyter", and tabs for "Files", "Running", and "Clusters". On the right, there are "Quit" and "Logout" buttons. Below the tabs, it says "Select items to perform actions on them." There are buttons for "Upload", "New", and a refresh icon. A file list table follows:

	Name	Last Modified	File size
0	/		
0	cleanText	7 minutes ago	
0	conda-bl	4 days ago	
0	semanticNetwork	7 minutes ago	
0	text	7 minutes ago	
1	EAS244-CLST344-INTG344-semanticNetworks.ipynb	7 minutes ago	7.55 kB
0	environment.yml	7 minutes ago	200 B
0	README.md	7 minutes ago	140 B

3. Upload the model downloaded from step 1 to the environment

- A.** Click on “Upload,” go to where you downloaded the file from Step 1 and select it.

This screenshot shows a portion of the Jupyter Notebook interface. At the top, there are buttons for "Upload", "New", and a refresh icon. Below them are sorting buttons for "Name", "Last Modified", and "File size". The file list shows several entries, with the last entry being a large file named "9 minutes ago" which is highlighted with a red box. The file size is listed as "9 minutes ago".

- B.** Press Ok after selecting the model to upload.

Large file size warning

The file size is 809 MB. Do you still want to upload it?

Cancel **Ok**

- 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. Once the upload has completed, click on the text link “EAS244-CLST344-INTG344-semanticNetworks.ipynb” as seen below to open a programming notebook.



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

 A screenshot of a Jupyter Notebook interface. The title bar says 'jupyter EAS244-CLST344-INTG344-semanticNetworks (unsaved changes)'. The toolbar includes File, Edit, View, Insert, Cell, Kernel, Widgets, Help, Run, Code, and other icons. The main area contains two code cells. The first cell's code is as follows:


```
In [ ]: from nltk import sent_tokenize, word_tokenize
from nltk.stem import WordNetLemmatizer as wnl
import nltk, gensim, re, string, glob
from itertools import islice, compress
import itertools
import matplotlib.pyplot as plt
import numpy
import networkx as nx
nltk.download("punkt")
nltk.download("wordnet")

model = "./enwiki_5_ner.txt"
word_vectors = gensim.models.KeyedVectors.load_word2vec_format(model, binary=False)

#####
# Initialize, config & define helpful functions #
#####

translator = str.maketrans(' ', ' ', string.punctuation.replace('-', ' ')) #filters punctuation except dash
lemmatizeCondition = 1
lemmatizer = wnl()

# Function for finding index of words of interest, like 'references'

def find(target):
    for i, word in enumerate(sents):
        try:
            j = word.index(target)
        except ValueError:
            continue
        yield i

# Function for handling the input for gensim word2vec

class FileToSent(object):
    def __init__(self, filename):
        self.filename = filename

    def __iter__(self):
        for line in open(self.filename, 'r'):
            ll = line.strip().split(',')
            ll = [''.join(c for c in s if c not in string.punctuation) for s in ll]
            ll = [num.strip() for num in ll]
            yield ll

    # Function for looking for element x occurs at least n times in list

    def check_list(lst, x, n):
        gen = (True for i in lst if i==x)
        return next(islice(gen, n-1, None), False)
```

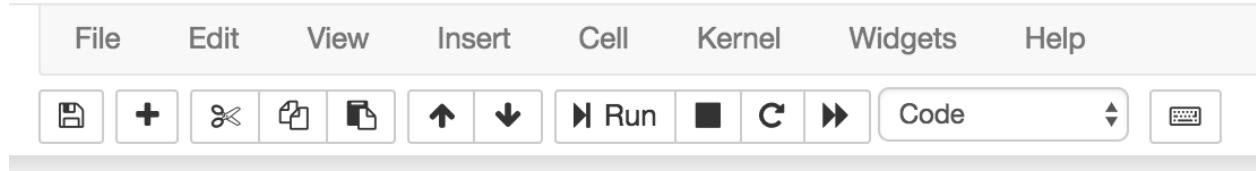
 The second cell's code is as follows:


```
In [ ]: #####
# Read in .txt file(s) from a specified directory #

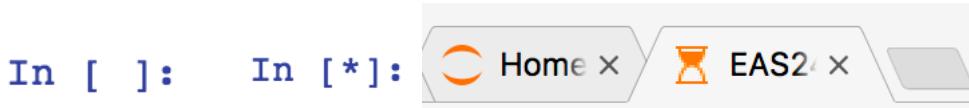
IDs = glob.glob('./text/*')
IDs_subIDs = []
for ID in IDs:
    IDs_subIDs += glob.glob(ID + '/*.txt')
print(len(IDs)) # Print number of files read
```

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_cons = 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 (one at a time) 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.

