Graphs and Network Analysis

1. **Install Gephi** Hopefully Gephi is already installed. If not, visit http://gephi.github.io/, download and install.

2. You will work in groups of 2-3. Take notes on your answers to the italicized questions and save figures where needed.

Social Network Analysis

3. **Download data** Go to the course website > Labs and download "Lab 8 Networks" and unzip it.

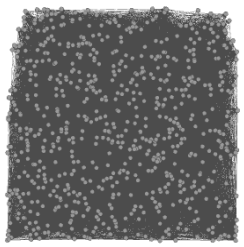
   This data contains friends from Facebook. Facebook data was collected from survey participants using an app. The data has been anonymized by replacing the Facebook-internal ids for each user with a new value.

   In this graph:
   nodes = people
   edges = friendship connections

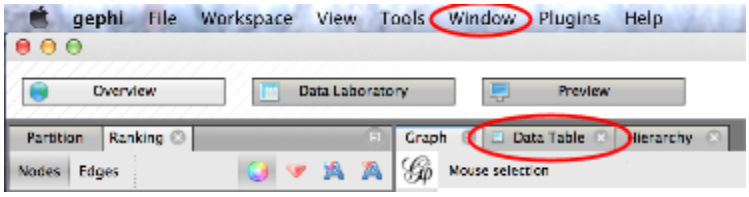
4. **Open Gephi**

5. **Open the data** you just downloaded
   - Control - O, or File > Open
   - Navigate to where you downloaded the “Lab 8 Networks” folder
   - There are three options: open **facebook_network.edges**

6. The data has many types of information, we will only use the edges. Initially, the graph will look like a large blob.



   If you see a tab at the top called “Data Table,” click it.
   If you don’t see this tab, click “Window” > “Data Table”



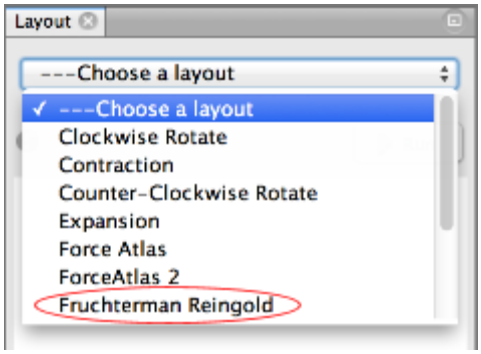
A. *Describe the kinds of data you see. Are relationships obvious in the table?*

6. **Degree** Click on “Statistics” on the right hand side, if it is not already open. If you don’t see it click “Window” > “Statistics” at the top.

   Calculate “Average Degree” by clicking “Run” next to that option.

B. *Examine the chart that pops up. Do most people in this social network have a similar degree? Did you have a guess from the graph? Probably not, because the graph is very crowded. We will update that in the next step. Why are the values different between In-Degree, Out-Degree, and Degree Distribution?*

7. **Clusters** Return to the Graph tab, where you can see the graph. On the left side, in the “Layout” tab (if you don’t see “Layout” click “Window” > “Layout”), click the button with “---Choose a layout” and select “Fruchterman Reingold”



   Keeping the default options, press “Run” and watch as the graph re-adjusts its layout. You may press stop at any point (and can restart it by pressing run again).

C. *Do you see clusters in the data? What do you think such clusters mean in a social network?*

D. *Based on the appearance of the graph, do you have a guess for the diameter?*

8. Under “Statistics,” run “Network Diameter.” Leave the default selections for now, though you can recalculate for the undirected version if you’d like later. A screen will pop up with information and graphs.

E. *What is the diameter of the graph? Are you surprised by that value? Or is it about what you expected? What is the average path length? Do you recall what average path length means? Why is the value so low?*

9. Play around with the other statistics you can measure, different layouts, filters, etc. Are there additional calculations that are relevant or can add insight?

10. **Bonus:** Tools exist to help you extract your own Facebook network data (Netvizz, http://givememydata.com/, http://socialnetimporter.codeplex.com/). If you have time, try extracting your network, analyzing it and comparing it to the example data.

Internet Analysis

1. **Download data** Again, the data is located on the course website > Labs under "Lab 8 Networks"

   This data has information about the internet. However, it does not contain web pages and links, but instead contains router information, that is, the hardware backbone of the internet. Routers are "networking devices, commonly specialized hardware, that forward data packets between computers or networks."[1]

   In this graph:
   nodes = routers
   edges = ethernet/broadband/etc connections

2. **Open Gephi**

3. **Open the data** you just downloaded
   - Extract the file (unzip it)
   - In Gephi, use Control - O, or File > Open
   - Navigate to where you downloaded the file
   - Open internet_routers-22july06.gml

4. **Clusters** In the graph tab, again look for the clusters in the data. On the left side, in the "Layout" tab (if you don't see "Layout" click "Window" > "Layout"), click the button with "---Choose a layout" and this time select "Force Atlas 2"

   Keeping the default options, press "Run" and watch as the graph re-adjusts its layout. You may press stop at any point (and can restart it by pressing run again).

