**Exercise- Data Visualization using Gephi**

**Student Name: Student Id:**

**Date:**

# Please use the screenshots ONLY as a reference. The instructions have to be followed AS-IS.

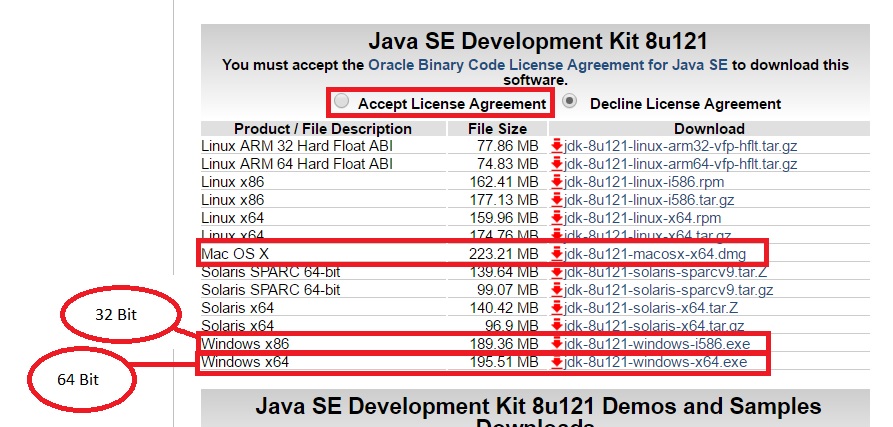
# **Objective**

The objective of this exercise is to develop skills on how to visualize and analyze large networks using Gephi. This exercise focuses on the relationship between websites. Here we would be emphasizing on the association of Apple with other websites and with itself.

**Prerequisites:**

1. Install Java SE Development Kit 8

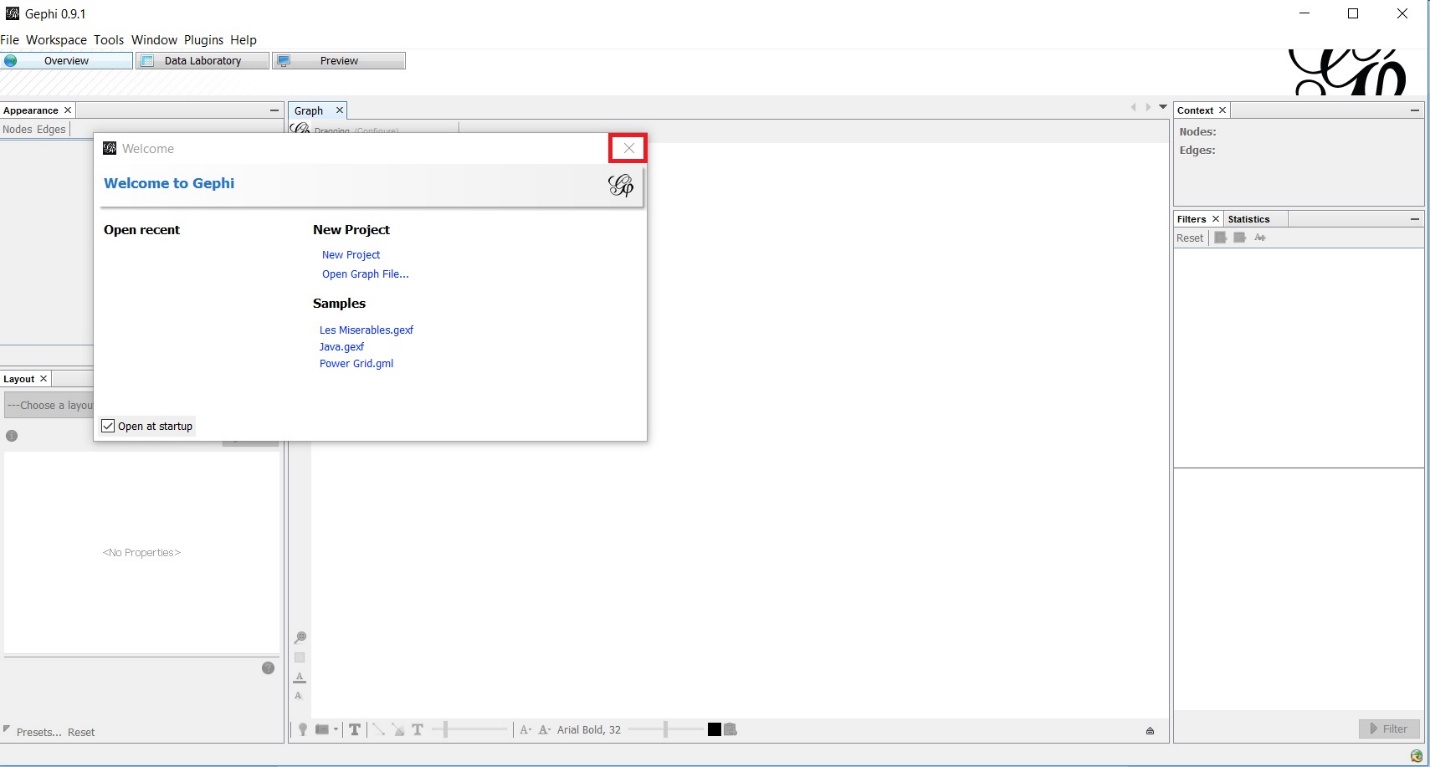
<http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html>



1. Install Gephi

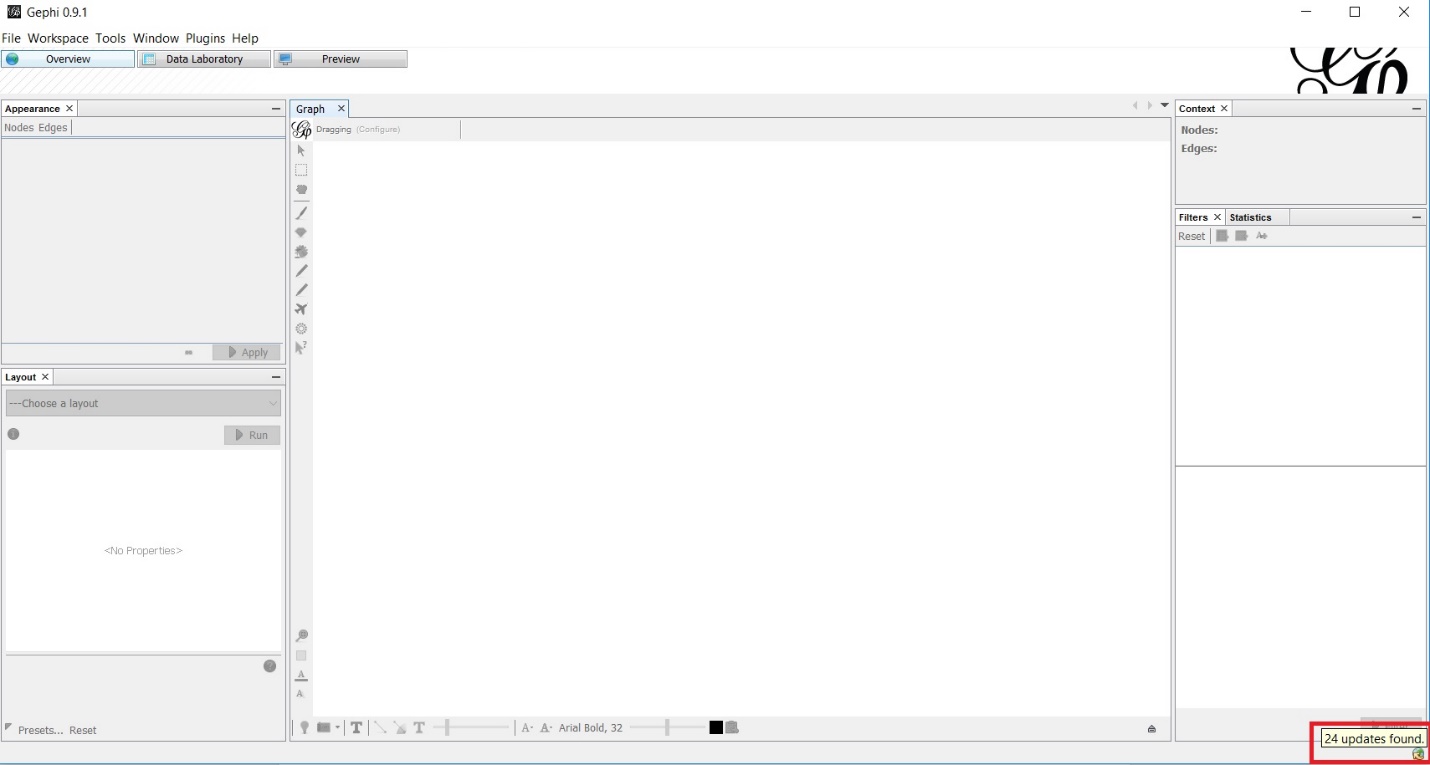
<http://gephi.github.io/users/download/>

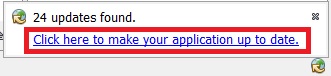
1. In case Gephi does not work after completing steps 1 & 2, follow the steps as below:
   1. Open your Gephi Installation folder (probably C:\Program Files (x86)\Gephi-0.8.2 ) and locate **etc** folder
   2. Within the **etc** folder, you will find a file named **Gephi.conf.** Open the file with notepad
   3. Search for “**#jdkhome="/path/to/jdk**"
   4. Remove **#**, otherwise the code will be considered as a comment and hence will not be executed
   5. Replace the text **/path/to/jdk** inside double quotations with the directory address of your java folder. In order to find the java folder, go to windows drive (Probably C:), go to Program Files and find Java folder. Inside the java folder you will find a folder that starts with **“jdk1.7”**. Copy the path of this folder and paste it instead of **/path/to/jdk** in the **Gephi.config** file.
   6. Save the config file, in case you are allowed to save it then save it somewhere else and then replace it by the original file.
2. Open Gephi. If the Welcome pop-up screen appears then click on cancel button.