A screenshot of a file browser interface showing three files listed: "plotAdjmat.m", "semanticNetwork.graphml", and "semanticNetworkAdjmat.txt". Each file has a checkbox next to it, with the first two checked. Below the files is a toolbar with buttons for Duplicate, Rename, Move, Download, View, Edit, and Delete. The "Download" button is highlighted in red.

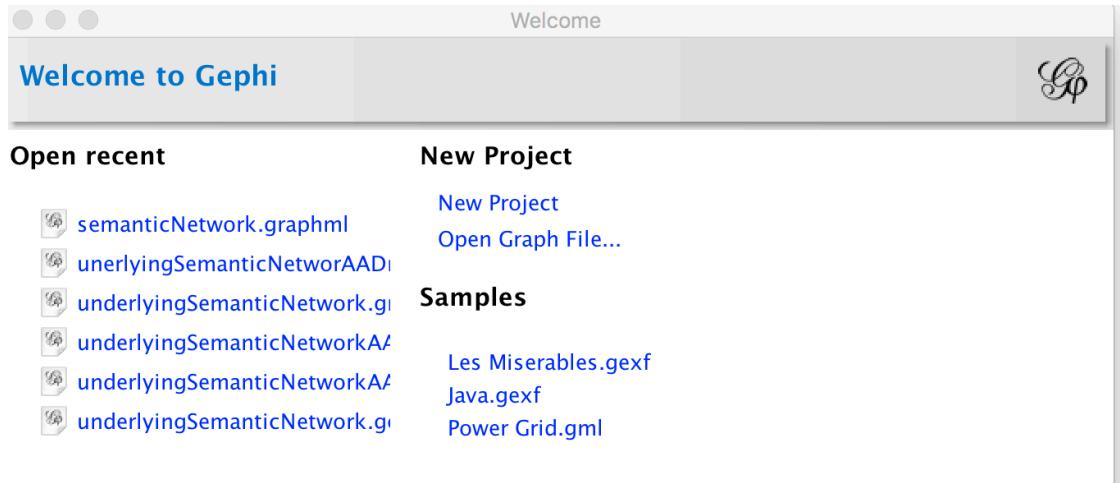
- C. Press the Quit at the top right.



