Riphah International University

Social Network Analysis Project

Submitted by: Mohammad Ayyaz Azeem

SAP ID: 12520

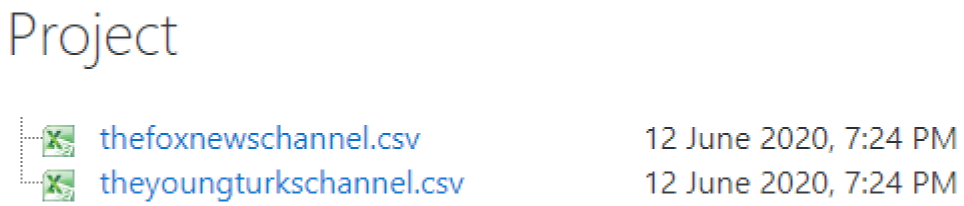
17-June-20

Contents

sdfs

Preprocessing

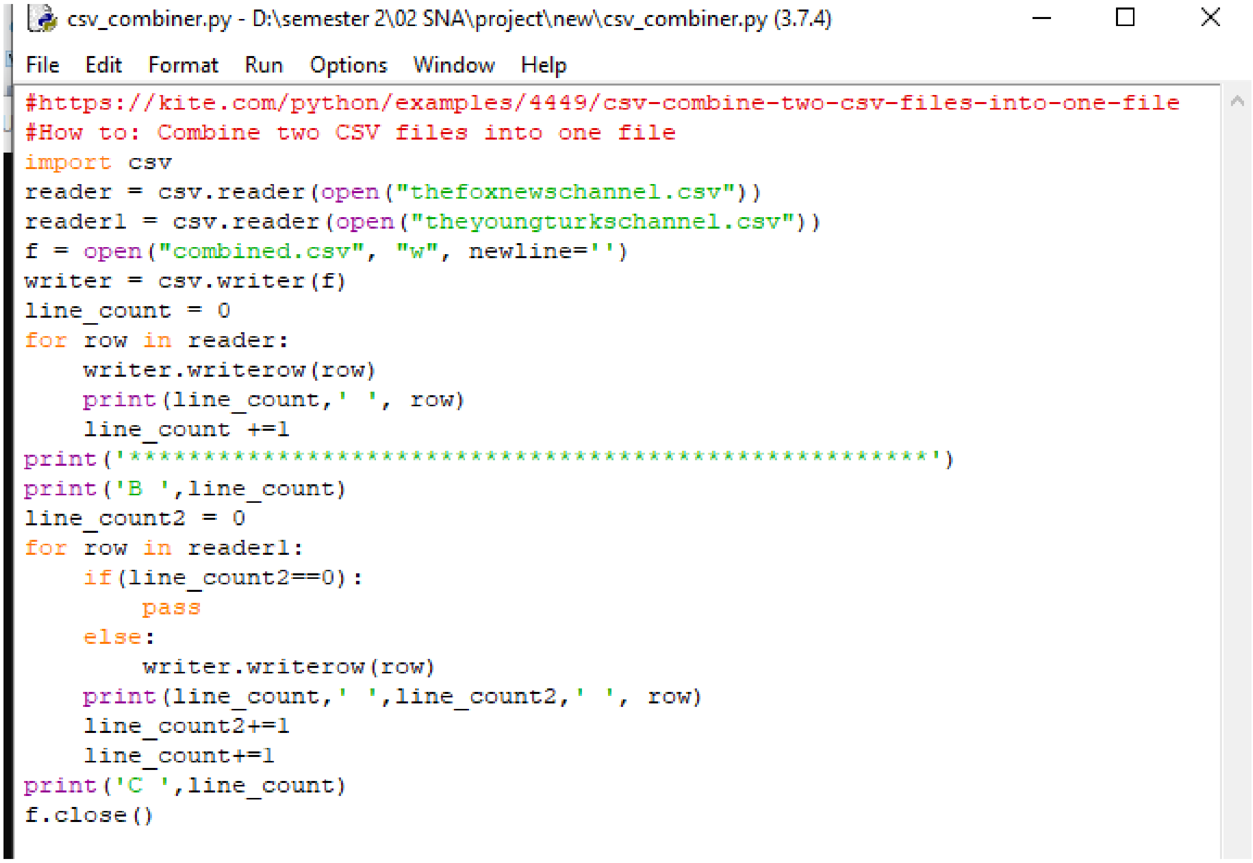
Two files given:



Project

thefoxnewschannel.csv	12 June 2020, 7:24 PM
theyoungturkschannel.csv	12 June 2020, 7:24 PM

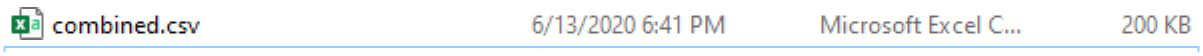
Combining both csv files using python code:



csv_combiner.py - D:\semester 2\02 SNA\project\new\csv_combiner.py (3.7.4)
File Edit Format Run Options Window Help

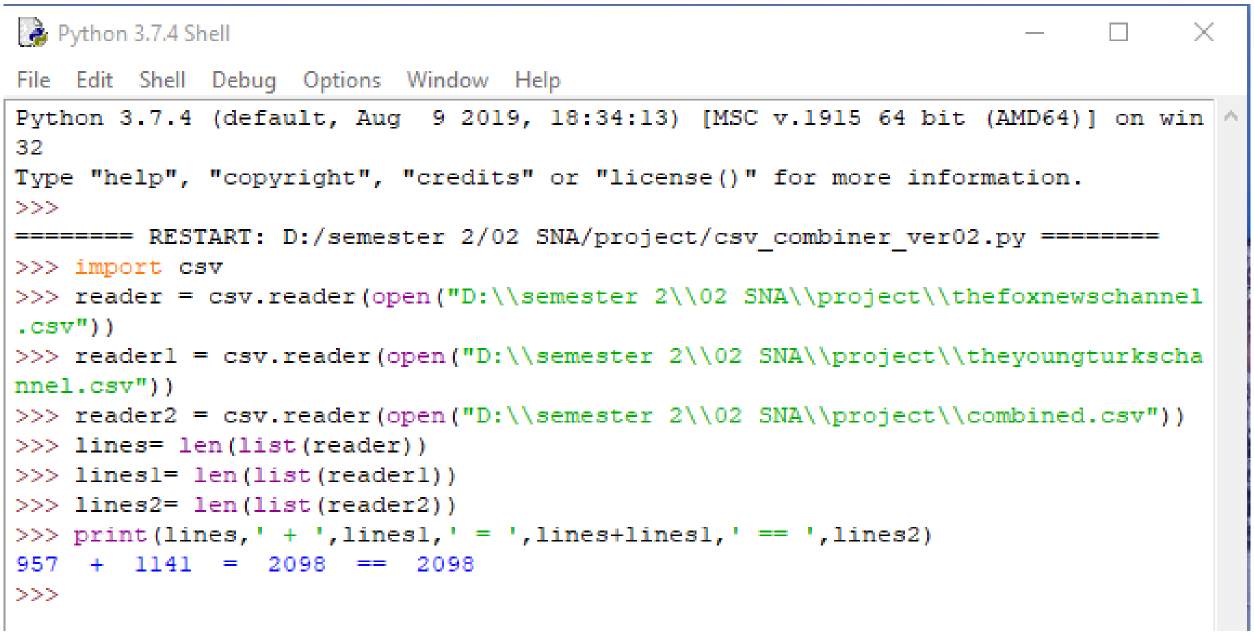
```
#https://kite.com/python/examples/4449/csv-combine-two-csv-files-into-one-file
#How to: Combine two CSV files into one file
import csv
reader = csv.reader(open("thefoxnewschannel.csv"))
reader1 = csv.reader(open("theyoungturkschannel.csv"))
f = open("combined.csv", "w", newline='')
writer = csv.writer(f)
line_count = 0
for row in reader:
    writer.writerow(row)
    print(line_count,' ', row)
    line_count +=1
print('******************************************************')
print('B ',line_count)
line_count2 = 0
for row in reader1:
    if(line_count2==0):
        pass
    else:
        writer.writerow(row)
    print(line_count,' ',line_count2,' ', row)
    line_count2+=1
    line_count+=1
print('C ',line_count)
f.close()
```