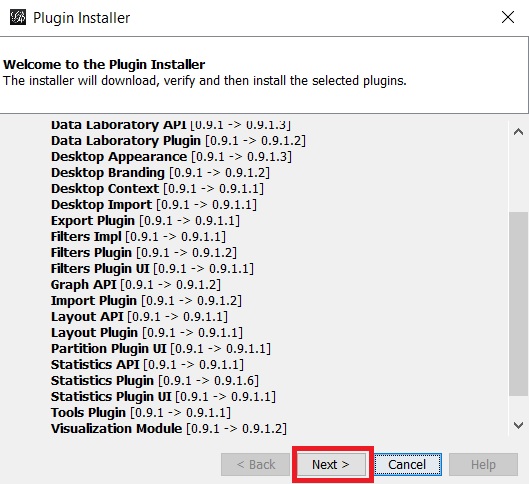
1. On the bottom right of screen there is an icon indicating updates C:\Users\Arron\Desktop\ROM SCREENSHOTS\EXERCISE 10\Assignment Fix\DF6.jpg. Click on the icon and install the updates as shown below:

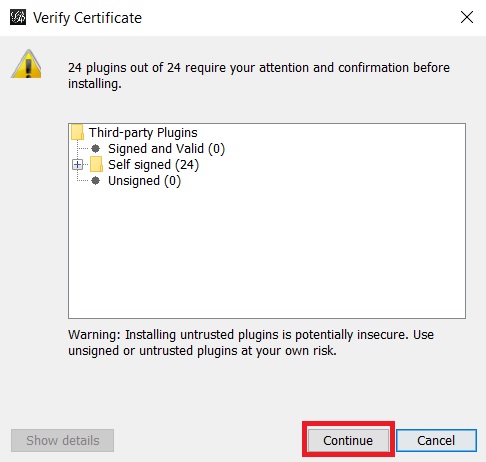
**[Please Note: If you cannot see the updates icon on the bottom right of the screen, follow step 6 to install updates. If you can see this icon, follow step 5 and move to step 7]**



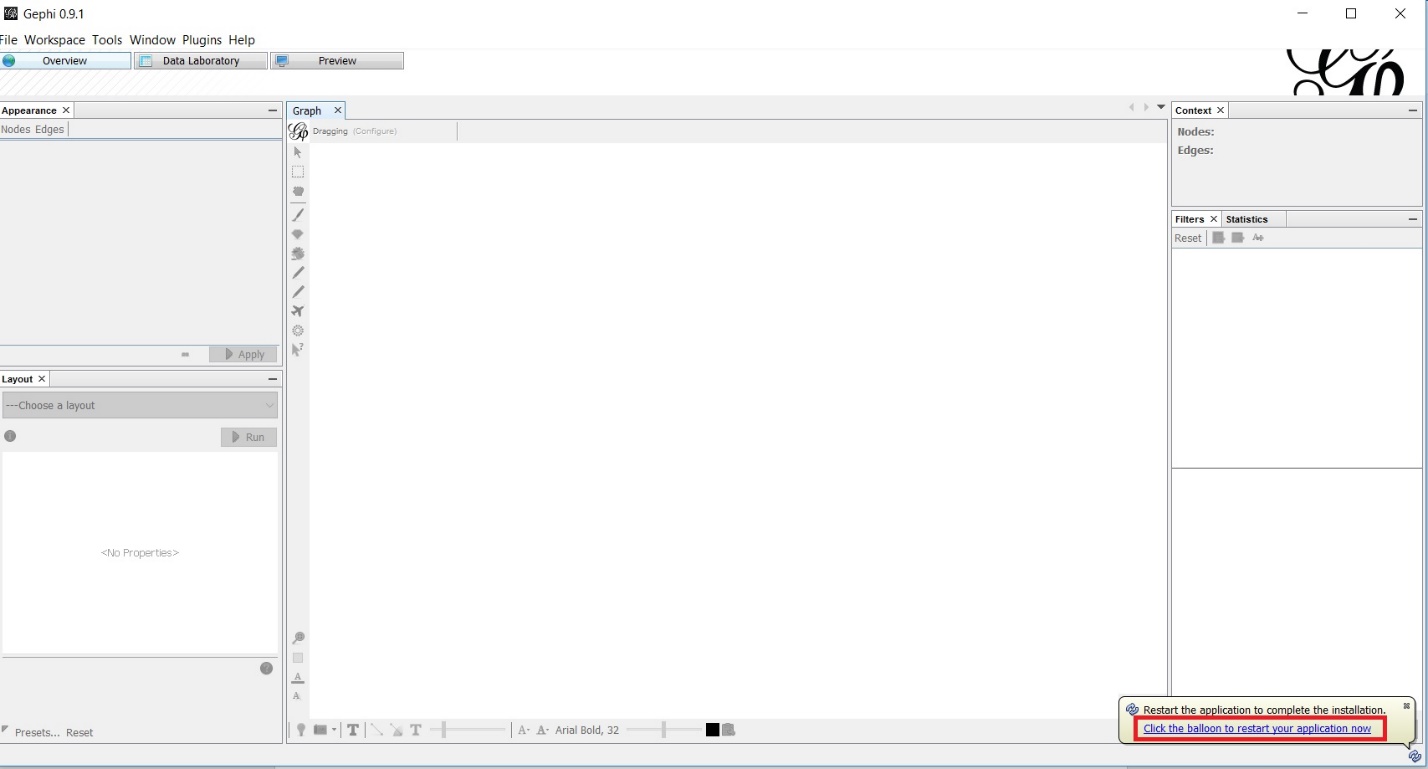


Click on Next in Plugin Installer screen. If ‘Verify Certificate’ screen pop ups. Click on Continue





After installing the plugins, a request for restart appears at the bottom right corner. Click on restart link and restart the applications.

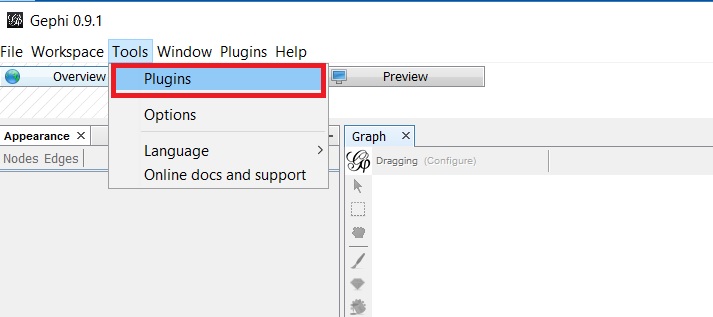


Your plugins are installed. Close the application for now.

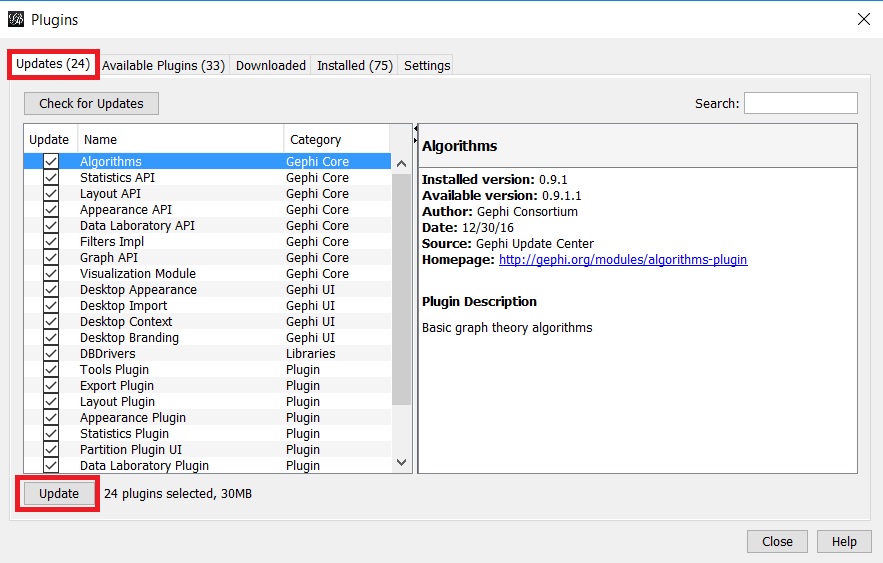
**[Please Note: You need to restart your Gephi in order to successfully install your updates. If you avoid restart you may face errors and will have to install Gephi all over again.]**

1. **[Please follow this step ONLY if you cannot install updates using the step 5 or else go to step 7]**

