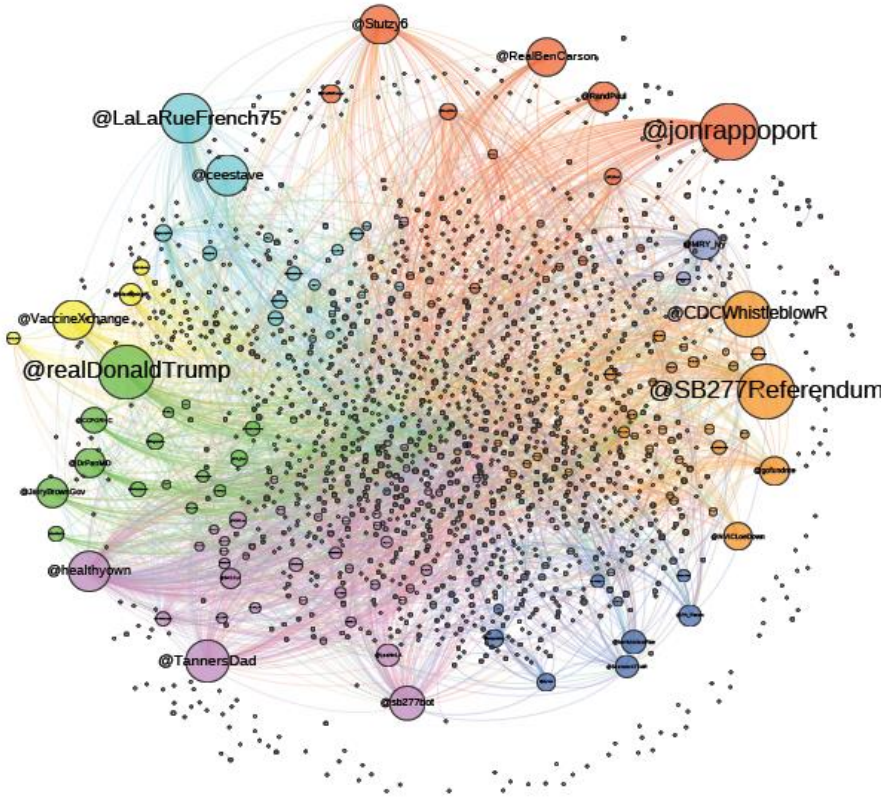
Web Exercise 06: Social Network Analysis (Gephi)

DUE Date: November 05 (Thursday), 5:30pm (on Blackboard).
Grade: 10 points (1.5 week)

Acknowledgement: Parts of this exercise was created by Elias Issa (GA of the HDMA Center).

1. **Download the Gephi Software** from http://gephi.github.io/users/download/ into your local computers.

Gephi is an open source network analysis tool written in Java on the NetBeans platform (cited from Wikipedia). It uses a 3D render engine to display large networks effectively and to speed up the exploration. A flexible and multi-task architecture brings complex data sets into valuable visual results.



There are three official tutorials available in the Gephi website.

https://gephi.org/users/

Please complete all three tutorials first and learn some basic concepts of Gephi. (Require 60 mins total).

The next task is to use the real world data (Twitter data) to create a Social Network Analysis. The example data we will use in this exercise is the Twitter data collected from the Search APIs with keywords related to the "vaccine exemption" topic. We will need to create the **EdgeList** table (the list of From Nodes and To Nodes) in order to create the social network.

STEP1: Use the Google Refine to create the EdgeList Table.

In the Google Shared Drive Folder (under the "Web-Exercises" folder), copy the "Tweets-SNA-examples.xlsx" into your local drive in a new working folder called "lab06".