Combined file created:



combined.csv	6/13/2020 6:41 PM	Microsoft Excel C...	200 KB

Checking if combined data contains all data using python code:



```
Python 3.7.4 (default, Aug  9 2019, 18:34:13) [MSC v.1915 64 bit (AMD64)] on win
32
Type "help", "copyright", "credits" or "license()" for more information.
>>> 
======== RESTART: D:/semester 2/02 SNA/project/csv_combiner_ver02.py ========
>>> import csv
>>> reader = csv.reader(open("D:\\semester 2\\02 SNA\\project\\thefoxnewschannel
.csv"))
>>> reader1 = csv.reader(open("D:\\semester 2\\02 SNA\\project\\theyoungturkscha
nnel.csv"))
>>> reader2 = csv.reader(open("D:\\semester 2\\02 SNA\\project\\combined.csv"))
>>> lines= len(list(reader))
>>> lines1= len(list(reader1))
>>> lines2= len(list(reader2))
>>> print(lines,' + ',lines1,' = ',lines+lines1,' == ',lines2)
957  +  1141  =  2098  ==  2098
>>> 
```

# of rows in thefoxnewschannel.csv = 957

# of rows in theyoungturkschannel.csv = 1141

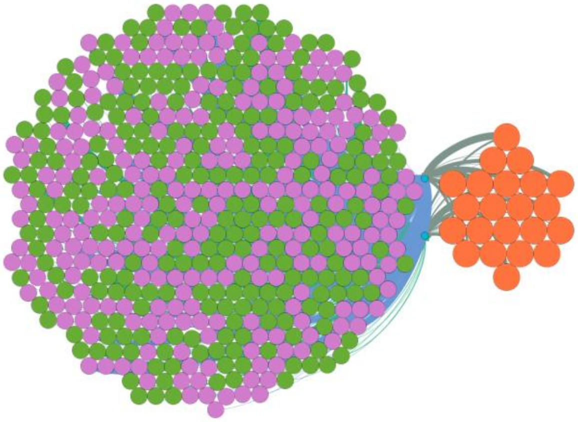
# of rows in combined file: 957+1141 = 2098

Task 1

Please attach your database that contains both of your YouTube channels in Gephi format. This means that you have to transform the two .csv files into a single one with only three columns.

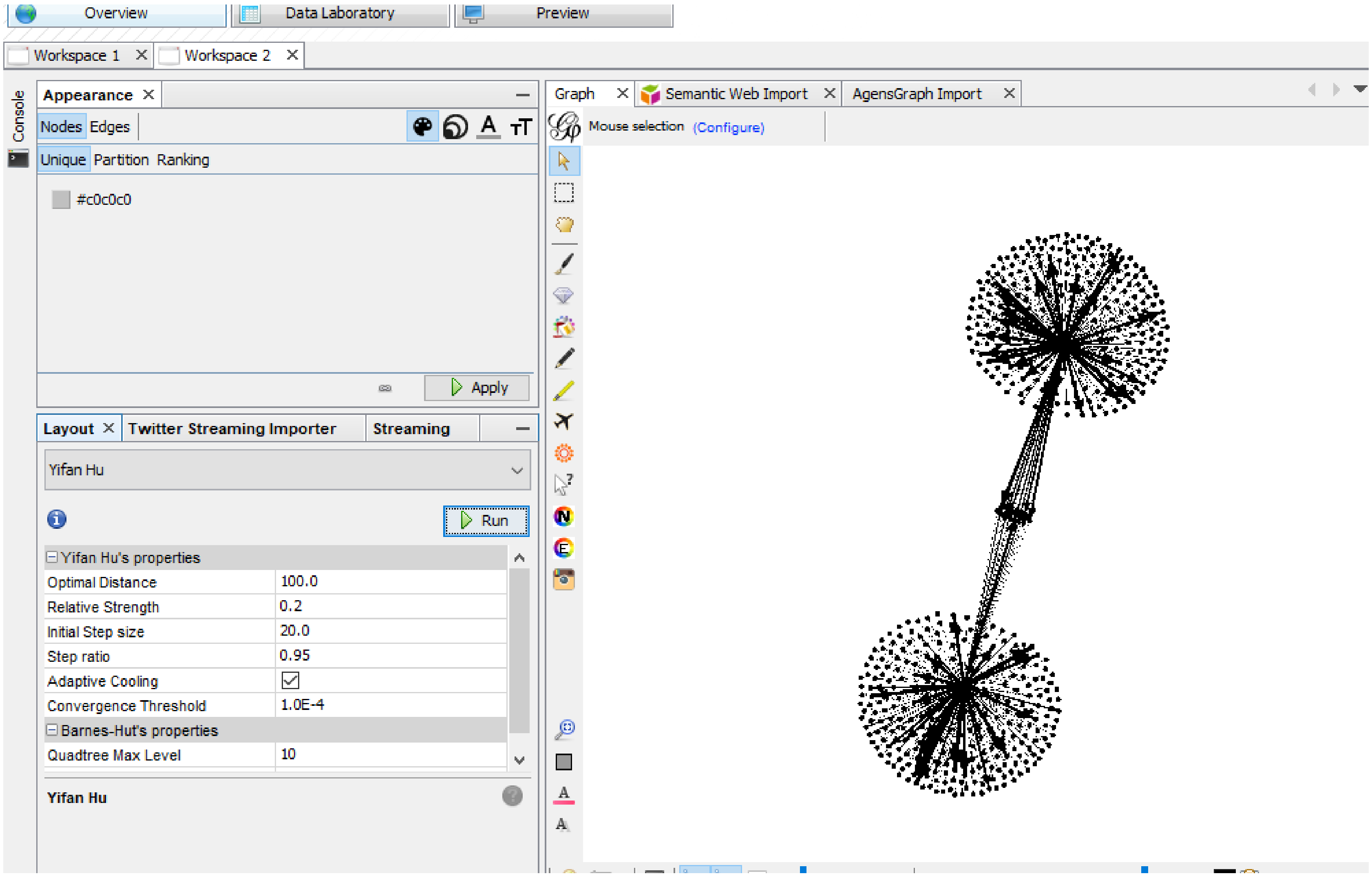
Pre-Processing step done. Combined.csv file attached with the zip file

Scenario explained: we have 2 channels and both publish videos with some of the videos are shared from both channels (channel # 1: Fox channel has 258 videos/nodes with 274 edges and channel # 2 Young Turk has 274 videos/nodes with 278 edges so in total we have 532 nodes), both have videos shared in it, 22 common videos (shown in red in below network) are shared from both channels. The test of 510 nodes is shown together below in green and pink color nodes.