   E. Do you see clusters in the data? What do you think such clusters mean in a router network? How dense is the graph compared to your expectations? Do you have a guess for the diameter of this graph?

5. Under "Statistics," run "Network Diameter." Leave the default selections for now, though you can recalculate for the undirected version if you'd like later. A screen will pop up with information and graphs. This graph is large and the calculation may be slower than last time.

   F. What is the diameter of the graph? Are you surprised by that value? Or is it about what you expected? What is the average path length? How does the diameter and average path length compare to the social network graph you just analyzed?

[1] http://en.wikipedia.org/wiki/Router_(computing)

6. **Degree** Click on “Statistics” on the right hand side, if it is not already open. If you don’t see it click “Window” > “Statistics” at the top.

   Calculate “Average Degree” by clicking “Run” next to that option.

   C. *Examine the chart that pops up. Do most routers in the 2006 Internet have a similar degree? Did you have a guess from the graph?*

7. Play around with the other statistics you can measure, different layouts, filters, etc. Are there additional calculations that are relevant or can add insight?

8. **Bonus:** See if you can find an additional data source online, either for websites with in- and out-bound link information or for routers with some geographic information. Visualize the data you find and see whether the information helps with understanding the graph, and whether it differs substantially from the router information provided here.

Language Analysis

2. **Download data** Data is located on the course website > Labs under "Lab 8 Networks"

   This data contains adjectives and nouns from “David Copperfield” by Charles Dickens, by co-occurrence.

   In this graph:
   nodes = words (nouns or adjectives - 60 most commonly occurring nouns in the novel and the 60 most commonly occurring adjectives)
   edges = co-occurrences (used in next to each other in text)

3. **Open Gephi**

4. **Open the data**
   - Extract the file (unzip it)
   - In Gephi, use Control - O, or File > Open
   - Navigate to where you downloaded the file
   - Open "word_adjacencies.gml"

7. **Clusters** In the graph tab, again look for the clusters in the data. On the left side, in the “Layout” tab (if you don’t see “Layout” click “Window” > “Layout”), click the button with “---Choose a layout” and this time select “Force Atlas 2”

   Keeping the default options, press “Run” and watch as the graph re-adjusts its layout. You may press stop at any point (and can restart it by pressing run again).

   Try again using “Fruchterman Reingold” with the default settings.

   Finally, use “Fruchterman Reingold” and adjust the “Gravity” setting to 200.

F. *Do you see clusters in the data? What do you think that means? How dense is the graph compared to your expectations? Do you have a guess for the diameter of this graph?*

8. Under “Statistics,” run “Network Diameter.” Leave the default settings. A screen will pop up with information and graphs.

G. *What is the diameter of the graph? What is the average path length? How do the values compare to the graphs you analyzed previously? What does this suggest about the graph you’re analyzing here?*

9. **Degree** Click on the “Ranking” tab on the left hand side, if it is not already open. If you don’t see it click “Window” > “Statistics” at the top.

   Click on “Nodes” and select “Degree” from the drop-down menu labeled “---Choose a rank parameter.” Then click “Apply”

   D. *Examine the changes to the graph. Do most words have a similar degree? How easy or difficult is this to interpret and understand, when compared to the charts you calculated for the previous graphs? Can you identify any interesting features from the data using this chart?*

10. Play around with the other statistics you can measure, different layouts, filters, etc. Are there additional calculations that are relevant or can add insight?

11. **Bonus:** The authors of the original paper include the following two visualizations. Can you explain the two? How do you interpret (b)? Can you reproduce these visualizations?

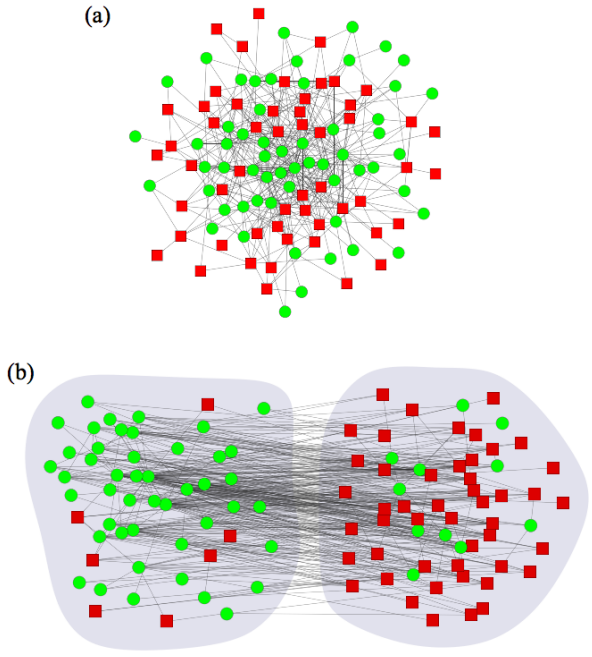


FIG. 7 (a) The network of commonly occurring English adjectives (circles) and nouns (squares) described in the text. (b) The same network redrawn with the nodes grouped so as to minimize the modularity of the grouping. The network is now revealed to be approximately bipartite, with one group consisting almost entirely of adjectives and the other of nouns.