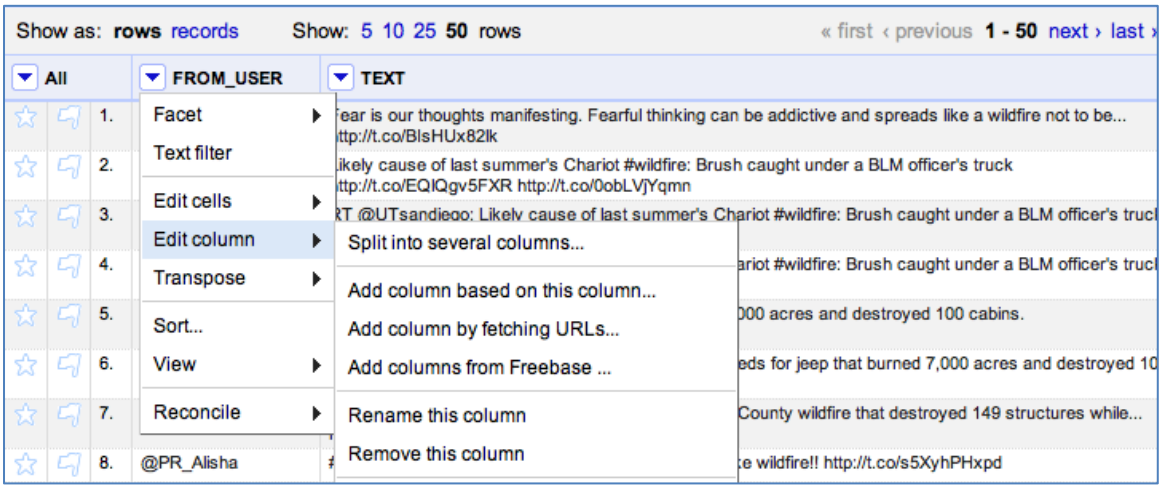
- Copy the columns "TEXT" and "FROM_USER" in the EXCEL file and paste it in a new excel file
- Save the new file as a "lab06SNA.csv". (CSV file format) in the lab06 folder.
- Download "**Google Refine**" using the website https://github.com/OpenRefine/OpenRefine/releases.
- Select the appropriate version (for Windows or for Mac) and copy the files into your local disk.

- Once you downloaded Google Refine, click on the "**openrefine.exe**" in the program folder. Then it will open a web-based user interface in a web browser.
- In **This Computer**, click on "**Choose File**" to import the "**lab06SNA.csv**" file that you created.
- Click on "**Next**", then in the "Columns are separated by", check the "commas (CSV)" box, then click "**Create Project**".

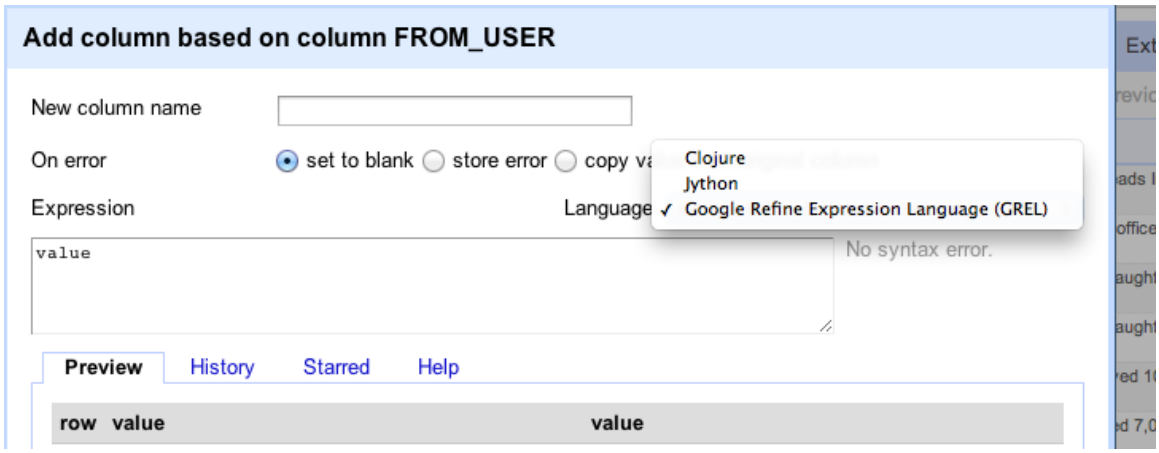
ADDING "@" IN FROM OF EACH USER

In the original tweet data, the FROM_USER name is stripped of the "@" in front of the username that is used in tweets - since we want to extract the usernames from tweet text later, we need to add a new user name column with @. This will involve a tiny bit of programming