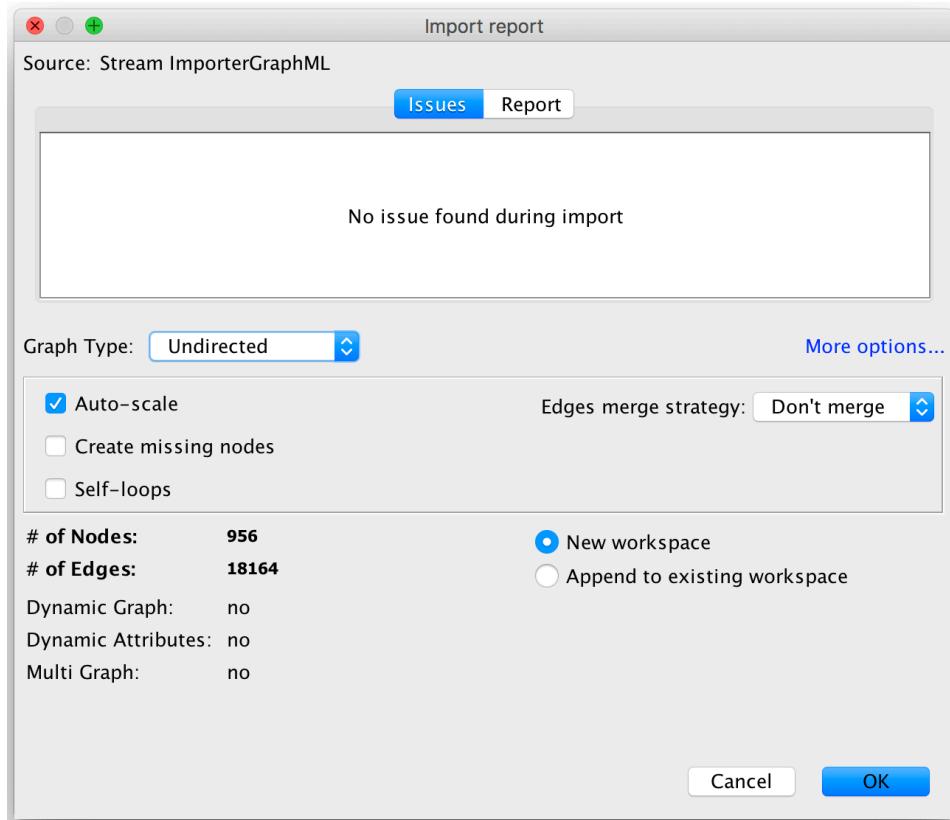
7. Visualize the network in Gephi

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

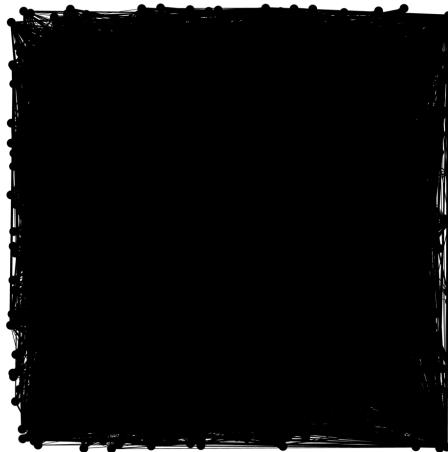
You will see a Welcome screen like below. Click on “Open Graph File...” If you don’t see the welcome screen, go to File and Open on your computer’s toolbar.



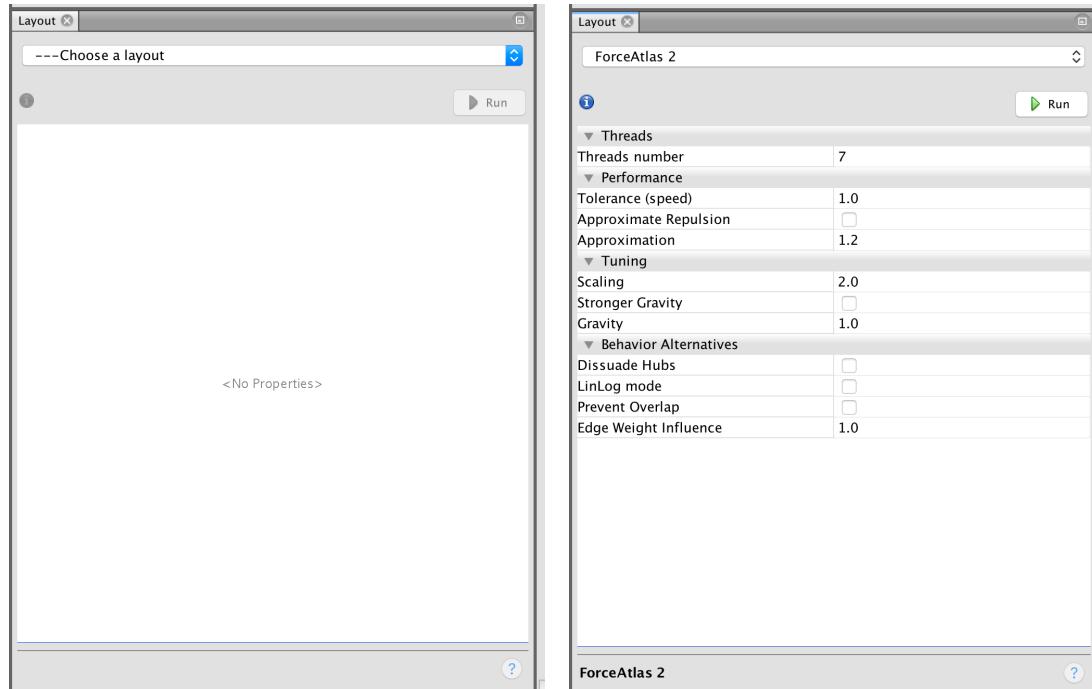
- B. Navigate to where you downloaded semanticNetwork.graphml and open this file.
- C. A new screen should pop up. Change the settings so they match with the ones below and then press OK.



