By the end of this activity, you will be able to:

1. Import a CSV file into Gephi
2. Perform statistical operations and layout algorithms on graph data in Gephi

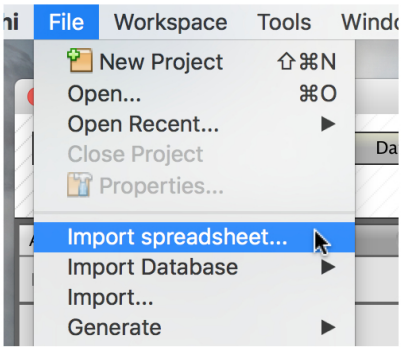
NOTE: Gephi should be run on your native hardware, not in the Cloudera VM. Instructions for downloading, installing, and running Gephi can be found at https://gephi.org/users/install.

Step 1. Download and import CSV file. In your web browser, go to the following link:

https://raw.githubusercontent.com/words-sdsc/coursera/master/big-data-2/graph/diseaseGraph.csv

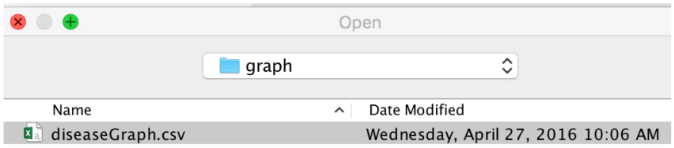
Click on File, and choose Save as to download the file

In Gephi, click on *File*, and choose *Import spreadsheet*:

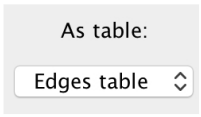


In the Import spreadsheet dialog, click on ... to choose the CSV file:

In the File dialog, choose the diseaseGraph.csv file you downloaded:

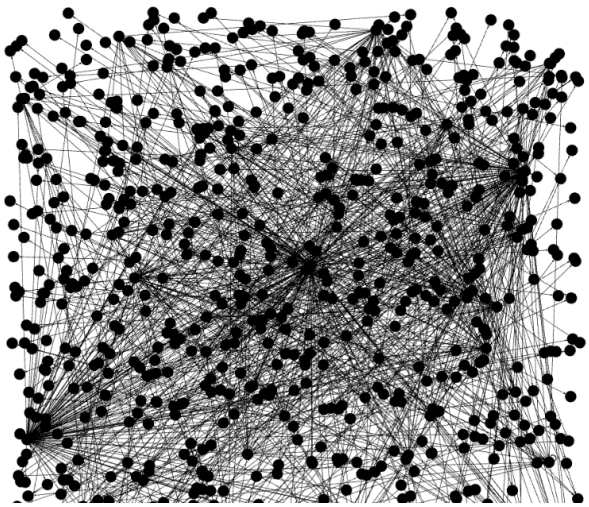


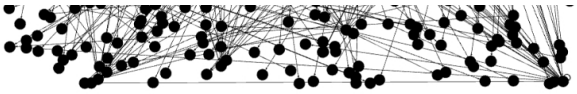
Next, in the dialog, make sure *As table* is set to *Edges table*:



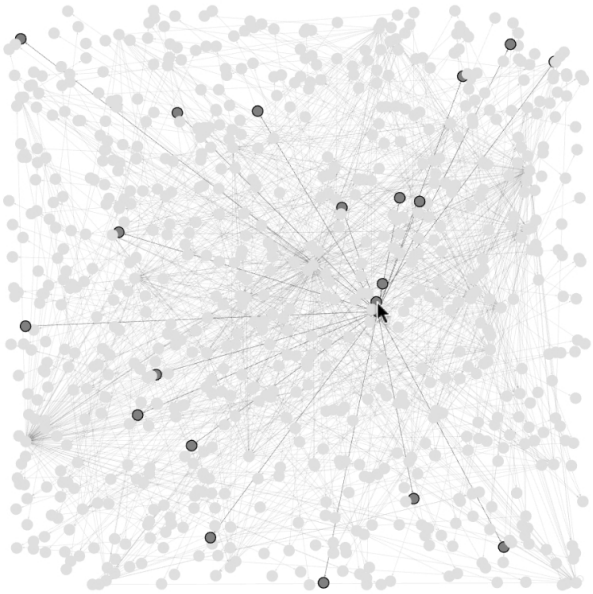
Click *Next*, and then click *Finish* to import the CSV data into Gephi.

Step 2. Examine graph properties. In the middle pane, Gephi displays the graph. The black circles are the nodes, and the lines between them are the edges.

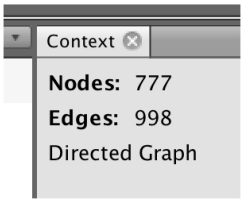




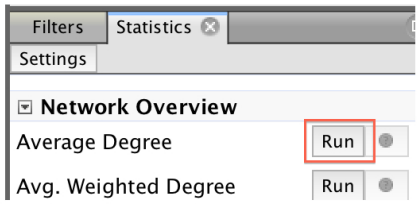
If you place the mouse on a node, then Gephi will highlight the nodes that connected to it:



In the top right is the *Context* pane, which says that the graph has 777 nodes and 998 edges.