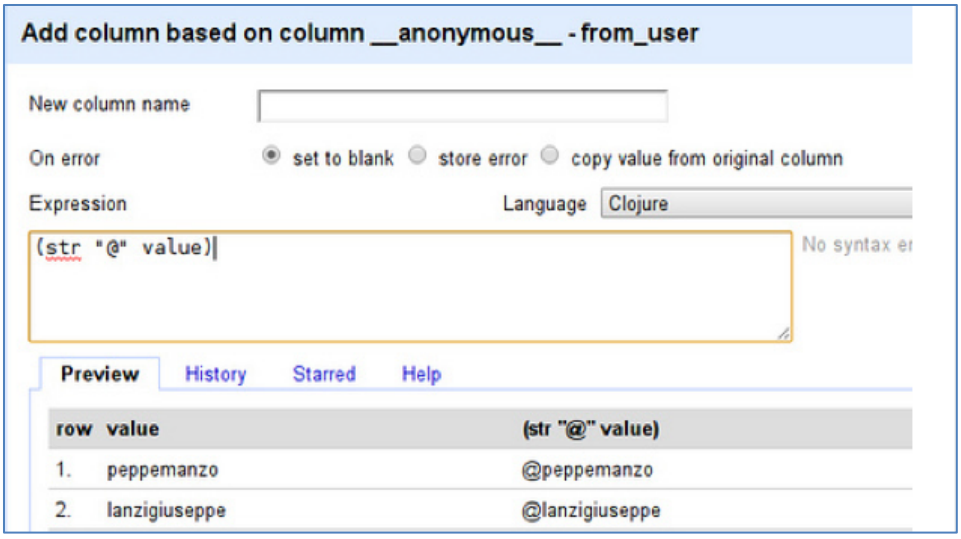
- On your FROM_USER column Select "Edit column → add column based on this column".



- Select **Clojure** as your **Language** to write a small code to filter your data



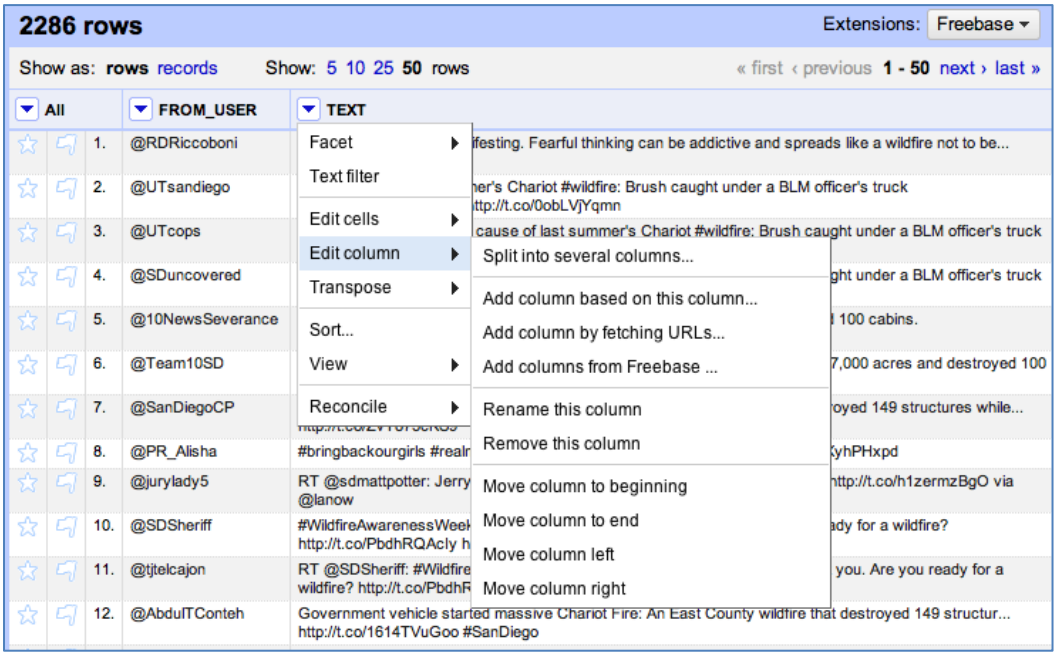
- We want to prepend "@" to each user name (here "value" refers to the value in each row).
- Enter **(str "@" value)** into the expression field (please include the "()").