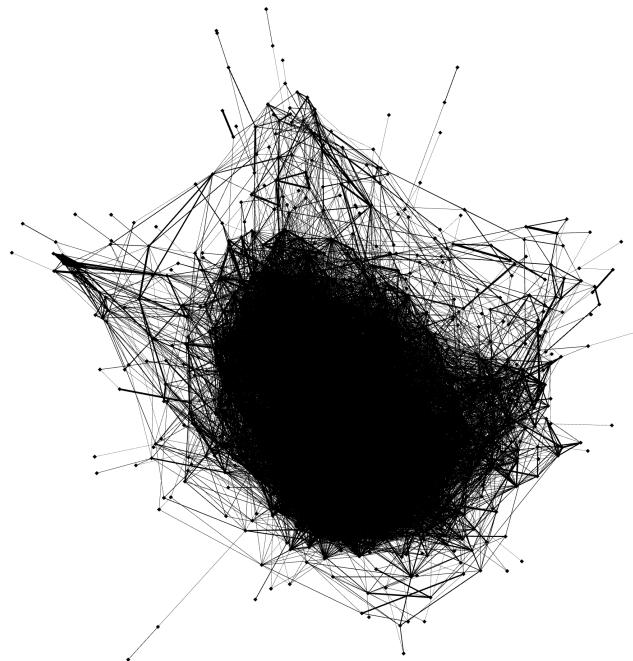
- D. The network should be visible. If not visible, you may need to go to Gephi's menu at the top and select Window > Graph. You can scroll your cursor over nodes to see connections. We'll now clean this up to make it more interpretable.



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.

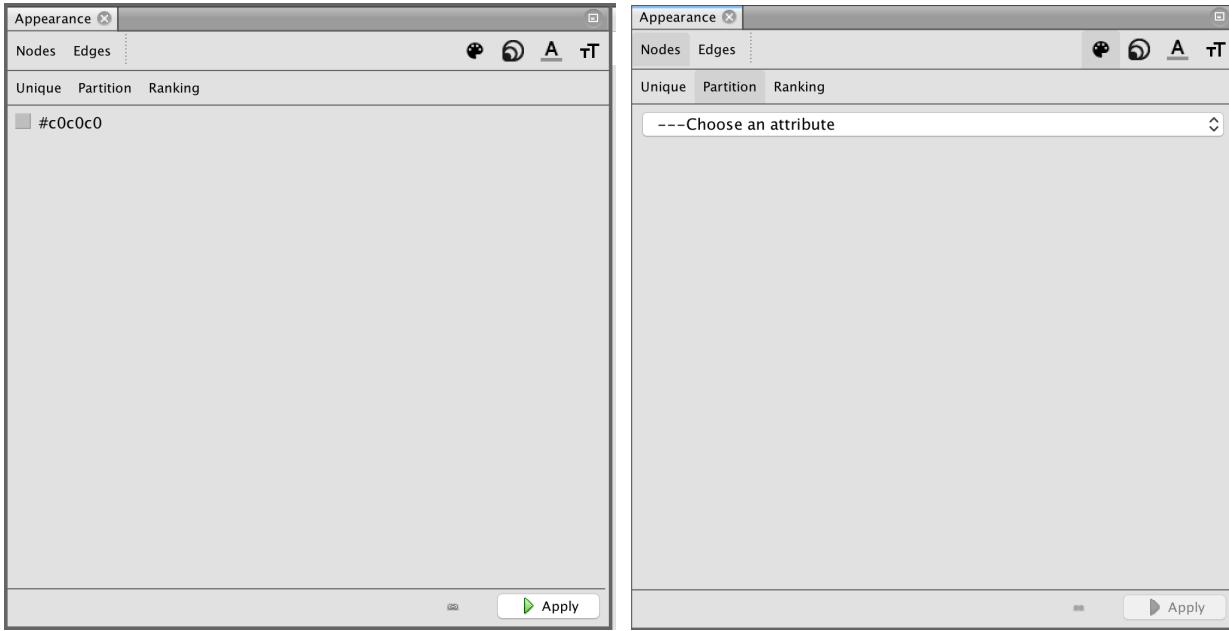


G. On the right-side toolboxes, click on the “Statistics” tab and then click “Run” for the Modularity measure. A box will pop up; click OK. Another pop-up labeled HTML Report will appear; click Close.