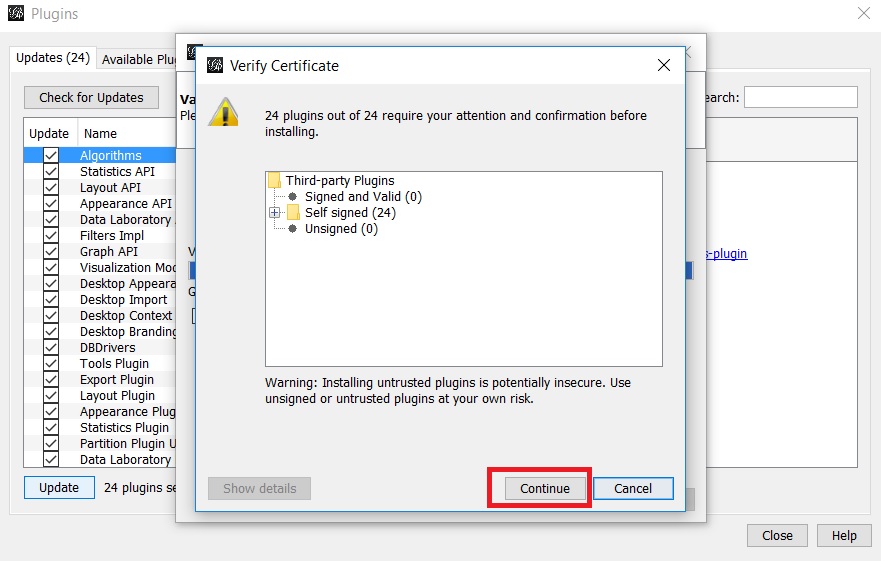
Go to Tools -> Plugins



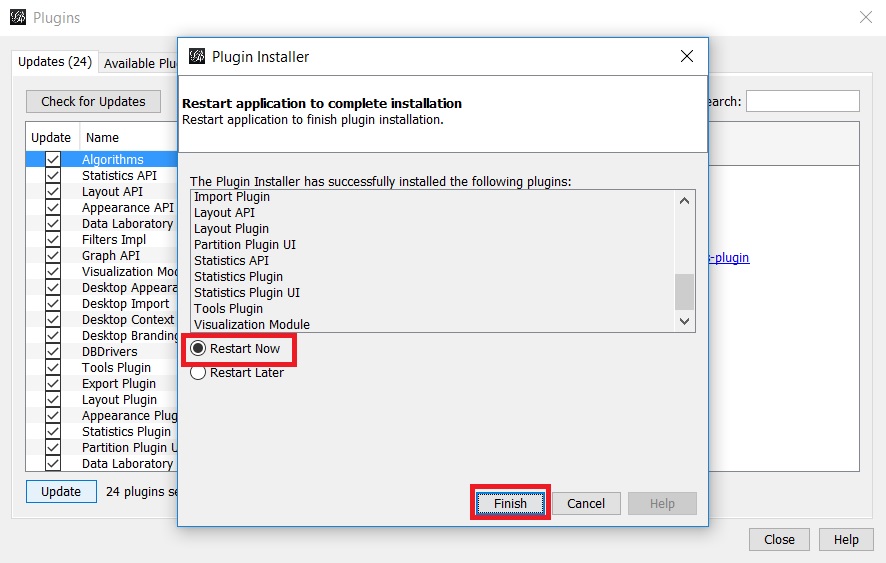
In the Plugins window, select Update Tab and make sure all the boxes are checked in that tab. Then click on Update Button



Click next and wait for installation to complete. If the Verify Certificate pop up appears, click on Continue.



Select Restart Now and click on Finish.

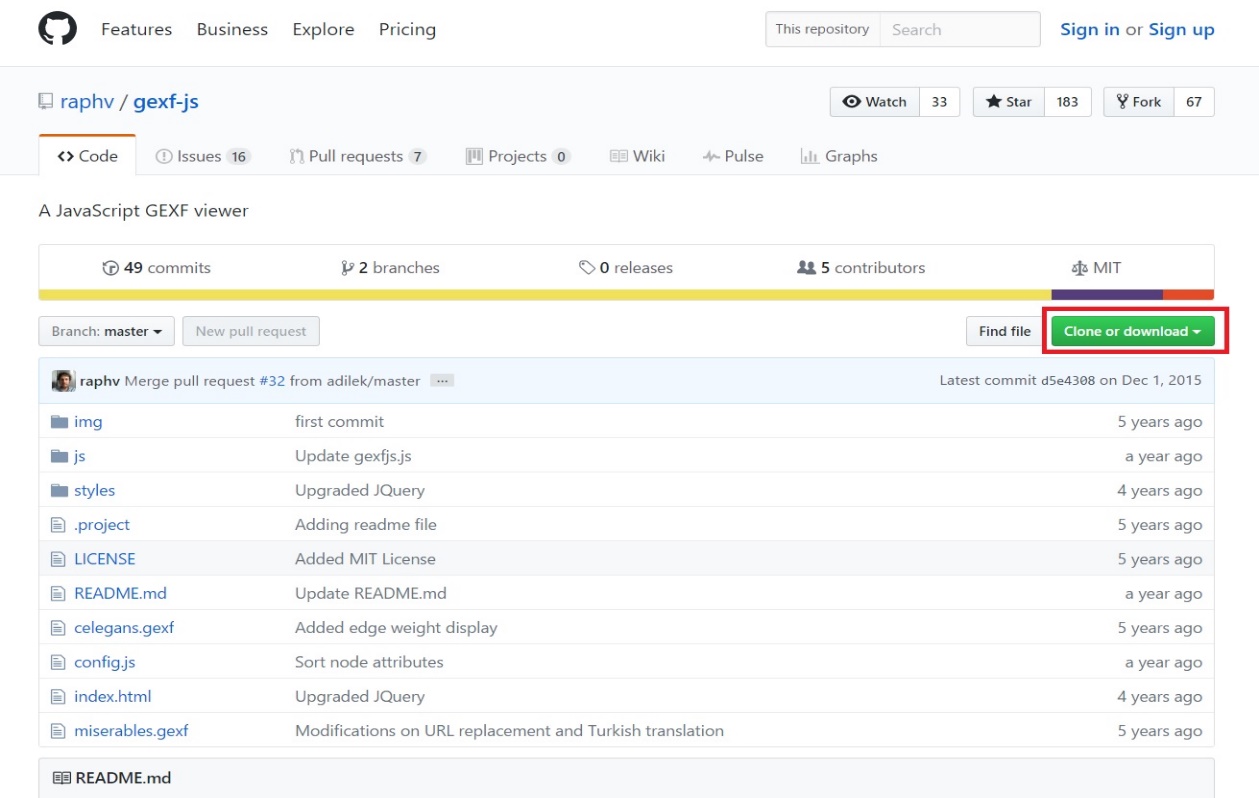
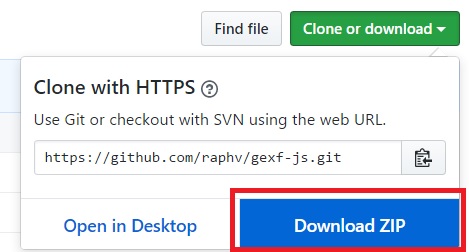


Your Updates are now installed. Close the Gephi for now.

**[Please Note: You need to restart your Gephi in order to successfully install your updates. If you avoid restart you may face errors and will have to install Gephi all over again.]**

1. Download Javascript Gexf viewer, to view the Gephi file in the web browser. After downloading the file make sure that you place the extracted folder in desktop or in any other know location and later on your exported Gephi file needs to be saved in this folder.

<https://github.com/raphv/gexf-js>

****

**Step 1: Download the provided CSV files from eLearning**

**About Gephi**

**Concept:**

Gephi is an open source software for network visualization. It can read many file formats including Gephi, GEXF, GML, GDF, CSV and many others. As part of this exercise we would be using csv files.

Before starting with Gephi, we should familiarize ourselves with 2 terms, one being the **Node** & the other **Edge**.

**Node:** A node is a unique identifier of an object within a data set.

**Edge:** An edge is a line that connects two nodes

In order to import data from excel or csv file into Gephi, you will usually need to prepare following 2 files:

1. **Node File** – Containing the nodes and its attributes
   1. The node file must include a column having name **‘ID’**
   2. ID column should contain unique entries
2. **Edge File** – Containing the edges and its attributes
   1. The edge file must include columnswith name **‘Source’** & **‘Target’** which contains the start and the end nodes for each edge

**Step 2: Manipulate the downloaded files from eLearning**

In this step we will make small additions to the node and edges files to understand how the nodes are created and how a relationship is established between 2 nodes. We will also be defining the strength of the relationship between 2 nodes which can be visualized by the thickness of the edges (links) between the nodes.

**Step 2.1: Create nodes and their relationships**

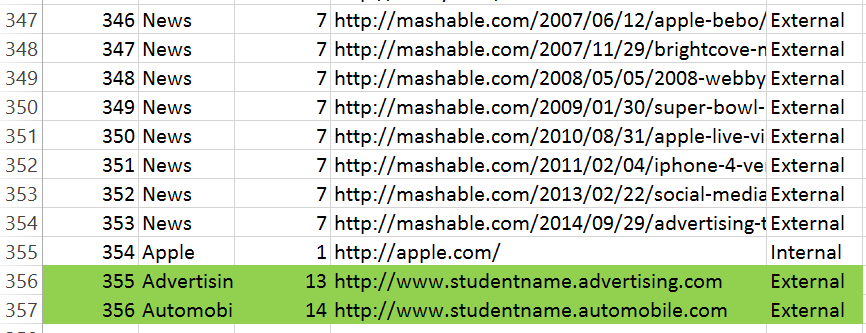
Open the **nodes table** and here you will be maintaining an entry at the end of your nodes file and make sure that you assign a unique node id to the new entry.

**Note:** It is better to follow the sequence for the node ids as in the node file.

Maintain the following entry in your node file:-

1. **Id no.**: Last + 1 record in your node file
2. **Industry**: Advertising
3. **Industrywise Weight:** 13
4. **Description:** http://www.studentname.advertising.com
5. **Origin:** External
6. Repeat steps **1 through 5** for Automobile industry, ensure that the node id for automobile industry is different from the one for advertising industry (see screenshot below)

After adding the records, your data should look as below.