Task 2

Attach a screenshot of your "Overview" tab in Gephi, which shows your network after you ran the "Yifan Hu" Layout algorithm.



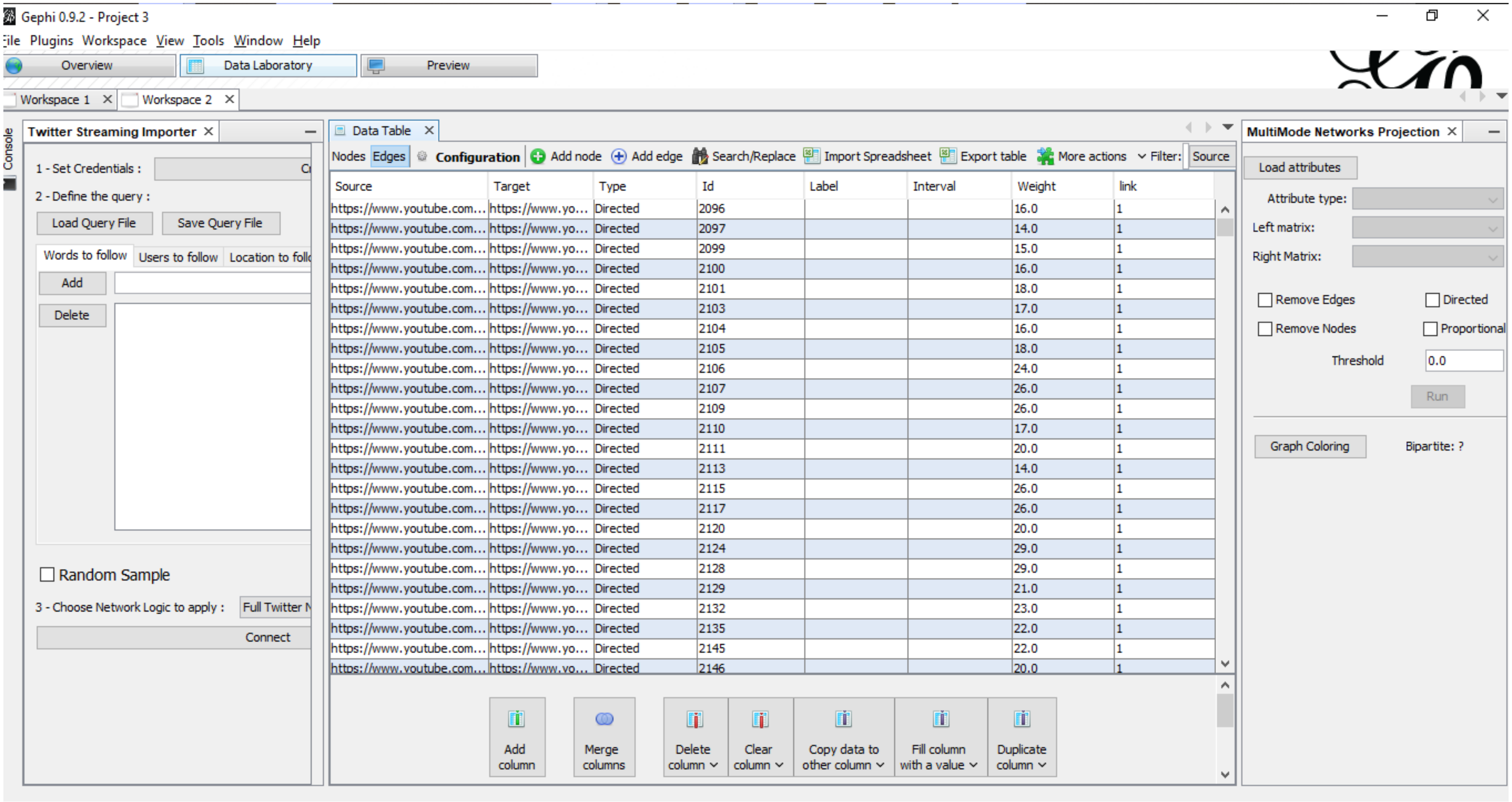
Yifan Hu

⊟ Yifan Hu's properties	
Optimal Distance	100.0
Relative Strength	0.2
Initial Step size	20.0
Step ratio	0.95
Adaptive Cooling	☑
Convergence Threshold	1.0E-4
⊟ Barnes-Hut's properties	
Quadtree Max Level	10

Yifan Hu

Task 3

Take a screen shot of the Data Table of Edges: separate screen shot in the zip folder as well

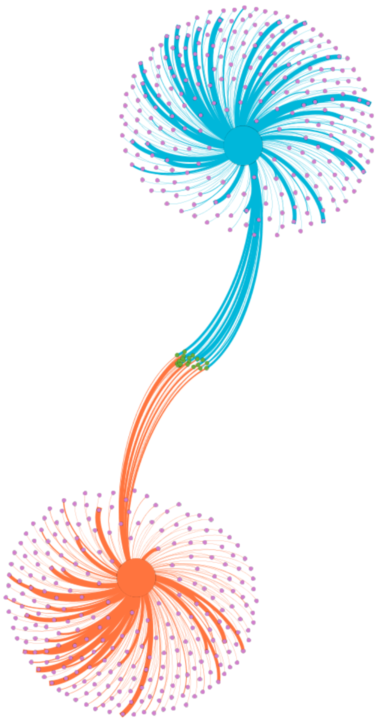


Source	Target	Type	Id	Label	Interval	Weight	link
https://www.youtube.com...	https://www.yo...	Directed	2096			16.0	1
https://www.youtube.com...	https://www.yo...	Directed	2097			14.0	1
https://www.youtube.com...	https://www.yo...	Directed	2099			15.0	1
https://www.youtube.com...	https://www.yo...	Directed	2100			16.0	1
https://www.youtube.com...	https://www.yo...	Directed	2101			18.0	1
https://www.youtube.com...	https://www.yo...	Directed	2103			17.0	1
https://www.youtube.com...	https://www.yo...	Directed	2104			16.0	1
https://www.youtube.com...	https://www.yo...	Directed	2105			18.0	1
https://www.youtube.com...	https://www.yo...	Directed	2106			24.0	1
https://www.youtube.com...	https://www.yo...	Directed	2107			26.0	1
https://www.youtube.com...	https://www.yo...	Directed	2109			26.0	1
https://www.youtube.com...	https://www.yo...	Directed	2110			17.0	1
https://www.youtube.com...	https://www.yo...	Directed	2111			20.0	1
https://www.youtube.com...	https://www.yo...	Directed	2113			14.0	1
https://www.youtube.com...	https://www.yo...	Directed	2115			26.0	1
https://www.youtube.com...	https://www.yo...	Directed	2117			26.0	1
https://www.youtube.com...	https://www.yo...	Directed	2120			20.0	1
https://www.youtube.com...	https://www.yo...	Directed	2124			29.0	1
https://www.youtube.com...	https://www.yo...	Directed	2128			29.0	1
https://www.youtube.com...	https://www.yo...	Directed	2129			21.0	1
https://www.youtube.com...	https://www.yo...	Directed	2132			23.0	1
https://www.youtube.com...	https://www.yo...	Directed	2135			22.0	1
https://www.youtube.com...	https://www.yo...	Directed	2145			22.0	1
https://www.youtube.com...	https://www.yo...	Directed	2146			20.0	1

Task 4

Calculate the average Degree of your network.
Display and analyze all three resulting network measures:

A. Degree
B. In-Degree
C. Out-Degree



Take screenshot of your network

Answer the following questions:

1. What is the difference between them?

Answer:

Graph is collection of vertices (V) and edges (E). Generic notation for representing graph is G(V,E). Every node is called as vertex and lines connecting the nodes are called as edges. There are two types of graph. An undirected graph is graph, i.e., a set of objects (called vertices or nodes) that are connected together, where all the edges are bidirectional. An undirected graph is sometimes called an undirected network. In contrast, a graph where the edges point in a direction is called a directed graph.