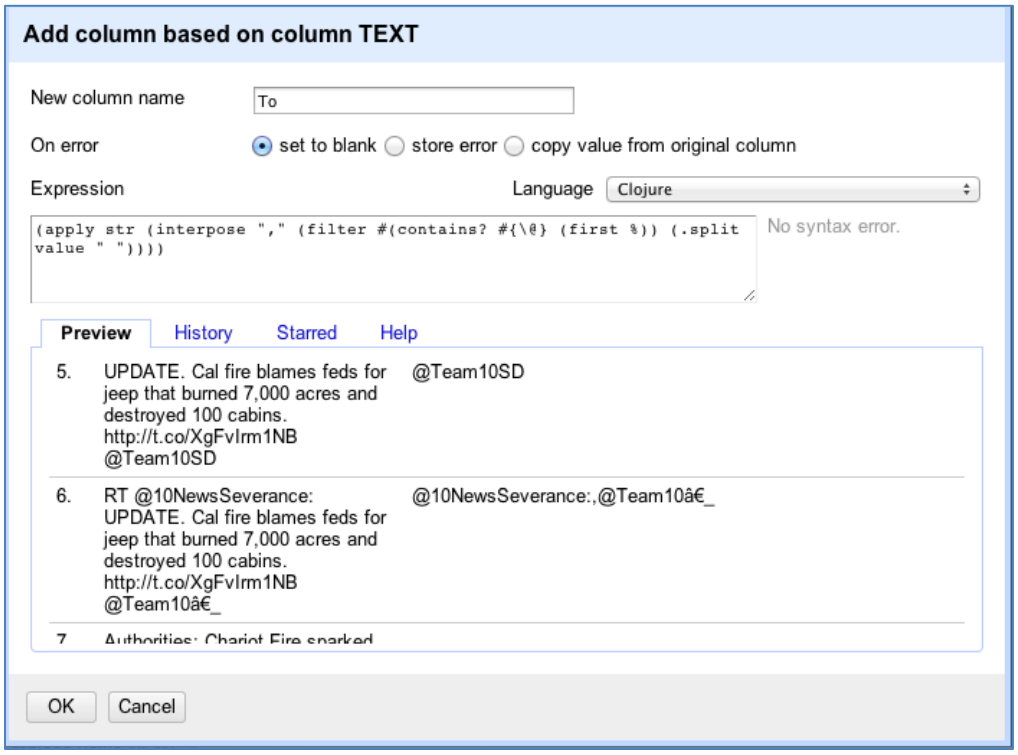
- See how the value has been changed from peppemanzo to @peppemanzo - what happened? In clojure "str" can be used to combine multiple strings: (str "@" value) therefore combines the string "@" with the string in value - what we wanted to do.
- Now simply name your column (eg. "**From**") and click on OK you will have a new column.

EXTRACTING USERS FROM TWEETS

- Let's start with adding a new column based on the "**TEXT**" column→ Edit column → "Add column based on this column".



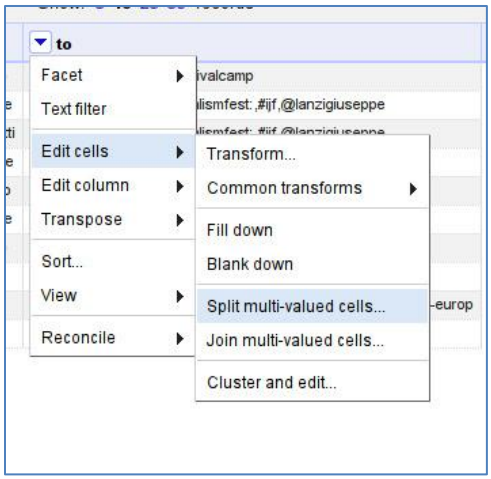
- What we want to do is to split the tweet into words then pick the "@ user" from the tweets- we can do so by entering **(apply str (interpose "," (filter #(contains? #{ \@} (first %)) (.split value " "))))** in the expression field (make sure your language is still **clojure**)



- We now extracted who talks to whom; name your column "**To**" and click "OK" to continue.

Cleaning up

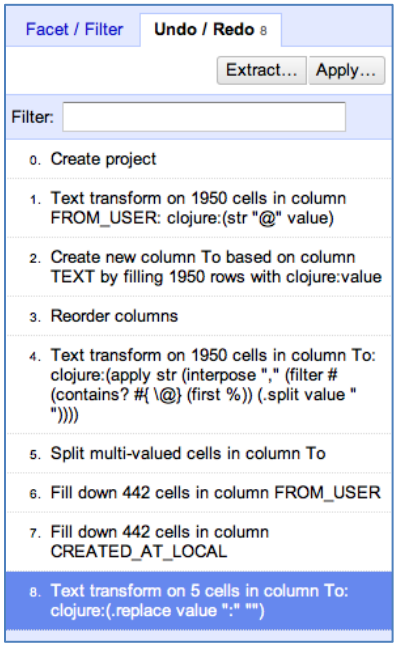
- First, let's remove the two columns we don't need anymore: the TEXT and the original FROM_USER column - do this with "all → edit columns → remove and reorder columns.
- Move your "From" column to the first column and the "To" column as the second column.
- Press "OK".
- Now, let's split up the "To" column so we have one row in each retweet or mention links : Select "**To**" → "**Edit cells**" → "**Split multi valued cells**" enter "**,**" as separator.