**Save the nodes file.**

Open the **edges table** and here you will be defining the relationship between the advertising and automobile websites with Apple website. Ensure that you map the same node ids maintained in the nodes file.

Scroll till the end of the edges file and then maintain the following:-

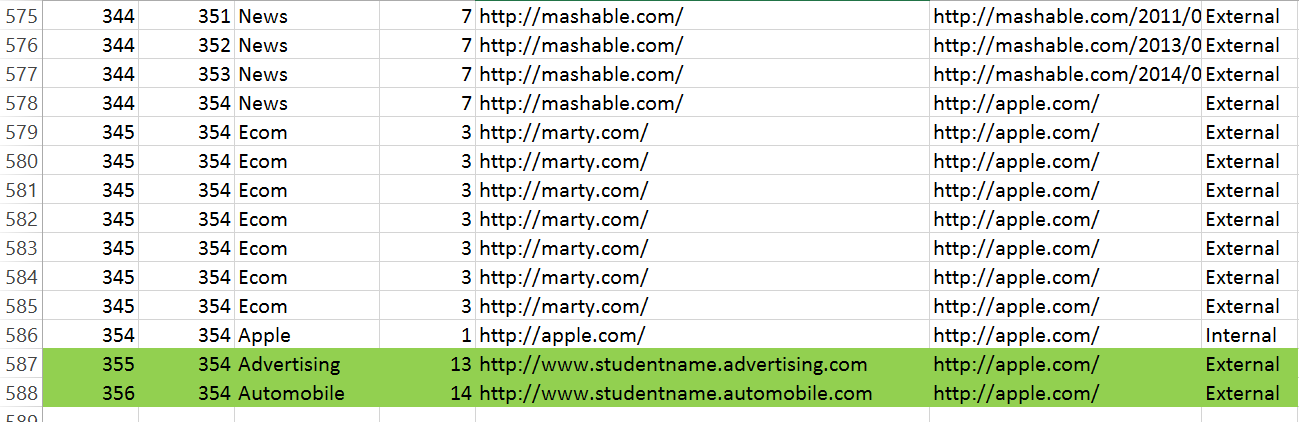
Maintain entry in the edges table for **advertising industry**:

|  |  |
| --- | --- |
| 1. **Source Column :** | Node Id for **advertising** industry which we have maintained in the node file. |
| 2. **Target Column :** | Node Id for Apple. Open the node file, filter by Apple and then copy its node id place in the target column against source column of advertising industry |
| 1. **Industry Column:** | Maintain industry as **Advertising** |
| 1. **Source URL :** | <http://www.studentname.advertising.com> |
| 5. **Target URL :** | http://apple.com |
| 6. **Origin :** | External |

Maintain entry in the edges table for **automobile industry:**

|  |  |
| --- | --- |
| 1. **Source Column :** | Node Id for **automobile** industry which we have maintained in the node file. |
| 2. **Target Column :** | Node Id for Apple. Open the node file, filter by Apple and then copy its node id place in the target column against source column of automobile industry |
| 1. **Industry Column:** | Maintain industry as **Automobile** |
| 1. **Source URL :** | <http://www.studentname.automobile.com> |
| 5. **Target URL :** | http://apple.com |
| 6. **Origin :** | External |

**After adding the 2 entries, your records should look as below**



**Question:**

1. **Paste the screen shot of your edges file after adding the 2 records as highlighted above.**

**Answer:**

**Step 2.2: Create relationship strength between 2 nodes.**

Select the record with industry as **Advertising** in the edges file and copy the record. Insert 50 duplicate records having industry as **Advertising**.



Select the record with industry **Automobile** in the edges file and copy the record. Insert 30 duplicate records having the industry as **Automobile**.

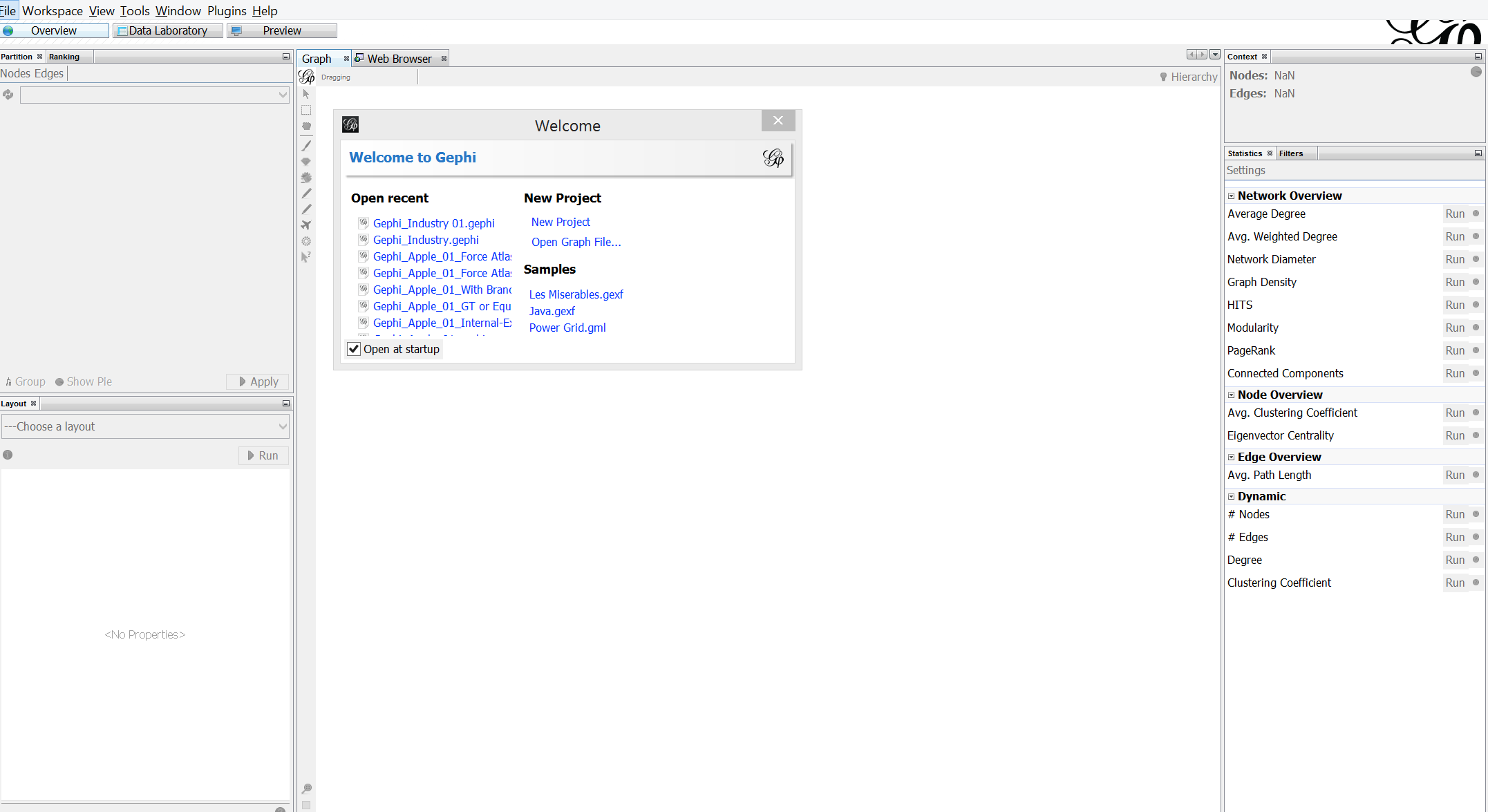


Save the edges file.

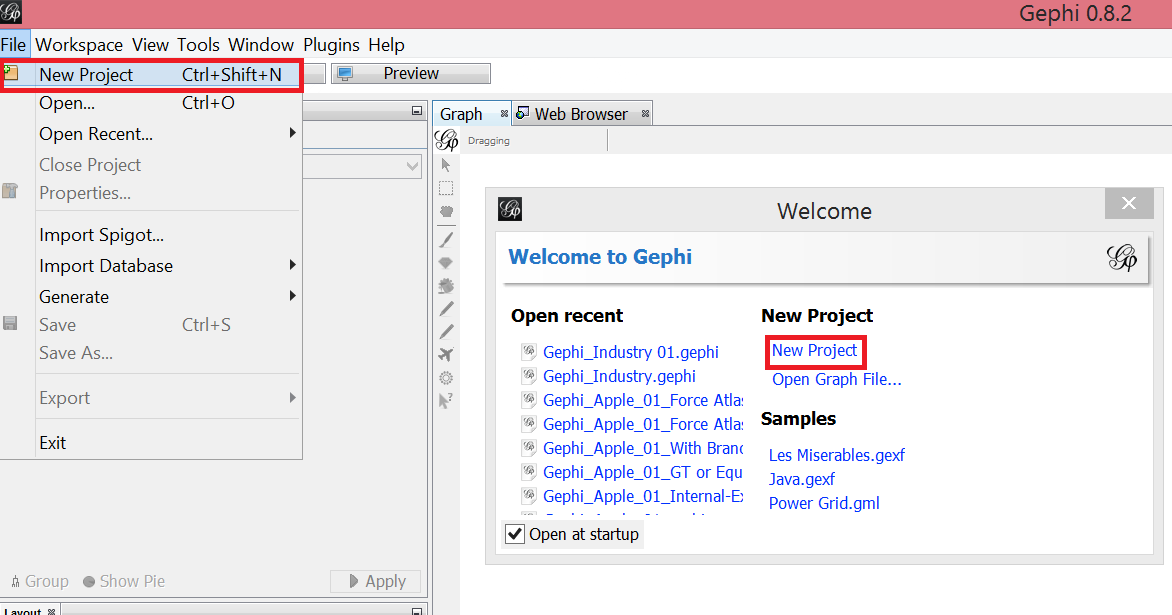
Once you import the data into Gephi, you will notice that link between the nodes advertising & automobile industry with apple is thick and which means that the relationship between the advertising & apple as well as automobile & apple is strong.