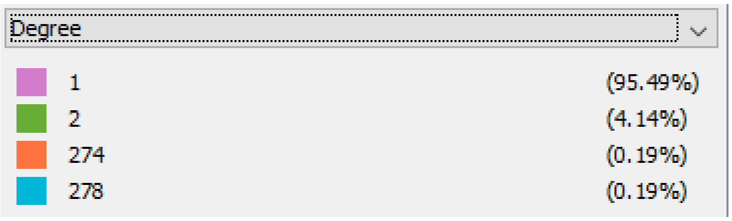
Degree: The degree of vertex is number of edges that are connected to the vertex. Degree is the edges incident on a node. According to Wikipedia, The degree of a node in a network (sometimes referred to incorrectly as the connectivity) is the number of connections or edges the node has to other nodes.

In Degree: This is applicable only for directed graph. This represents the number of edges incoming to a vertex.

Out degree: This is also applicable only for directed graph. This represents the number of edges outgoing from a vertex.

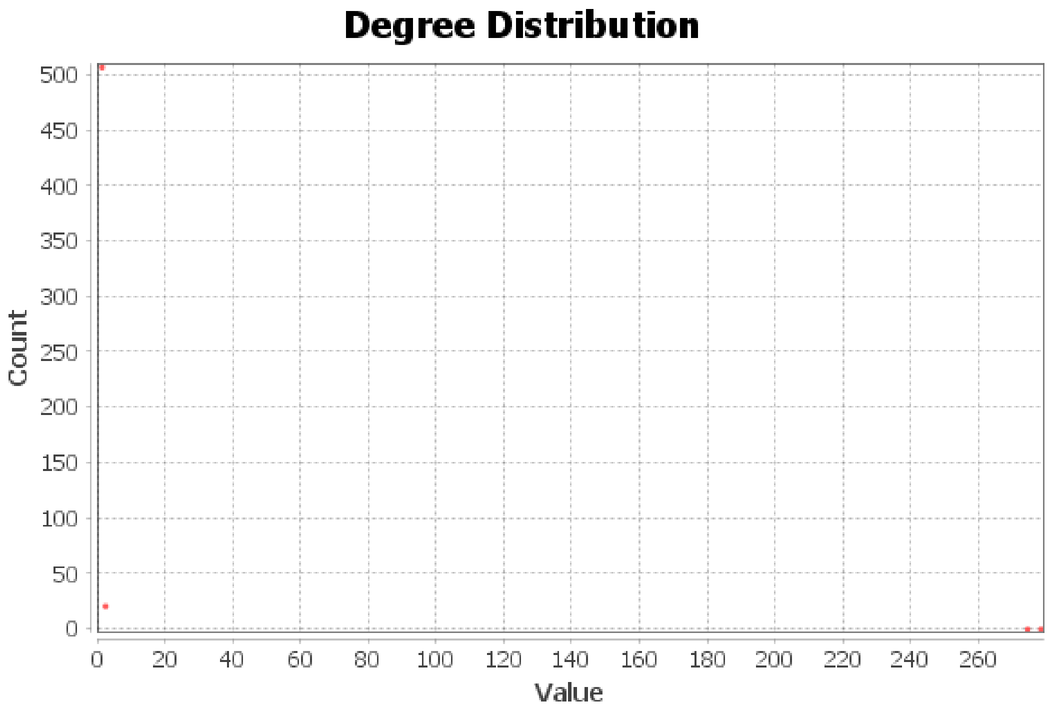
2. How many categories do you get for each?

Degree: 4 {1, 2, 274, 278}

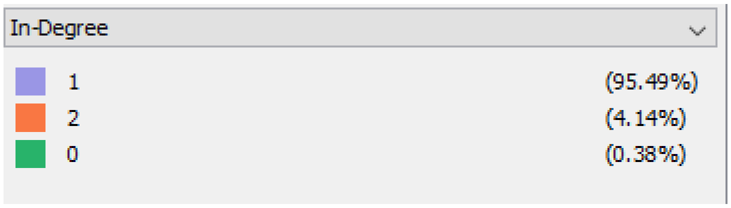


Degree

🟪	1	(95.49%)
🟩	2	(4.14%)
🟧	274	(0.19%)
🟦	278	(0.19%)



In-Degree: 3 {0, 1, 2}

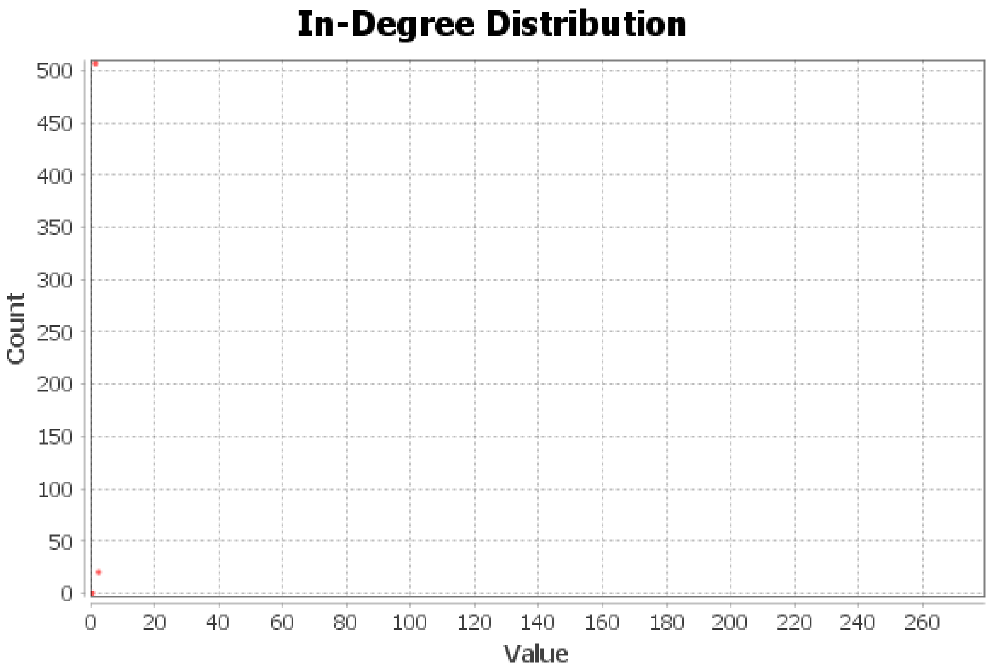


In-Degree

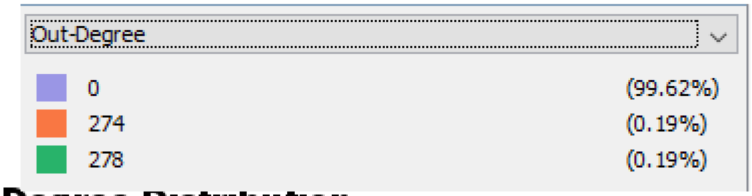
🟪 1 (95.49%)

🟧 2 (4.14%)