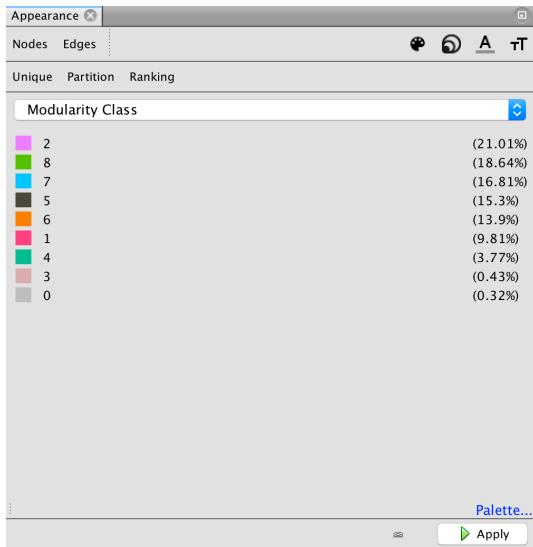
The screenshot shows the NetworkMiner interface with the following components:

- Context Window:** Displays network statistics: Nodes: 928, Edges: 18164, and Type: Undirected Graph.
- Toolbox:** Located on the right, it contains tabs for Filters, Statistics (selected), and Settings. The Statistics tab lists various metrics with "Run" buttons:
 - Average Degree
 - Avg. Weighted Degree
 - Network Diameter
 - Graph Density
 - HITS
 - Modularity
 - PageRank
 - Connected Components
- Modularity Settings Dialog:** A modal window titled "Modularity settings" with the following options:
 - Modularity:** Community detection algorithm.
 - Randomize: Produce a better decomposition but increases computation time.
 - Use weights: Use edge weight.
 - Resolution:** Lower to get more communities (smaller ones) and higher than 1.0 to get less communities (bigger ones). Set to 1.0.
 Buttons at the bottom: Cancel, OK.
- HTML Report Dialog:** A modal window titled "HTML Report" with buttons for Print, Copy, Save, and 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.”



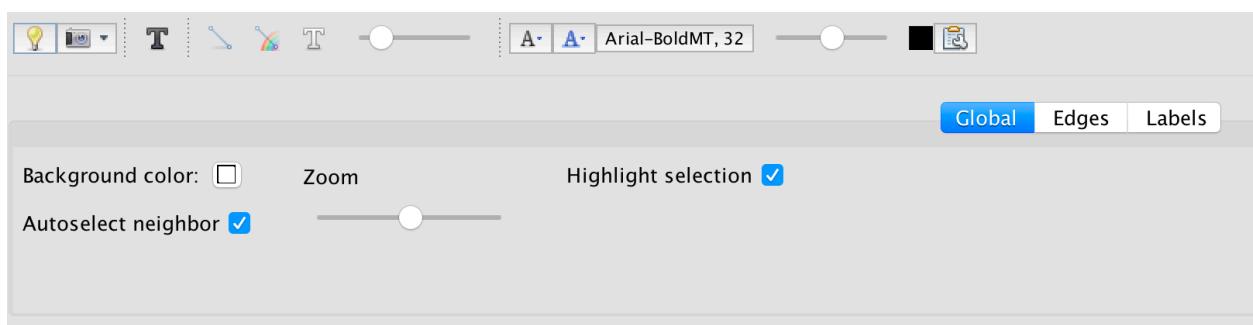
Your graph should now be colored by module.