Step 3. Perform statistical operations. Below the *Context* pane, is the *Statistics* pane, where you can perform various statistical calculations. We can calculate the average degree by clicking on *Run* next to *Average Degree*.

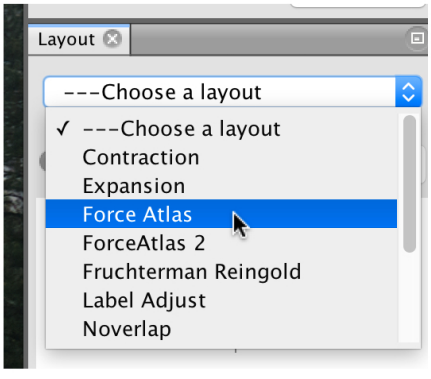


The dialog that pops up says that the average degree is 2.569. Click on *Close* to close the dialog.

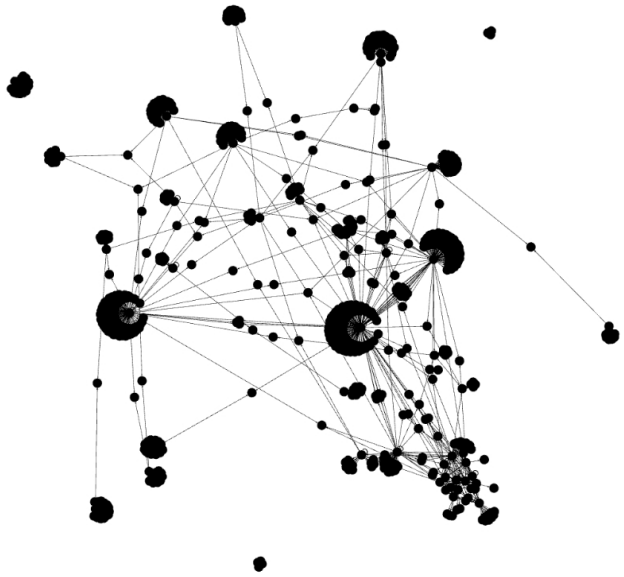
Next, we can calculate the connected components by clicking on *Run* next to *Connected Components*: This will present a dialog box with the title, "Connected Components settings" and ask for either "Directed" or "Undirected" calculations.

The resulting dialog says there are 5 weakly connected components and 761 strongly connected components. Click *Close* to close the dialog.

Step 4. Run layout algorithms. Gephi can perform different layout algorithms on the graph. In the *Layout* pane on the bottom left, click on the *--Choose a layout* combo box and select *Force Atlas*, and click on the *Run* button.

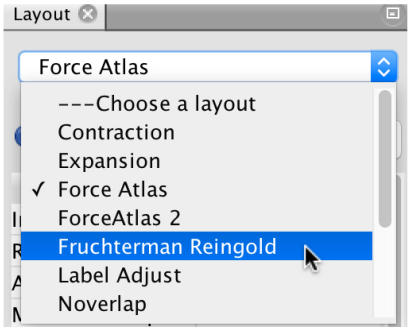


Gephi will change the layout of the graph, and after some time, click on the *Stop* button.



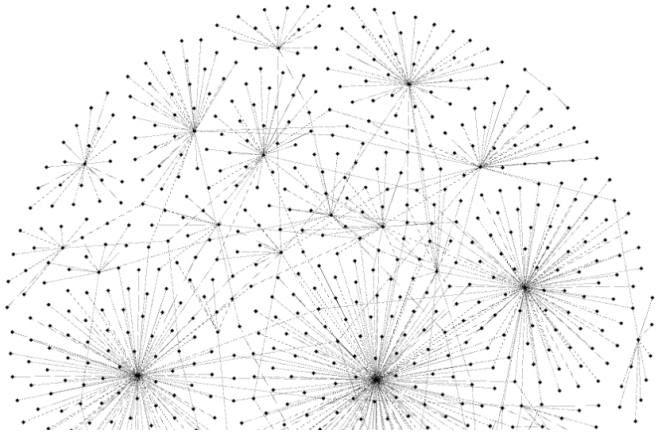
In this layout, strongly connected nodes are clustered together, and we can also see several clusters that are disconnected from the rest of the graph.

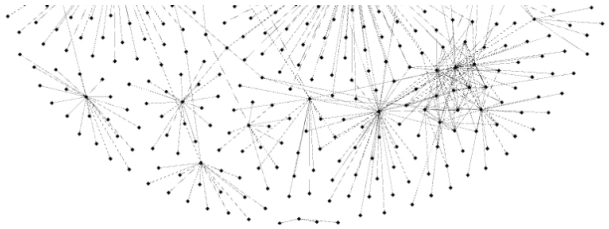
In the *Layout* combo box, select *Fruchterman Reingold* and click on the *Run* button.



The graph layout will change to make the nodes evenly spaced. After the graph stops moving, click on the Stop button, and then on the magnifying glass icon in the middle-left bottom to center the graph.

In this layout, we can better see which nodes have a lot of edges.





Mark as completed