🟩 0 (0.38%)

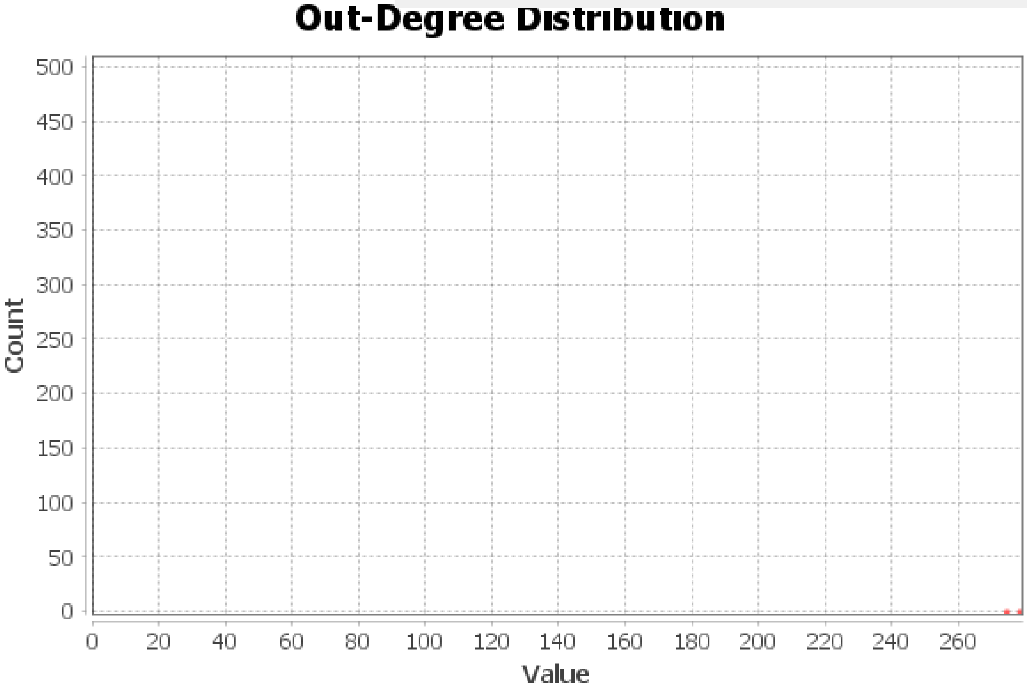


Out-Degree: 3 {0, 274, 278}



Out-Degree

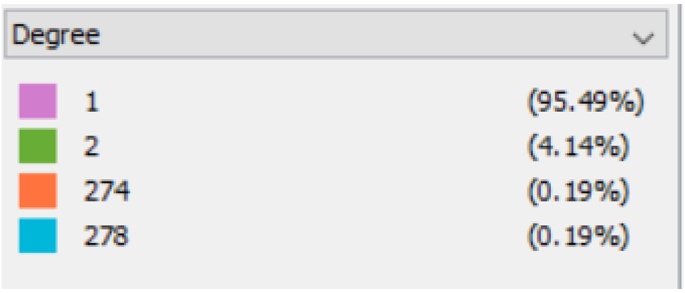
🟪	0	(99.62%)
🟧	274	(0.19%)
🟩	278	(0.19%)



3. Can you make sense of the numbers it indicates the number of degree per category for each of the three measures? Why or why not?

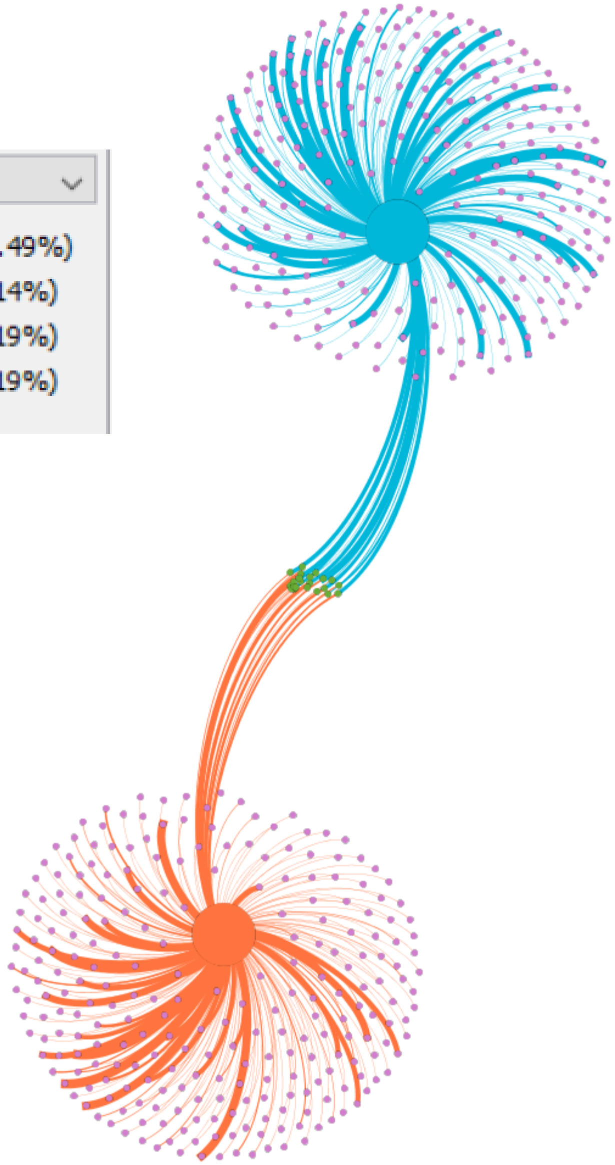
Degree: Videos shared by 2 channels are represented here. Total videos shared are 532 from both channels.

1. Pink nodes have degree 1 represents the nodes/videos that are shared from each channel. Whereas 95.49% means 508 out of 532 nodes/videos have degree 1 means these videos are shared from their respective channels.
2. Green nodes represents videos shared from both channels (common videos on both channels), 4.41% means 22 out of 532 videos are common in both channels therefore have degree of 2.
3. Orange means channel 1: shared 274 videos whereas 0.19% means 1 out of 532 nodes have degree 278. This represents channel Young Turk.
4. Blue means channel 2: shared 258 videos whereas 0.19% means 1 out of 532 nodes have degree 274. This represents channel Fox.



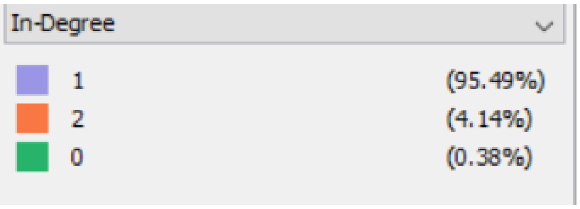
Degree

🟪	1	(95.49%)
🟩	2	(4.14%)
🟧	274	(0.19%)
🟦	278	(0.19%)