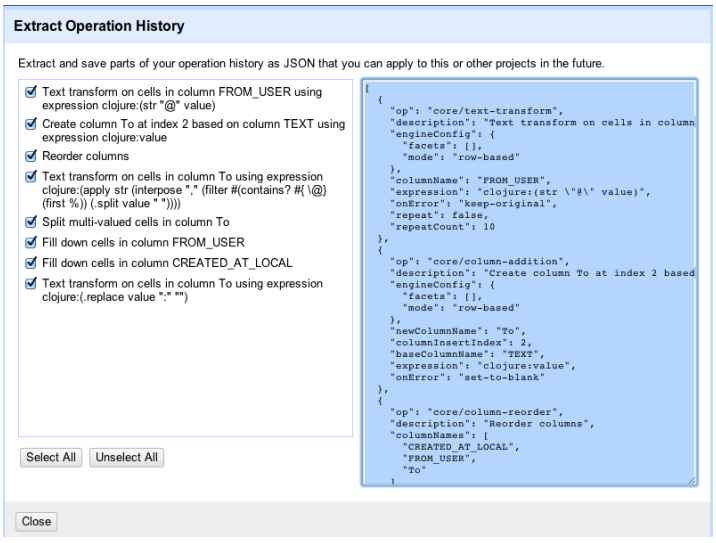
- Make sure to switch back to “rows” mode.
- Now let’s fill the empty rows: Select “**From**” → “Edit cells” → “Fill down”
- Notice that there are some characters in there that don’t belong to names (e.g. “:” ?) Let’s remove them.
- Select “**To**” → “**Edit cells**” → “**Transform...**”
  - To replace our transformation is going to be **(.replace value ":" "")**
  - Press “OK”.

You can export your final network file by clicking on “Export” -→ “Excel (.xls)” → It will save an “lab06SNA-csv.xls” in the Web Download folder. Copy the new file into your working folder (lab06).

- Google refine allows you also to extract the code of all the previous steps and apply it to different files that have the same headings.
- Click on the “Undo/Redo” menu on the top left of the window. Then Press “**Extract**”.
-



- A frame will be shown including a code on the right side. Right click on the code as illustrated in the figure.



- Close this frame, and click on the button “Apply” next to the “Extract” button and paste the code.
- The procedure can be applied to all other files that have same format.

Use the EXCEL for the final data cleaning procedure.

Now you can open the **lab06SNA-csv.xls** file in Excel. You may notice that some "To" fields are empty. The reason is that some tweets don't have any mentions or retweets. We should remove these empty rows before we conduct the social network analysis for Retweets or Mentions.

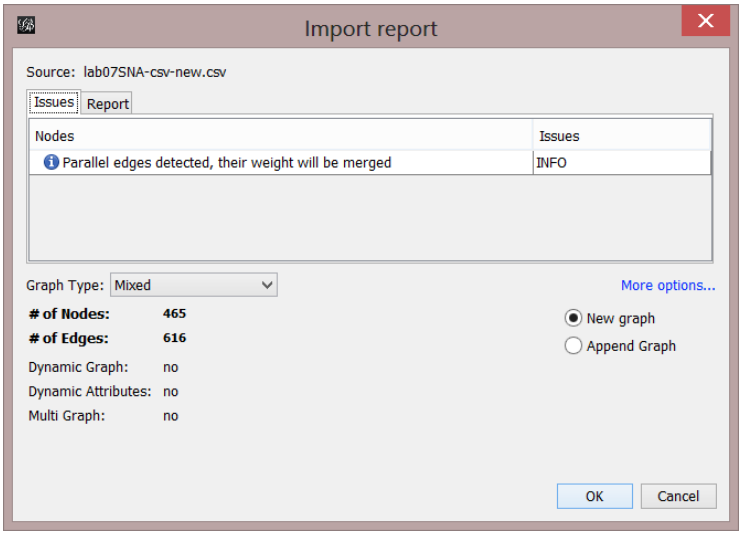
To **clean the data**, select the "To" column first. Then select "**Sort & Filter**" icon in the top icon function, select "Sort A-Z" – choose "**Expand the selection**". -→ "Sort". Then **delete these empty rows** (in To column), and the rows with only "@" inside the "To" column. Scroll down to the bottom of the row, delete the last row from "To" column.