**Step 3: Import data into Gephi**

Once Gephi is installed properly, the following screen should appear:



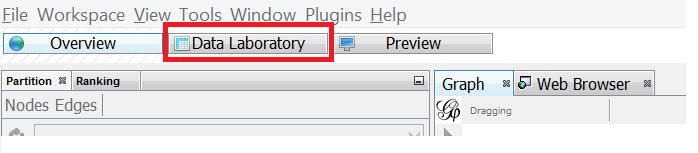
Click on **New Project** in the welcome screen as highlighted in the below screen shot or follow the menu path **File -> New project**



There are 3 major sections within Gephi:

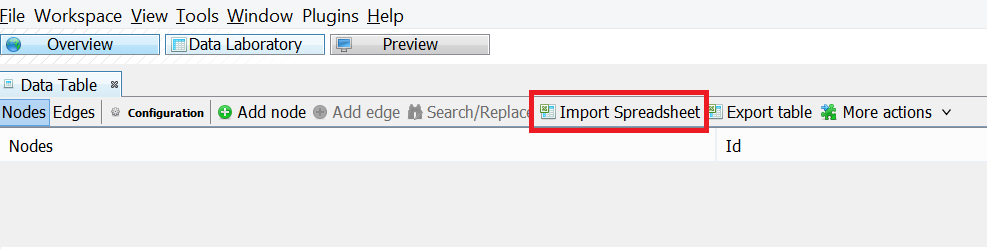
1. **Overview Tab** : Setup the network visualization
2. **Data Laboratory Tab** : Import, Examine & Edit network data ( Node & Edges table)
3. **Preview Tab** : Configure the rendering settings, for instance color, label sizes etc and also preview the visualization.

In order to upload the data, click on the data laboratory tab as highlighted below

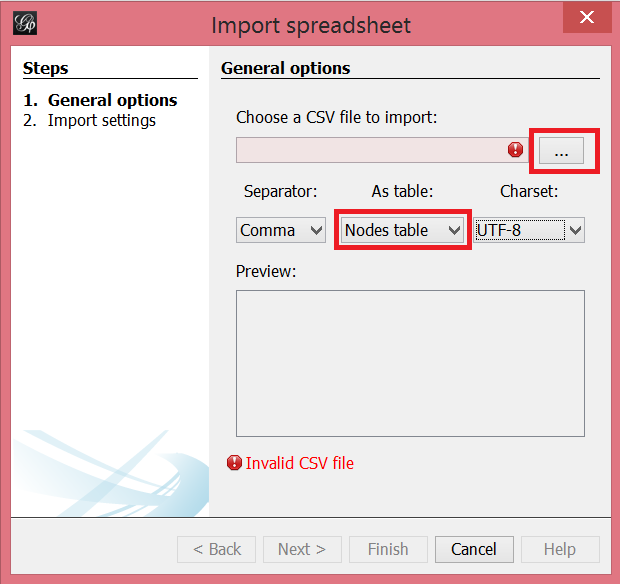


**Import Node file:**

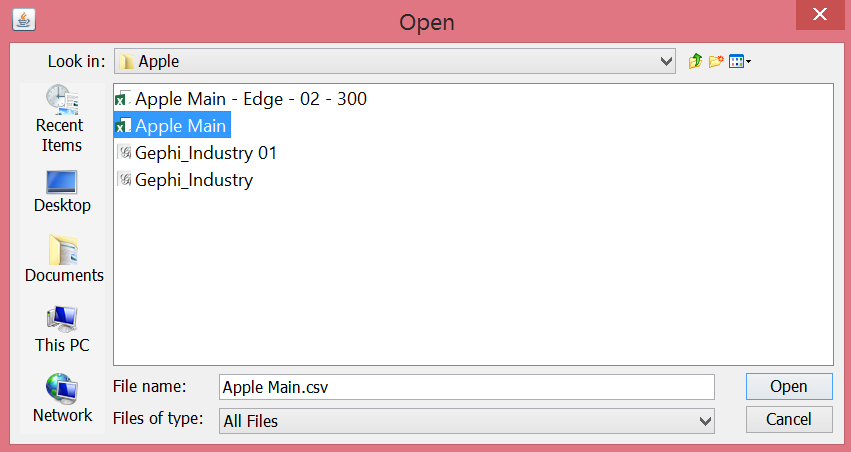
Click on Import Spreadsheet under Data Laboratory tab



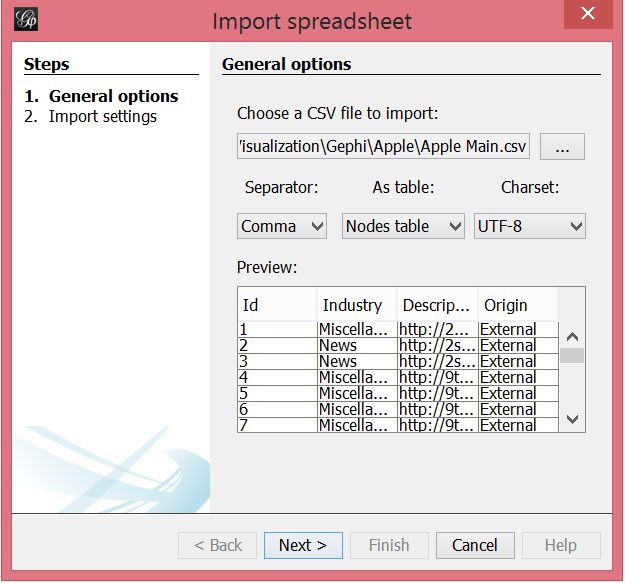
A popup will appear, click on the icon below the **“Choose a csv file to import”** to import the data also ensure that under **“As Table:”** option **nodes table** is selected from the dropdown as highlighted in the below screenshot.



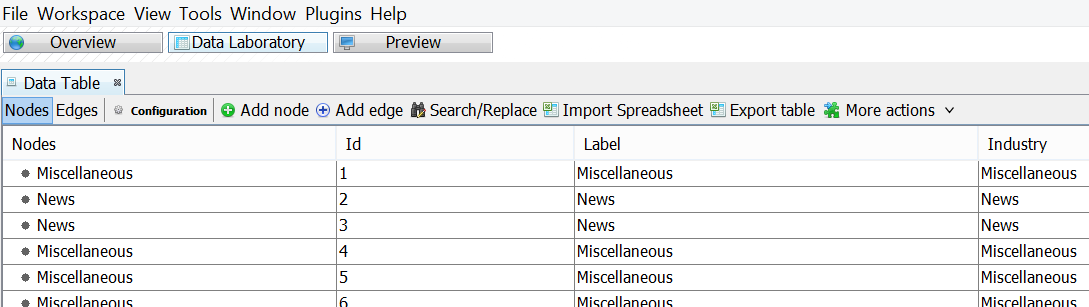
When you click on icon , a popup appears, select your node file downloaded from eLearning and then click on **Open.**



If case of no errors, the next button will be enabled, click next.



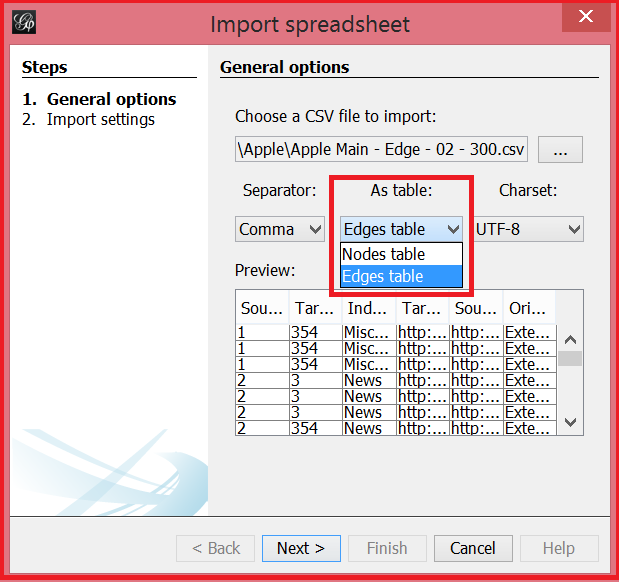
Then click finish. You will be able to view the contents of the node file by clicking on the **nodes** tab under **Data Laboratory** tab.



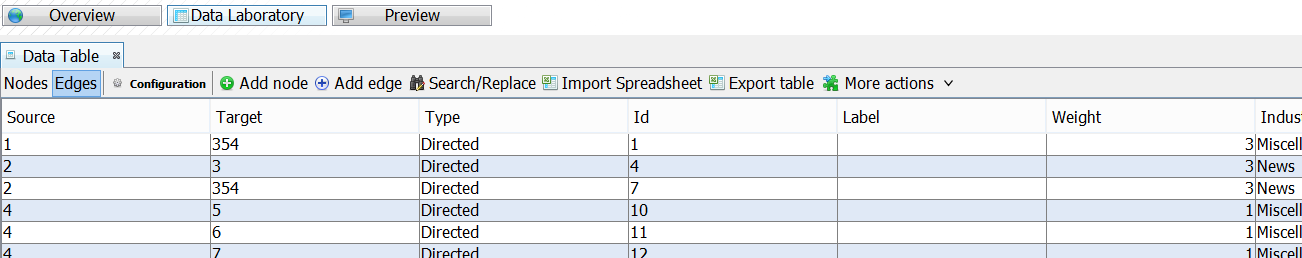
Import Edges Table:

Follow the same steps are followed while uploading the nodes file.