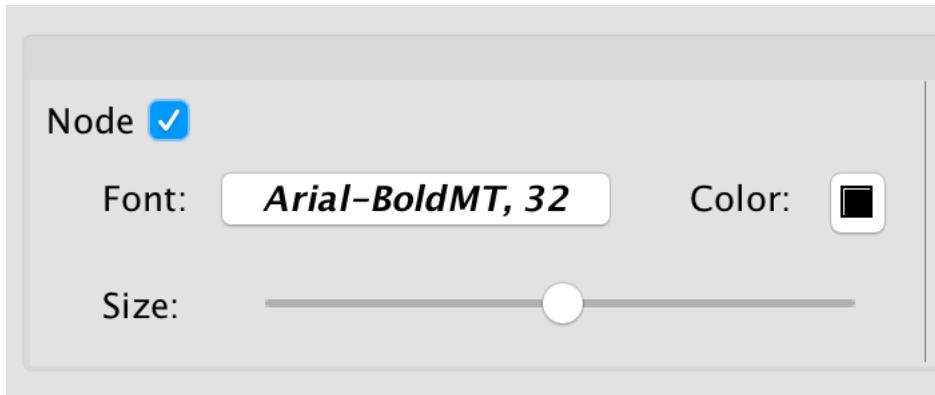
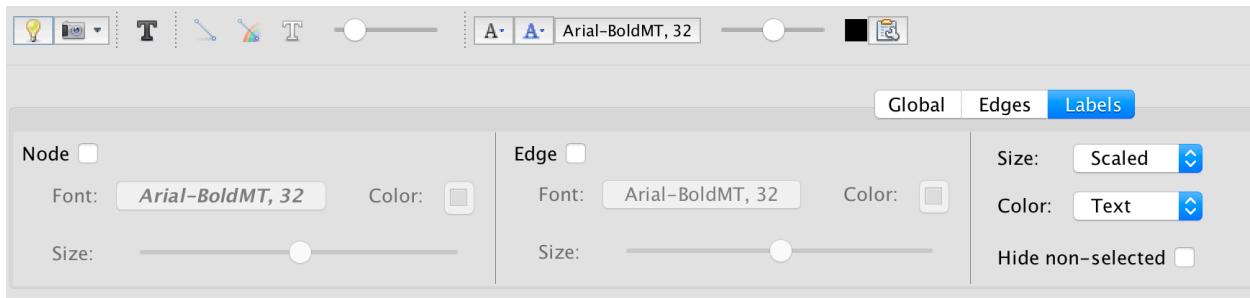


- J. Label each node with the unique word it represents. At the bottom, there is a toolbar; click the small button all the way to the right.

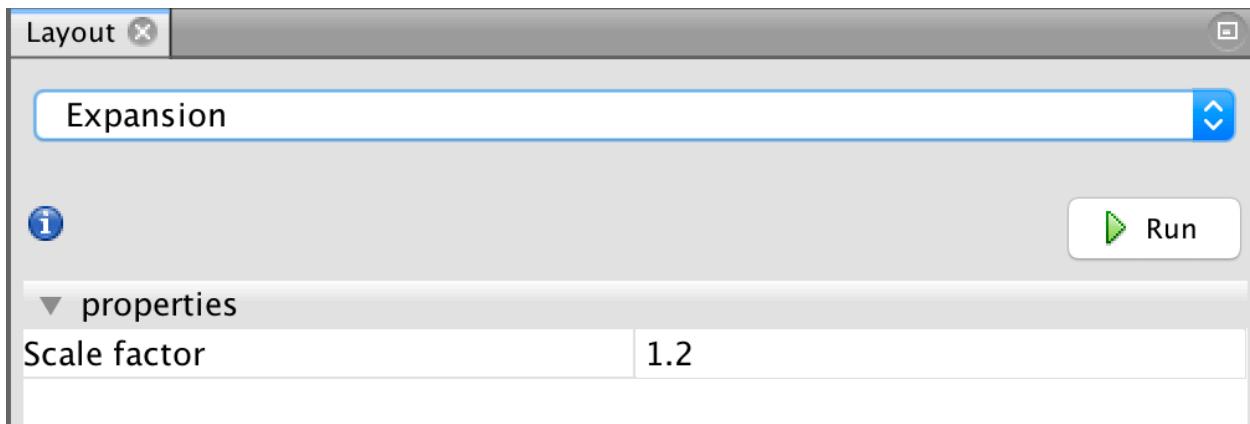


- K. Click on the “Labels” tab and click the checkbox for Node.