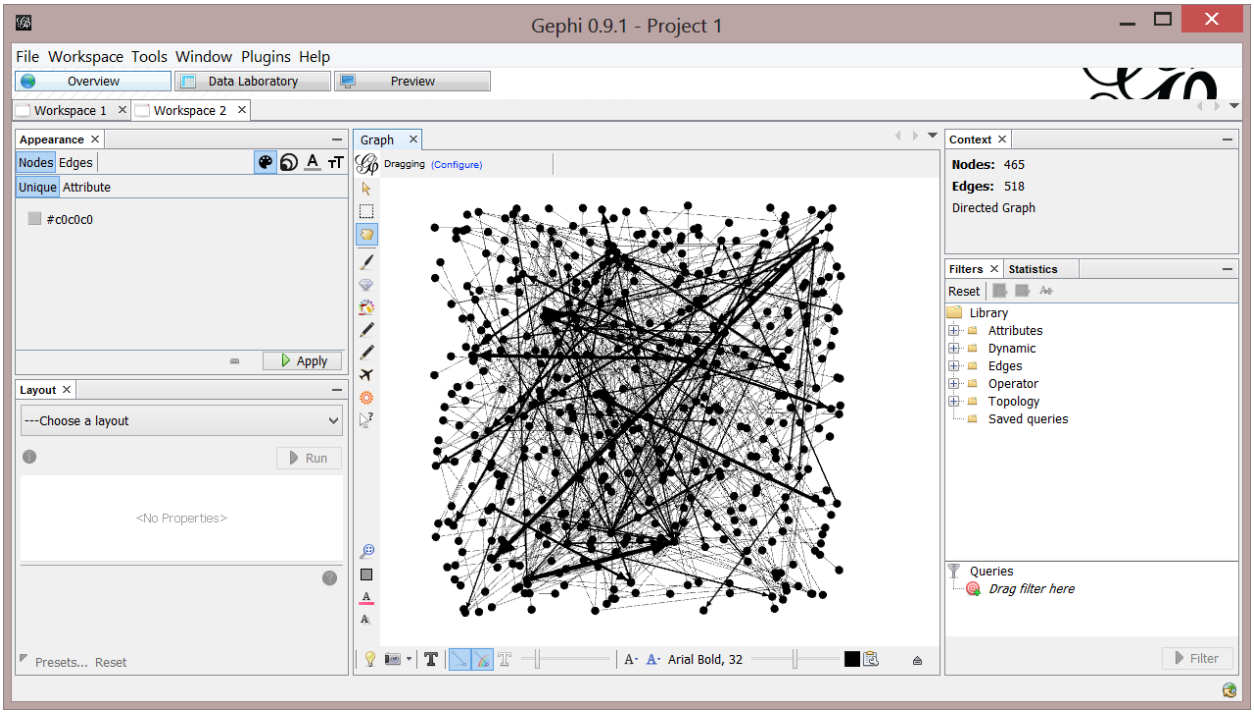
Save the new file as "**lab06SNA-new.csv**" (using the CSV format as output format in Excel).

Now your new dataset (EdgeList) is ready for the Gephi to create social networks.

STEP2: Use the Gephi to Create Social Network Maps

First, launch the Gephi software. Then click on "**New project**" first, then click on "File" menu on the top → "Open…" → select the "**lab06SNA-new.csv**". Click on "OK".





Switch Gephi to the "data laboratory" view.

If you click on the "Data Laboratory" tab near the top of the window, you'll see a table with all of your contacts. Click on "Edge" on the top left-hand side to see the connections between them.