* + - 1. Click on **Import spreadsheet** icon under Data Laboratory tab.
      2. Click on icon , to import the downloaded edge file
      3. Select Edge Table from dropdown, under **“Table as”**



* + - 1. Click on **next** and then click **Finish**.
      2. View the contents of the **edges** file by clicking on edges tab, under data laboratory.

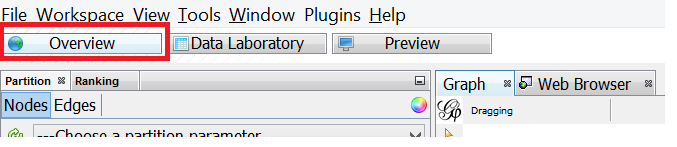


**Step 4: Prepare Visualization**

As part of this exercise, we will be creating three different visualizations using the imported data set.

1. Clustering based on websites
2. Clustering based on Industry
3. Clustering based on Apple’s Internal & External relationships with different websites.

In order to create visualizations click on the overview tab.



Notice that Gephi has already provided us with a default visualization based on our data set once you click on the **overview tab** as highlighted below

The dark thick links which you notice depicts the websites having stronger relationship with Apple website.

**Question:**

1. **Paste the screen shot of visualization in the overview tab.**

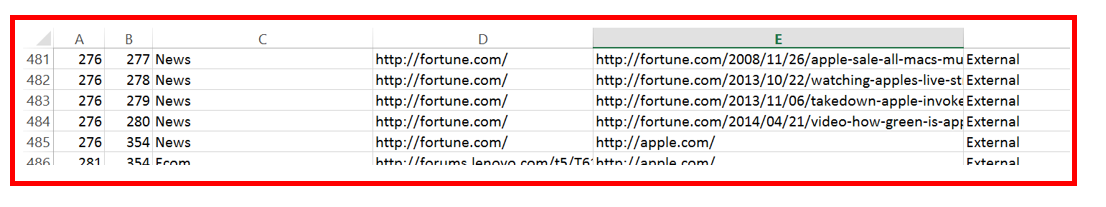
**Answer:**

**Step 4.1: Clustering based on websites**

In the data set we have grouped the web pages by website names i.e. the web pages belonging to the same website are grouped together.

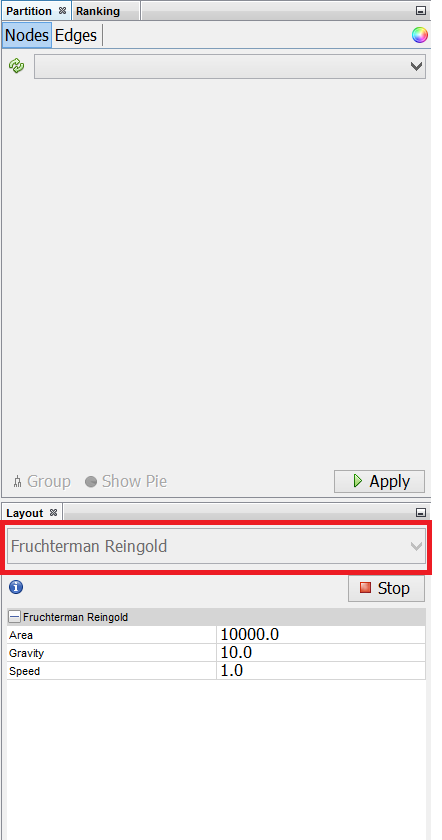
For instance, all different fortune website page links are grouped under http://fortune.com and which is then linked to apple.com. With this instead to having 1:1 linkage between apple and any other web page we are trying to create a pattern.

Below screen gives a snap shot of the grouping done based on website.



**Step 4.1.1 Prepare Visualization**

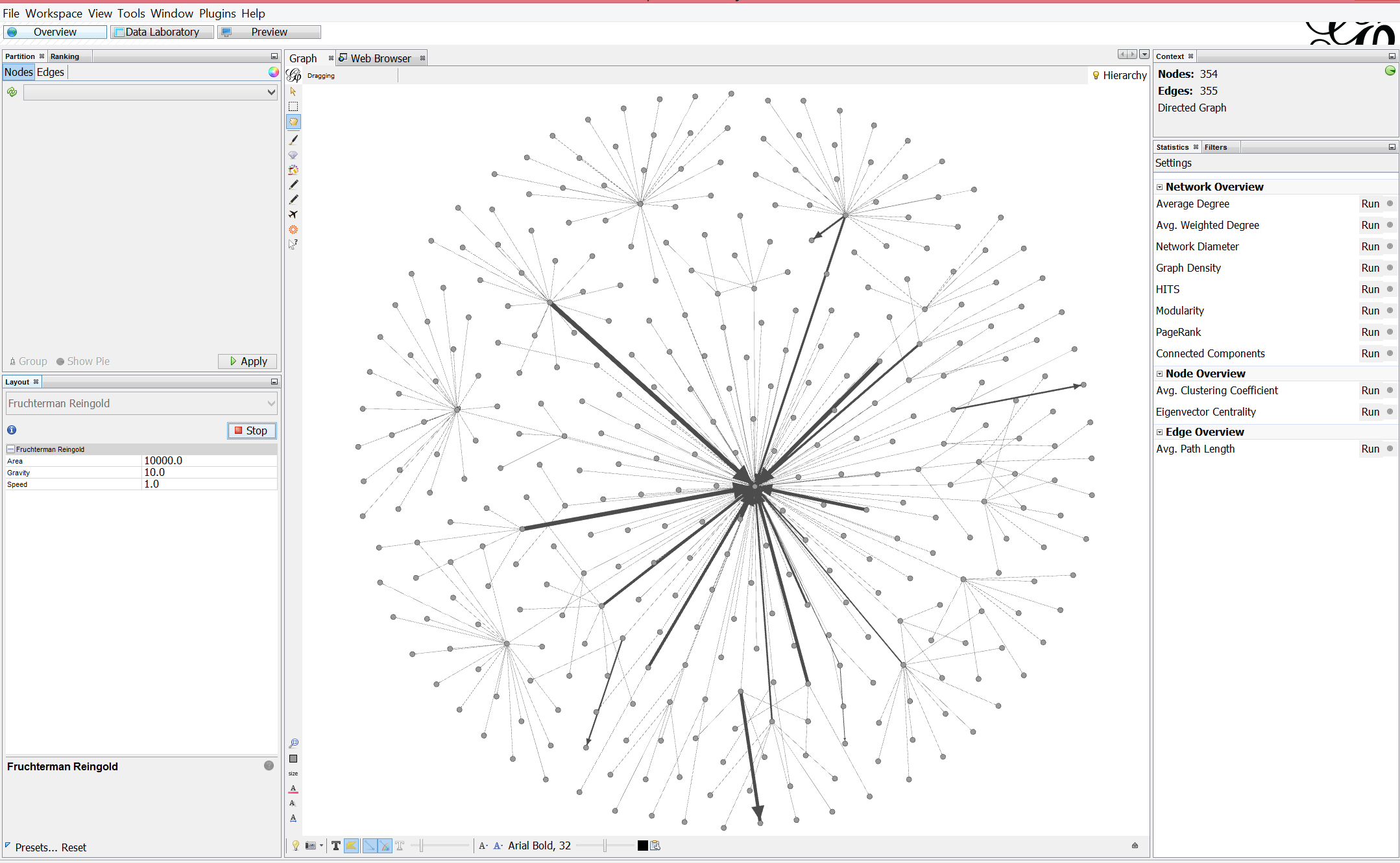
Choose algorithm **Fruchterman Reingold** under layout section on the left panel. And then click on Run. Let it run for a 1 minute and then once the visualization is expanded and clearly visible, stop the algorithm by clicking on the stop button.



Notice that the visualization is changed now. Clustering has now become more prominent and few of the websites have stronger relationship than the other.

You can notice the strength of the relationship by the thickness of the linkage between the nodes.

Your output should look as below:

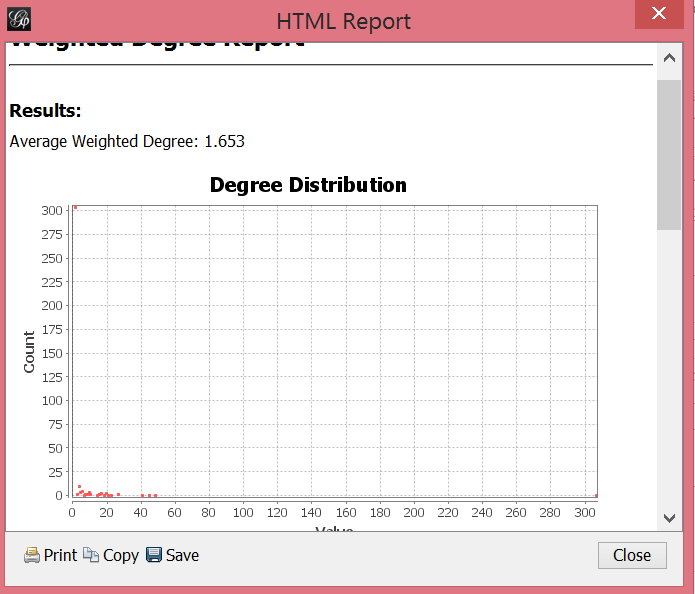


Now we will assign weights to different nodes using Avg. Weighted Degree parameter present in the right panel under network overview within statistics tab.