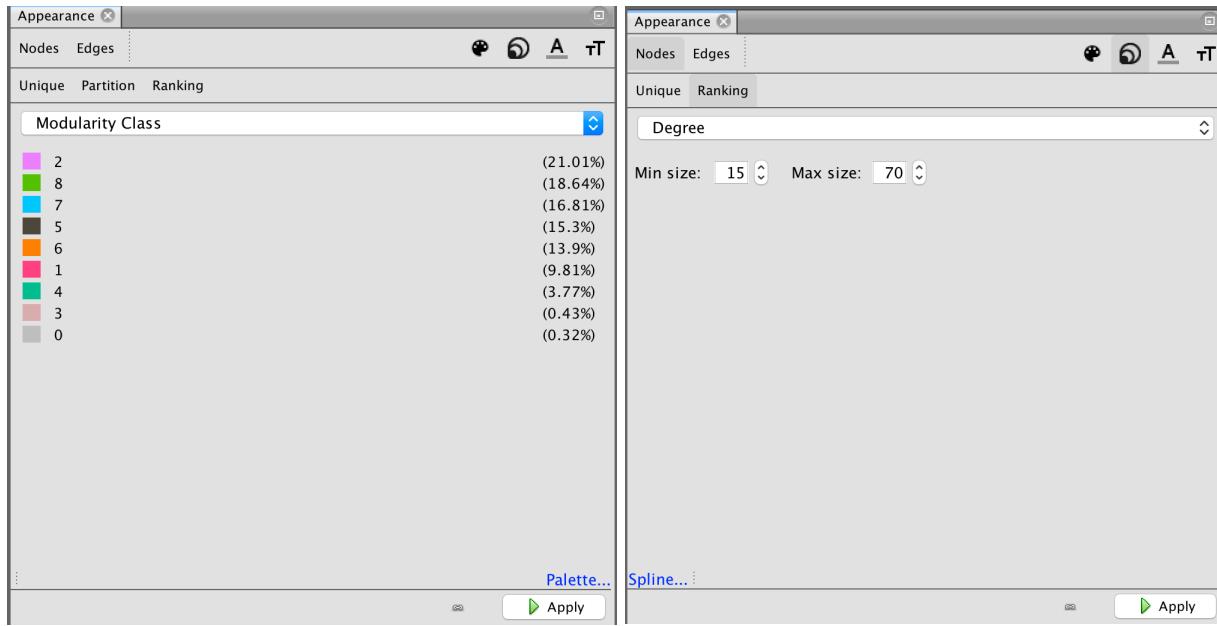




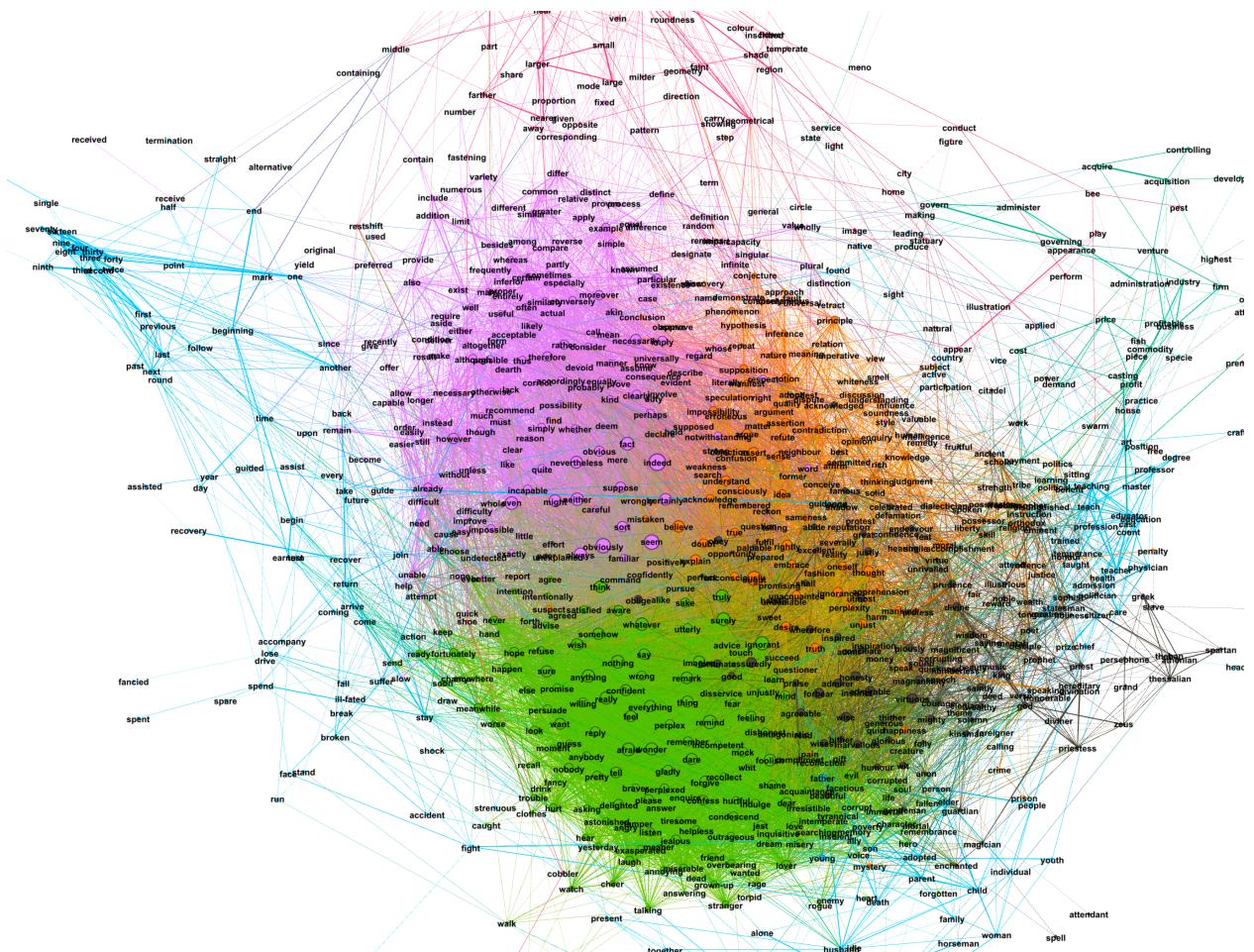
- L. Spread out the graph to read the text more easily by going to the “Layout” pane again (the same place where you applied “ForceAtlas 2”) and select “Expansion.” Click on Run a few times until you’re satisfied with the spread of the text labels. Then, zoom out on the graph using your mouse wheel.



- 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”.



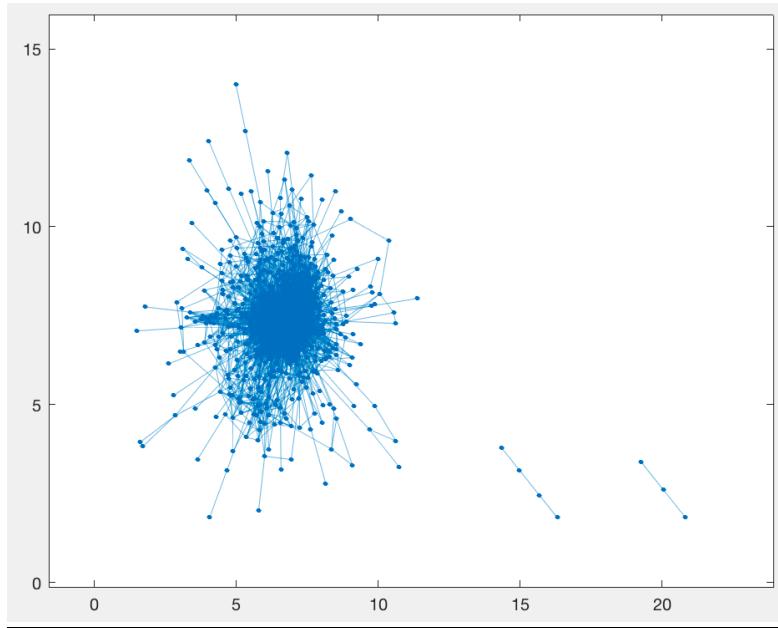
N. Explore your graph! You can scroll over individual nodes to see what words are connected.



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

8. **Visualize/analyze the network in MATLAB.**

- 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.”

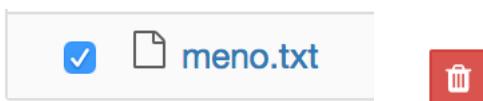
Constructing a network matrix from your own text

1. **Follow the same steps as above, but on Step 3 also do the following:**

- 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>