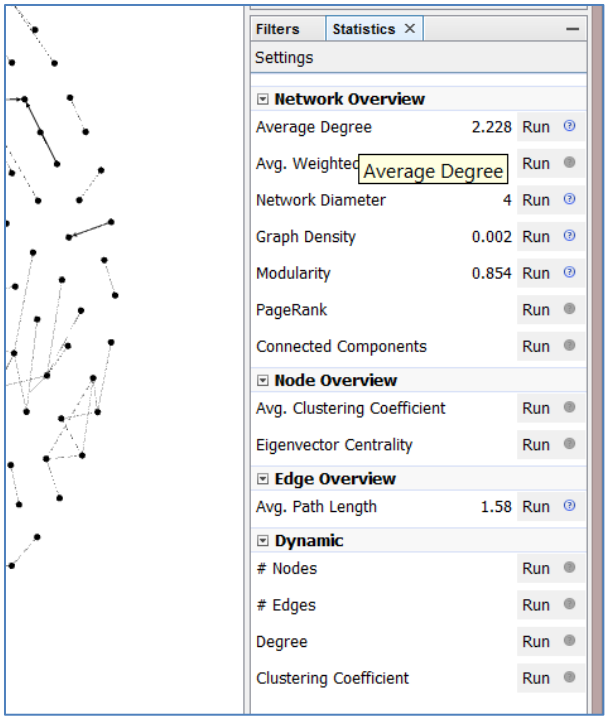
- **Let's start the design.** Click on "**Overview**" at the top left-hand side to return to the SNA graph.

Setup the layout

- It's pretty messy right now so let's add some nice layout display. To create a new layout, simply select the algorithm in Layout option and click "**Run**" - see how the graph changes.
- You can try "**Force Atlas**" or "**Fuchterman Reingold**" to get different display results. You can select "label adjust" or "noverlap" as an additional step of the previous selected layout to make sure texts or nodes do not overlap.
- Select "**Fuchterman Reingold**" (it is one of the popular and easy-to-use algorithm in Gephi).

Calculate the Network Characteristics

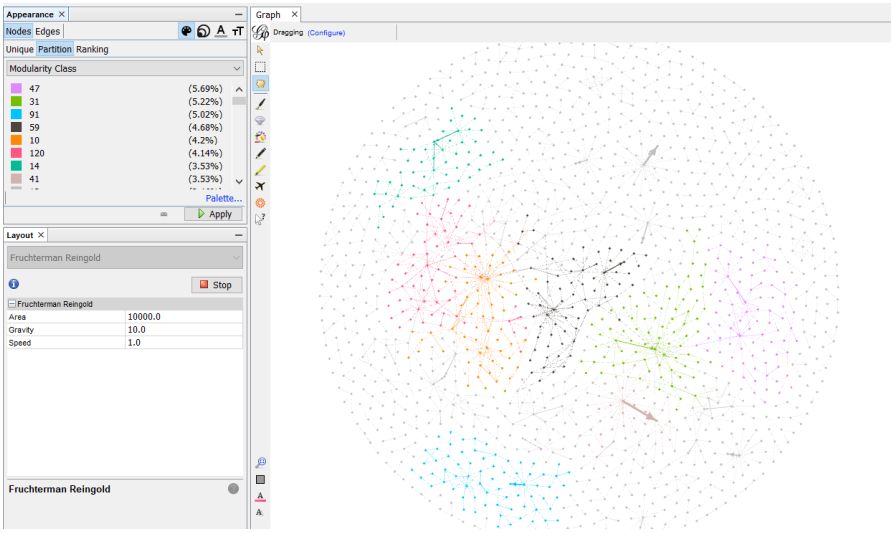
Before we create the actual display of Social Networks, we need to calculate a few key characteristics of this network. There are some important statistical analysis tasks, "**Average Degree**", "**Network Diameter**", "**Graph Density**", and "**Modularity**". You can find the list of statistics on the right side of the window. Click on the "Run" button for each function and get the results.



After conducting these statistical analyses for the social network, we will work on the color and labels.

Assigning Color and Symbols to the Network

- Go to "Appearance" on the left side window → select "Nodes" → "Partition" → choose " Modularity Class" → then click "apply".

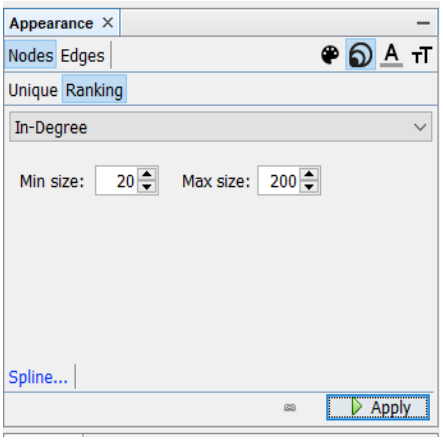


- After clicking, apply you will notice that the color of you graph changed based on the color shown in the figure above.