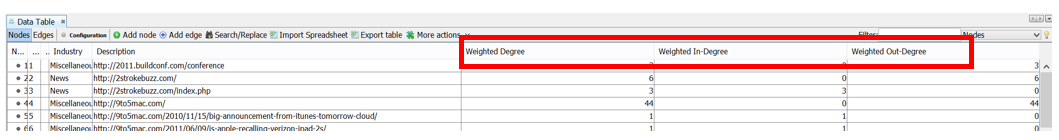
**Note**: In case the statistic tab is not visible right side of the window then go to context menu **window -> statistic**. After this step the statistic should be visible on the right side panel.

Click on **Run** beside the **Avg Weighted Degree** parameter. This will assign weights to the node id’s based on the no. of edges or relationships present irrespective of the type of entry i.e. whether it is duplicate or distinct. It will count all the entries against that node id in the edges table.

Once you click on run, you will get the following popup just click on close.

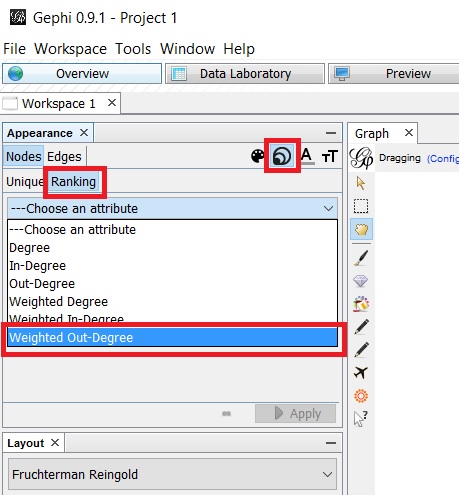


After this step click on the **nodes tab** under **data laboratory tab**. Notice that 3 more columns have been added to the nodes table with different weights assigned to different records as highlighted below.

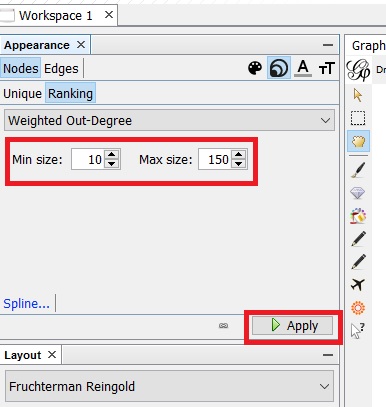


Go back to the visualization section by clicking on the overview tab. Now we will be applying these weights to different nodes.

Click on the **nodes** tab under **Appearance** on the left panel and select **Size Icon** in Nodes tab, Select **Rankings** under node tab, click on the drop down and select **Weighted out Degree** as highlighted below.



Set min size and max size as below and then click on **apply**



After this step you will notice changes in your node sizes.

**Question:**

1. **Place the screen shot of your visualization section below.**

**Answer:**

1. **Which website has got the maximum node size?**

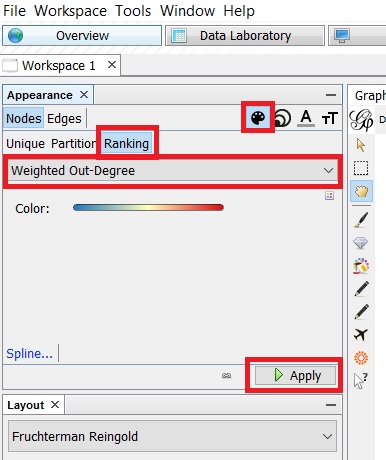
**Hint**: Find the node with the biggest circle, right click on it and then click on select in data laboratory. Go the data laboratory and mention the selected website name in the answer.

**Answer:**

**Coloring the nodes by weight.**

Select nodes tab under Appearance pane. Select color icon as highlighted below, select rankings tab under node tab and then click on apply.

**Note:** You can also change the color coding by right clicking on the area under color section.

****

After this step you will notice change in the color of the nodes.

**Question:**

1. **Paste the screen shot of your visualization/graph section.**

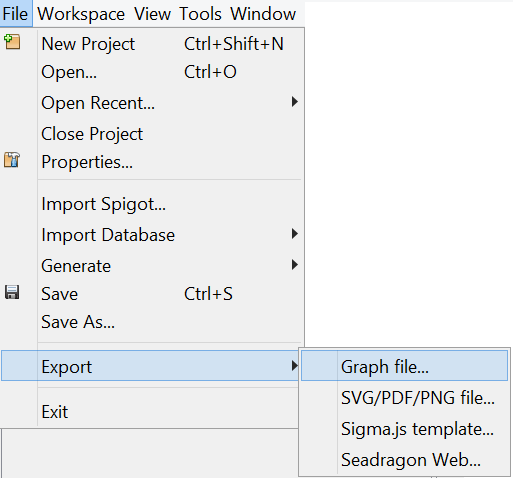
**Answer:**

**Step 4.1.2 Save Gephi file**

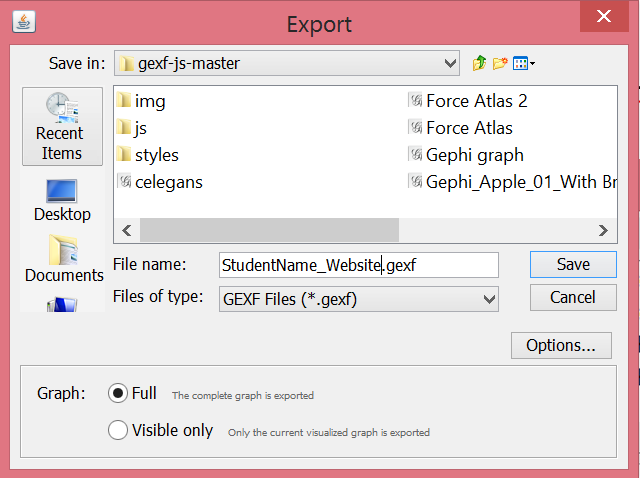
Go to context menu **File -> Save** and then save the file as **StudentName\_Website.Gephi**

Also **export** the file as a Graph file ( \*.**GEXF** file), in order to view it in the web browser.

Follow the below context menu path.

****

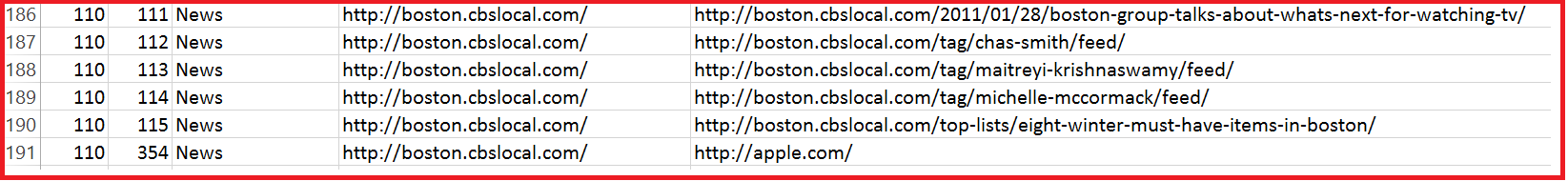
A popup window appears, maintain the filename as **StudentName\_Website** and select file type as .**GEXF file** and then click **Save.**



**Step 5.2: Clustering based on Industry**

In the data set we have grouped the web pages by Industry i.e. the web pages linked to a news website is assigned **Industry** as **News** whereas a web page linked to an educational website is assigned **Industry** as **Education**.

For instance, all different website pages in the screen shot are of a news website http://boston.cbslocal.com and which is then linked to apple.com. With this type of clustering we are trying to view the data from a different perspective and trying to analyze the pattern and see which industry is most strongly linked with Apple.

****

**Step 5.2.1 Import data into Gephi**

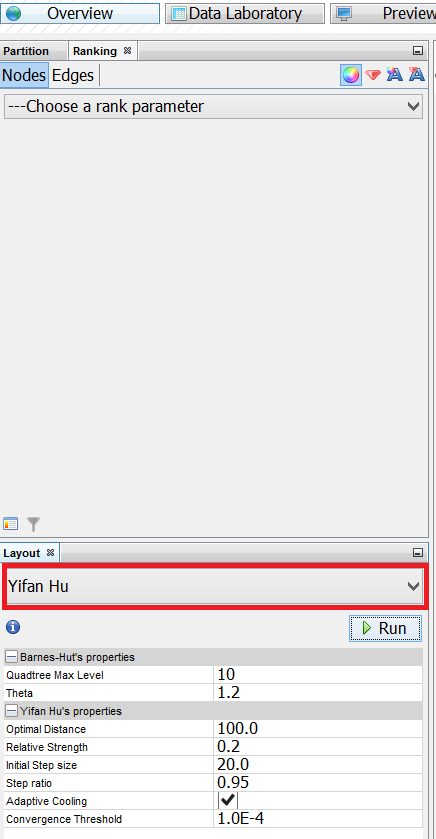
Create new project and follow exactly the same steps as defined in Step 3.

**Note:** You need to import the same files as imported in step 3.

**Step 5.2.2 Prepare Visualization**

Choose algorithm **Yifan Hu** under layout section on the left panel. And then click on Run. Let it run for a minute and then once the visualization is expanded and clearly visible, stop the algorithm by clicking on the stop button.