Modify the size of nodes and add labels

On the "**Appearance**" window. Select "**Nodes**" → click on the **SIZE** icon (the icon with multiple circles).

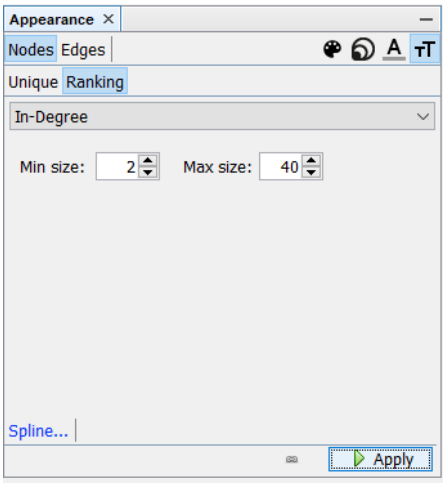
Select "**Ranking**" → Select "**In-Degree**". Type "**20**" for Min size and "**200**" for Max size. Click on "**Apply**".



The next step is to add labels. First of all, click on the "**T**" icon (the dark color one) in the bottom of Graph window to turn on the labels. Then adjust the scale-bar next to the text fonts to reduce the smallest size.

The next step is to adjust the label size using the In-Degree number for each node.

In the "**Appearance**" window, click on "**Nodes**" → select the "**tT**" icon (for Label size). Then select **Ranking** → In-Degree → change the size to Min "**2**" and Max "**40**".



Select a Best Layout

The final step is to select a best layout for this network. Please try several layout and find out the best one you like. You will need to adjust the Node size and Label text size along with different types of layout.

There are several options help defining the shape of your network in different layouts. For example increasing **gravity** bring the nodes closer one from each other. "**Noverlap**" is as its name is indicating preventing any overlap. Click "Run". You will now see the nodes' positions change and evolve. Once your network has reached a shape that satisfies you, just click "Stop". You can also "Drag" and drop some nodes using the "Drag" icon in the Graph window to adjust the layout if you are not satisfied with their exact positions. Pick up one good network layout theme and apply it to your data.

Once you finalize the layout, click on the **Take Screenshot** icon at the bottom of the Graph window (the icon looks like a camera).

Save the screen shot into your working folder (lab06) as "**SNA-name-of-layout-InDegree.png**"

Now you should create another layout screen shot of "OutDegree" network and adjust node sizes and text sizes. Save the screenshot as "**SNA-name-of-layout-OutDegree.png**".

Finally, save your Gephi design results into the working folder as "**yourname.gephi**".

Attached the two screenshots when you submit the lab-7 assignment.

Additional Social Network Analysis Resource:

https://github.com/jdwilson4/Network-Analysis-I

https://seinecle.github.io/gephi-tutorials/generated-html/simple-project-from-a-to-z-en.html

After finishing this Web Course, Please use your own words to answer the following questions (next page): (DO NOT COPY any web resources or Wikipedia texts. We will check your answers with Blackboard tools to verify that your responses are uniquely yours.) By submitting your answers (paper) to Blackboard, you agree: (1) that you are submitting your paper to be used and stored as part of the SafeAssign™ services in accordance with the Blackboard Privacy Policy; (2) that your institution may use your paper in accordance with your institution's policies; and (3) that your use of SafeAssign will be without recourse against Blackboard Inc. and its affiliates.

SafeAssign accepts files in .doc, .docx, .docm, .ppt, .pptx, .odt, .txt, .rtf, .pdf, and .html file formats only. Files of any other format will not be checked through SafeAssign.

LAB-6 Questions:

1. Explain what is a "node"? What is a "Edge" in this exercise example? Please provide a social network example for **at least** FIVE nodes and SIX edges. Draw the network graph in your report. (You can draw on a paper and then take a photo using your phone, or scan it into a digital picture. You can also use any graphic software to draw the network graph.)

2. Display the In-degree network screenshot (one layout) and the Out-degree network screenshot (another layout) in Gephi using different layouts. Explain which layout scheme do you select for the In-degree network and the Out-degree network in this exercise. (Note: DO NOT use the same layout for both networks).

3. Explain the meanings of "**Network Diameter**", "**Graph Density**", and "**Modularity**". What are the values of these parameters in the lab-06 example network? Is this example network similar with random networks, or small world networks, or scale-free networks? WHY?

Please submit your LAB-6 Answers (in a MS Word or a PDF file format only) to the Blackboard System BEFORE the DUE DATE/TIME.