**Note:** In case the algorithm stops automatically then, keep clicking **Run** until it expands to a level where it has completely covered the graph section of your window.

****

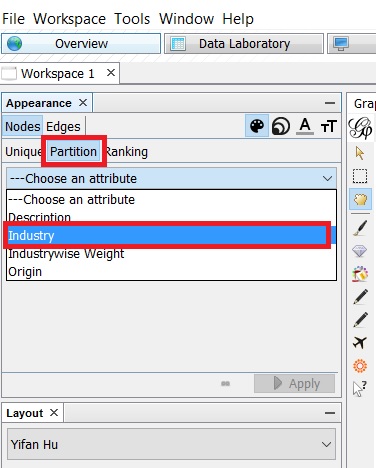
**Question:**

1. **Paste the screen shot of your Graph/Visualization section.**

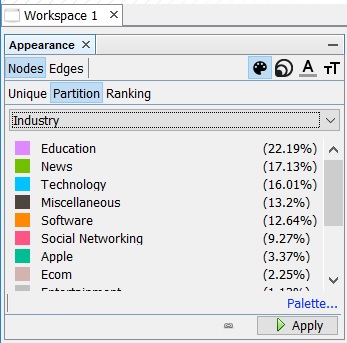
**Answer:**

**Segregate the Industries by color.**

Select **Partition** section under **Nodes** tab on the left appearance panel. Select **Industry** from the drop down and then click on **apply** button.

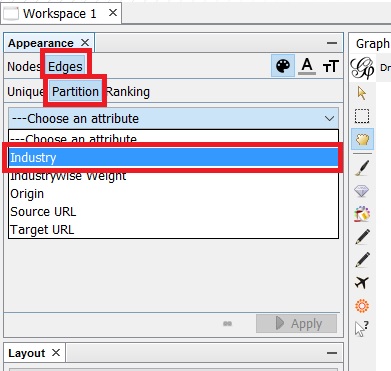


**Note**: You can change the color of a specific node by right clicking on the color next to the industry.

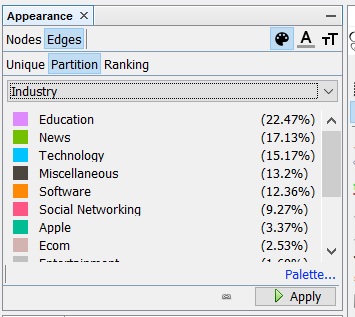


**Assign distinct color to the Edges connecting the nodes**

Select **Partition** section under **Edges** tab on the left appearance panel. Select **Industry** from the drop down and then click on **apply** button.



**Note**: You can change the color of a specific node by right clicking on the color next to the industry.

****

After this step you will notice change in the color of the nodes.

**Question:**

1. **Paste the screen shot of your Graph or Visualization Section.**

**Answer:**

1. **Which website has got the strongest connection with Apple website? Also name its corresponding industry.**

**Answer:**

Hint: Select the 2 edges, you think are the thickest among other edges

**Step 5.2.3 Save Gephi file**

Follow the same steps as highlighted in **Step 5.1.2.**

**Note:** Ensure that the file name for both the type of files is **StudentName\_Industry**

**Step 5.3: Clustering based on Internal or External Linkage**

In the data set we have grouped the web pages by **Origin** i.e. the web pages linked internal to the apple website i.e. web pages that are a part of the apple website have their origin defined as Internal. And Web pages linking to the Apple website from outside have their origin as External.

For instance, all different website pages in the screen shot below are of a news website <http://boston.cbslocal.com>. They are externally linked to the Apple website and hence **origin** is defined as **external**.

****

However, the web pages in the below screen shot are from within the apple website and hence assigned **origin** as **Internal**.

****

**Step 5.3.1 Import data into Gephi**

Create a new project and follow exactly the same steps as defined in Step 3.

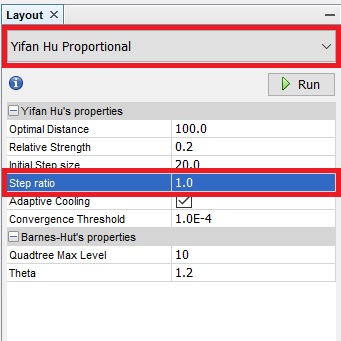
**Note:** You need to import the same files as imported in step 3.

**Step 5.3.2 Prepare Visualization**

Choose algorithm **Yifan Hu Proportional** under layout section on the left panel. Maintain **Step Ratio as** **1** as highlighted in the below screen shot in Red.

Then click on **Run**. Let it run for a minute. While it is running notice how different levels are exploding. Once the visualization is expanded and clearly visible, stop the algorithm by clicking on the stop button.

**Note:** In case the algorithm stops automatically then, keep clicking **Run** until it expands to a level where it has completely covered the graph section of your window.

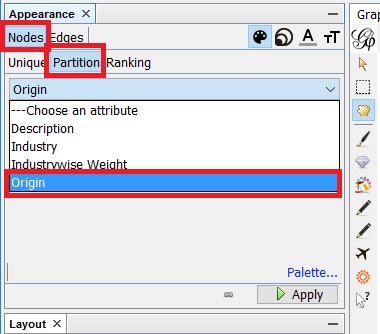
****

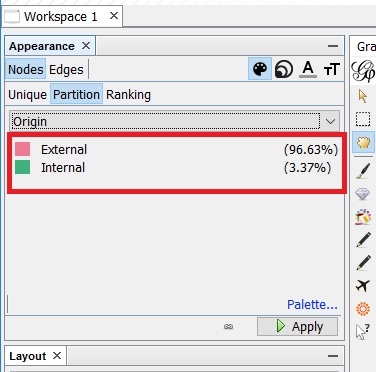
**Assign weight to different nodes id’s**

Click **Run** placed beside the **Average Weight Degree** option under **Network Overview** section of the **Statistics** tab in the right panel. A popup will appear, close the popup by clicking on the close button.

**Assign color to the nodes based on Origin**

Under Partition section in **Nodes** tab, choose **Origin** from the dropdown box and then click on Apply.





After this you will notice change in the node colors, clearly distinguishing the internal and external links.

**Question:**

1. **Paste the screen shot of your Graph or Visualization section.**

**Answer:**

1. **How many internal nodes can you see from the graph excluding the central node?**

**Answer:**

**Step 5.3.3 Save Gephi file**

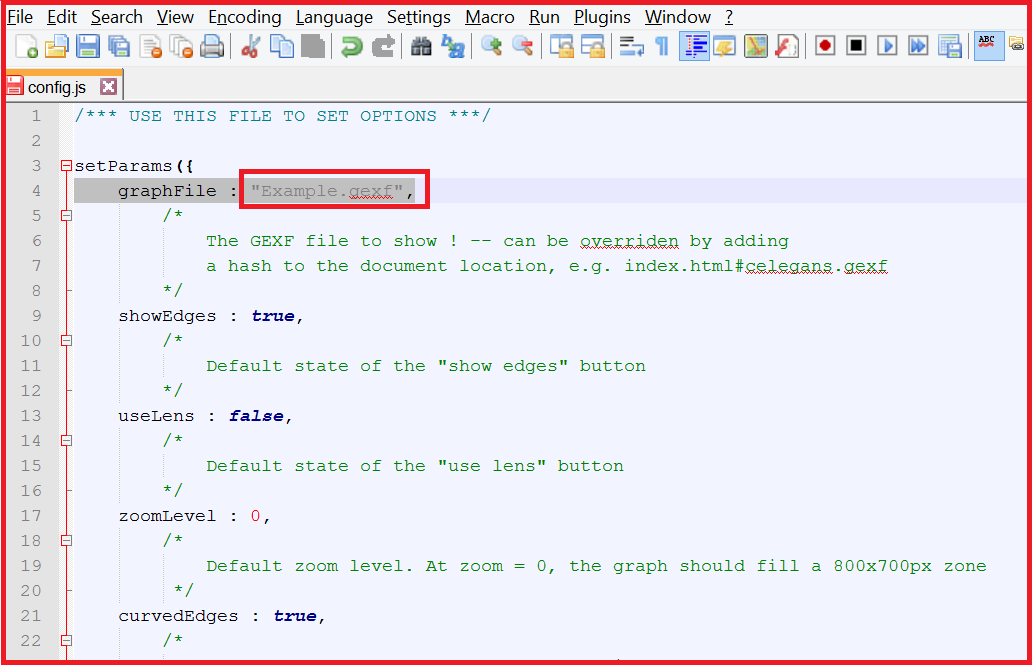
Follow the same steps as highlighted in **Step 5.1.2.**

**Note:** Ensure that the file name for both the type of files is **StudentName\_IntExt**

**Step 6: Display the Gephi file in the browser**

* Double click on the **gexf\_js\_master** folder created in the **prerequisites section under step 3** and then open the config.js files using text editors (Sublime, Notepad++, Brackets etc.)
* Copy and Place all the .gexf files created from above steps into **gexf\_js\_master** folder.
* Maintain your exported .gexf file as highlighted below

**Note**: You will have to do this step for each of the .gexf files

****

* Right click on Index.html in JS folder and then select open with any web browser of your choice. You should be able to view the Gephi file on the browser.
* Note: Use Mozilla Firefox, if it doesn’t open in any browser

**Question:**

1. **Display each of the 3 exported .gexf files in the browser and paste the screen shot below.**

**Answer:**

**Step 7: Attach assignments in elearning**

1. Attach the **assignment document** (only with the answers)in Microsoft Word
2. Attach all 3 f**iles** with extension **.gephi**