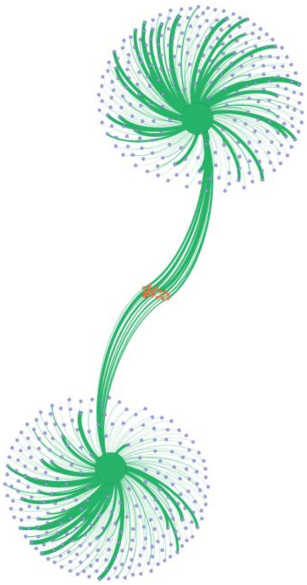


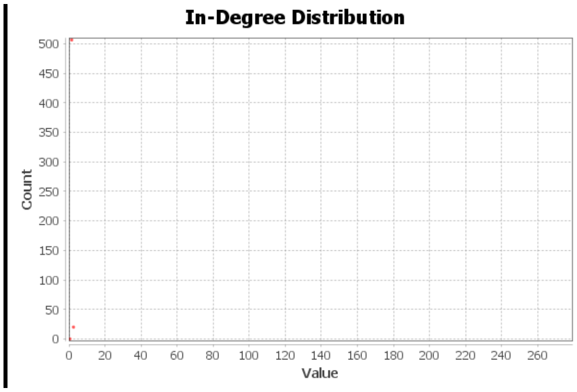
In-degree:

1. Purple Nodes: 95.49% means 508 out of 532 nodes have in-degree = 1, represents the videos that are published by respective channel.
2. Orange Nodes: 4.14% means 22 out of 532 nodes have in-degree = 2, represents videos that are published by both channels.
3. Green Nodes: 0.38% means 2 out of 532 nodes have in-degree = 0, represents both channels that publish videos.



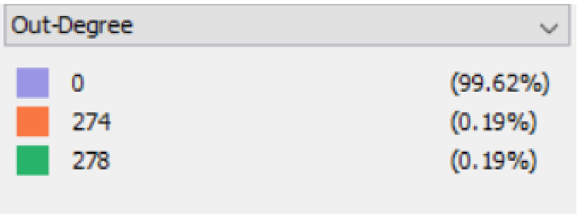
In-Degree		
🟪	1	(95.49%)
🟧	2	(4.14%)
🟩	0	(0.38%)



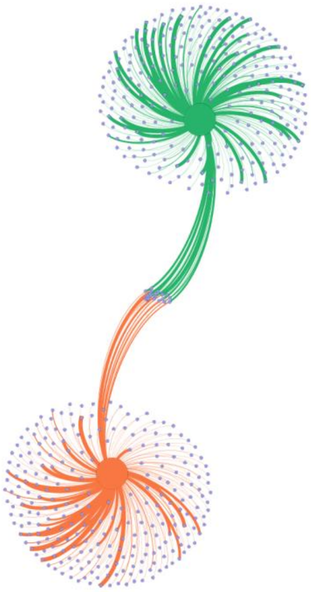


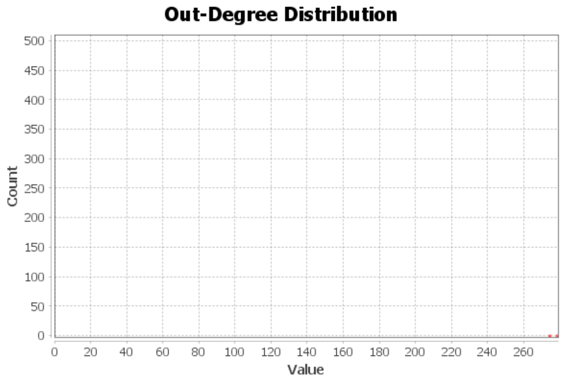
Out-degree

1. Purple Nodes: 99.62% means 530 out of 532 nodes have out-degree = 0, represents the videos that are published by respective channel.
2. Orange Nodes: 0.19% means 1 out of 532 nodes have out-degree = 274, represents channel #1 Young Turk that has published 274 videos.
3. Green Nodes: 0.19% means 1 out of 532 nodes have out-degree = 258, represents channels # 2 that publish 278 videos.



Out-Degree		
🟪	0	(99.62%)
🟧	274	(0.19%)
🟩	278	(0.19%)





Task 5

1. How many nodes (videos) are shared by both YouTube channels? Count them or calculate them.

   Solution:

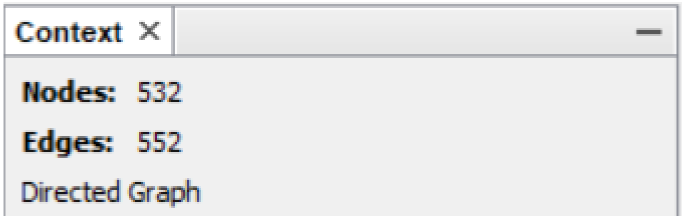
   As obvious from the Context tab:

   Total videos shared: 532

   Channel # 1 published: 51.5% which is 274 out of 532 videos

   Channel # 2 published: 48.5% which is 258 out of 532 videos

   Both channels published: 22 same videos



Nodes: 532

Edges: 552

Directed Graph

   Now by looking at the graph shown we can see that the 22 videos are published from both channels. The network shows that 17 videos belong to 1st channel and 5 to the second.

   So new calculations for both channels are (calculations made by analyzing visually):

   Channel # 1: 257 individual videos plus 17 videos that are also shared on other channel that makes a total of 274 videos has 278 edges

   Channel # 2: 253 individual videos plus 5 videos that are also shared on other channel that makes a total of 258 videos has 274 edges

2. Calculate the network Modularity and take screen shot of the network.

   Solution:

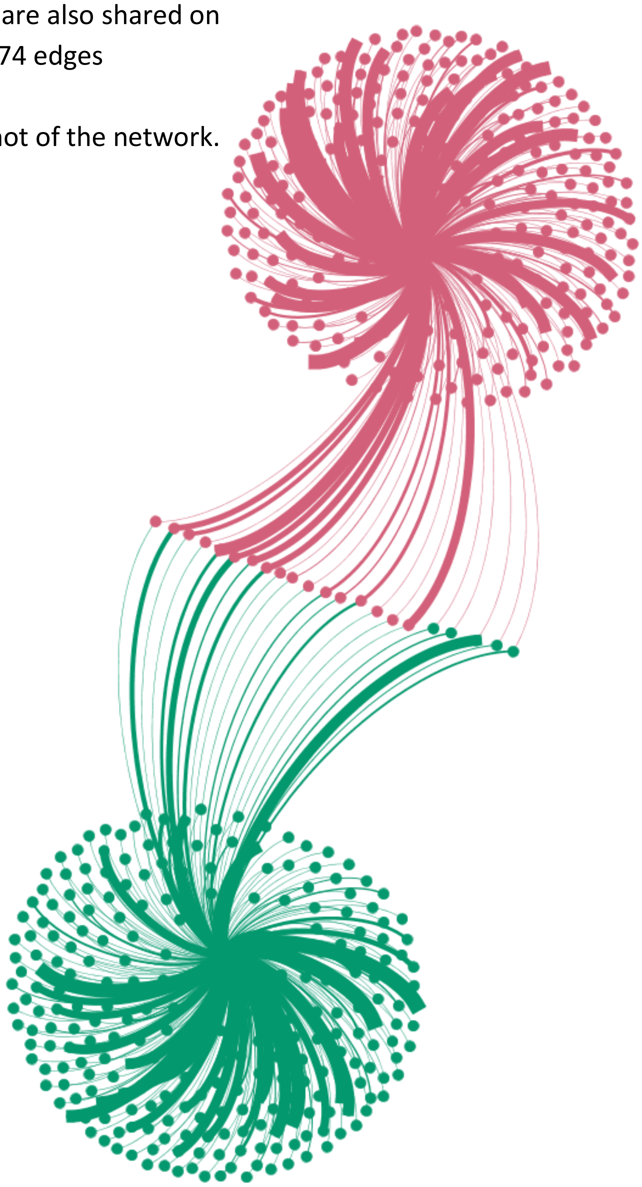
   Modularity Report

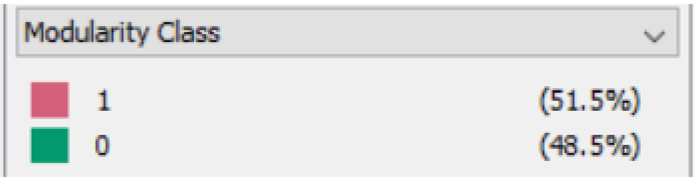
   Parameters:

   - Randomize: On
   - Use edge weights: On
   - Resolution: 1.0

   Results:

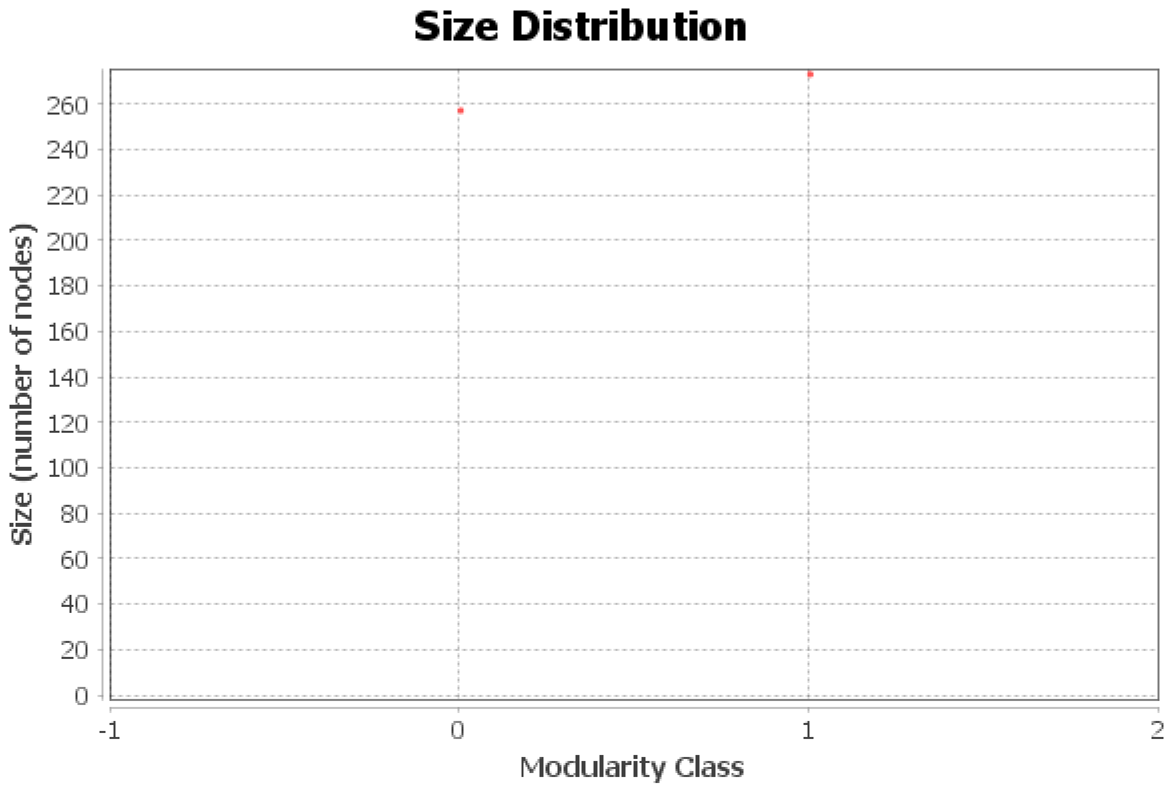
   - Modularity: 0.465
   - Modularity with resolution: 0.465
   - Number of Communities: 2





Modularity Class

🟥	1	(51.5%)
🟩	0	(48.5%)



Task 6

Calculate the "Undirected Closeness Centrality" for your network, through "Average Path Length" (attach network screen shot) and then answer the questions:

1. How many groups of nodes do you get?

Solution: 5 groups with values shown below

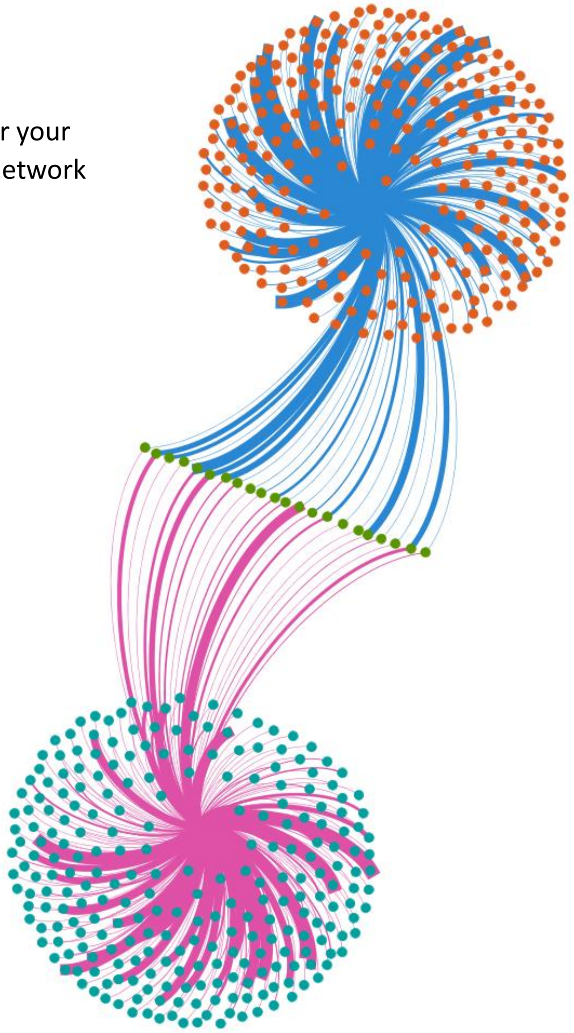
Orange Nodes: 48.12% is 256/532 nodes

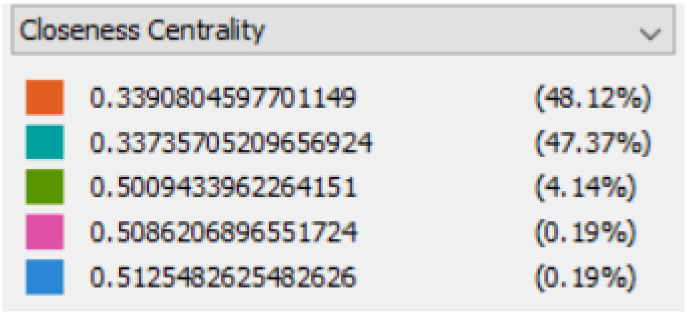
Teal Nodes: 47.37% is 252/532 nodes

Green Nodes: 4.14% is 22/532 nodes

Pink Nodes: 0.19% is 1/532 nodes

Blue Nodes: 0.19% is 1/532 nodes





Closeness Centrality

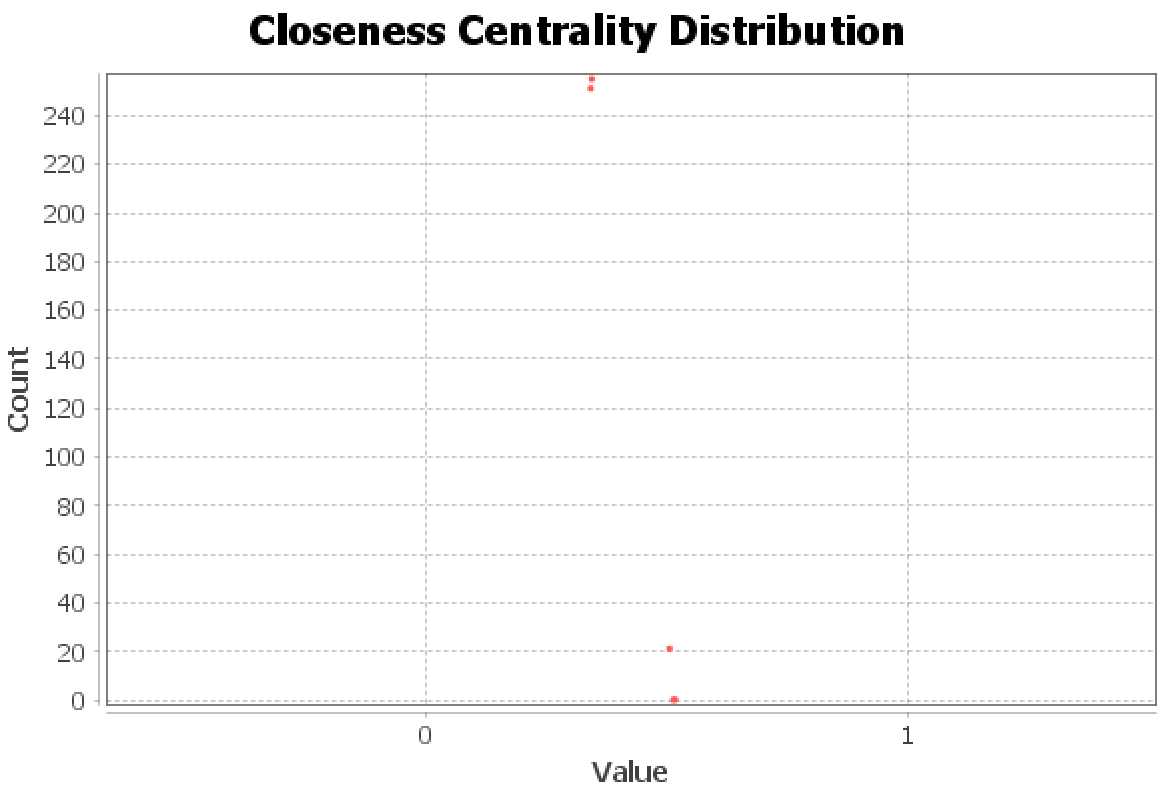
🟧 0.3390804597701149 (48.12%)

🟩 0.33735705209656924 (47.37%)

🟩 0.5009433962264151 (4.14%)

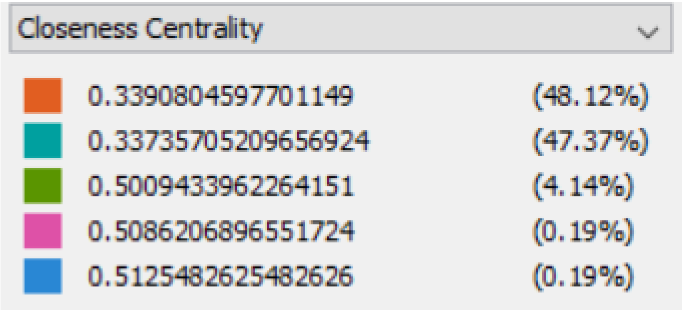
🟪 0.5086206896551724 (0.19%)

🟦 0.5125482625482626 (0.19%)



2. Please interpret the different groups. Which nodes are part of which group and why?

Solution:


Closeness Centrality

🟧	0.3390804597701149	(48.12%)
■	0.33735705209656924	(47.37%)
🟩	0.5009433962264151	(4.14%)
🟪	0.5086206896551724	(0.19%)
🟦	0.5125482625482626	(0.19%)

Closeness centrality is dependent on the position of the node in the network

Closeness centrality measures the mean distance from a vertex to other vertices. Simply it is representing the node that is more close to all the other nodes. Here,

1. Highest closeness centrality: Blue Nodes: 0.19% is 1/532 nodes: meaning 1 out of 532 nodes is closest to majority of nodes that's why it has the highest closeness centrality of 0.5125: this represents channel # 1: Young Turk channel
2. 2nd highest closeness centrality: Pink Nodes: 0.19% is 1/532 nodes: meaning 1 out of 532 nodes is 2nd closest to majority of nodes that's why it has the 2nd highest closeness centrality of 0.5086. this represents channel # 2: Fox channel
3. 3rd highest closeness centrality: Green Nodes: 4.14% is 22/532 nodes: meaning 22 out of 532 nodes are 3rd closest to majority of nodes that why it has the 3rd highest closeness centrality of 0.5009. These 22 are those videos that are published from both channels.
4. 4th highest closeness centrality: Orange Nodes: 48.12% is 256/532 nodes: meaning 256 out of 532 nodes are 4th closest to majority of nodes that why it has the 4th highest closeness centrality of 0.3390. These are 256 videos from Fox Channel.
5. 5th highest closeness centrality: Teal Nodes: 47.37% is 252/532 nodes: meaning 252 out of 532 nodes are least close to majority of nodes that why it has the lowest closeness centrality of 0.3373. These are 252 vides from Young Turk Channels.

3. Calculate the "directed Closeness Centrality" for your network, through "Average Path Length"(Also attach screenshot)

Solution:

Parameters:

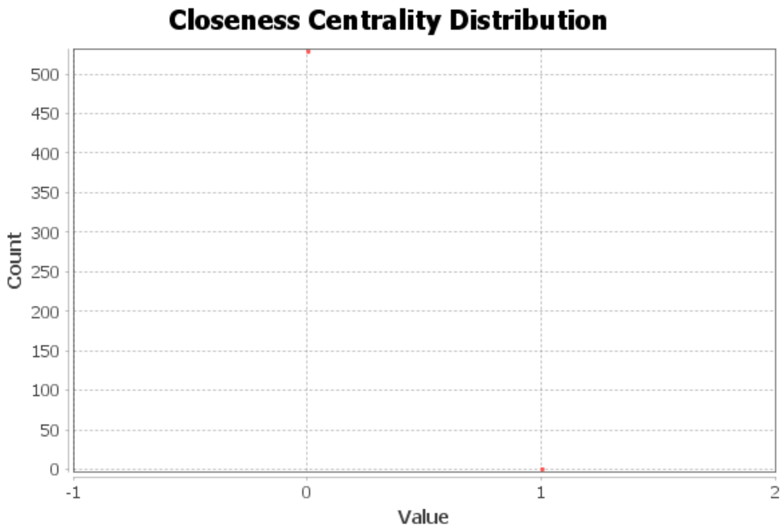
- Network Interpretation: directed

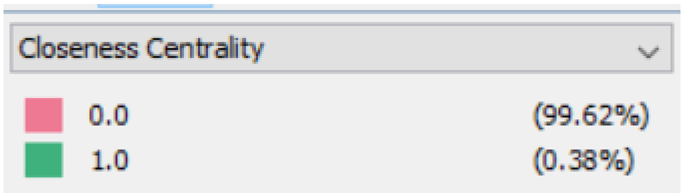
Results:

- Diameter: 1
- Radius: 0
- Average Path length: 1.0

Explanation:

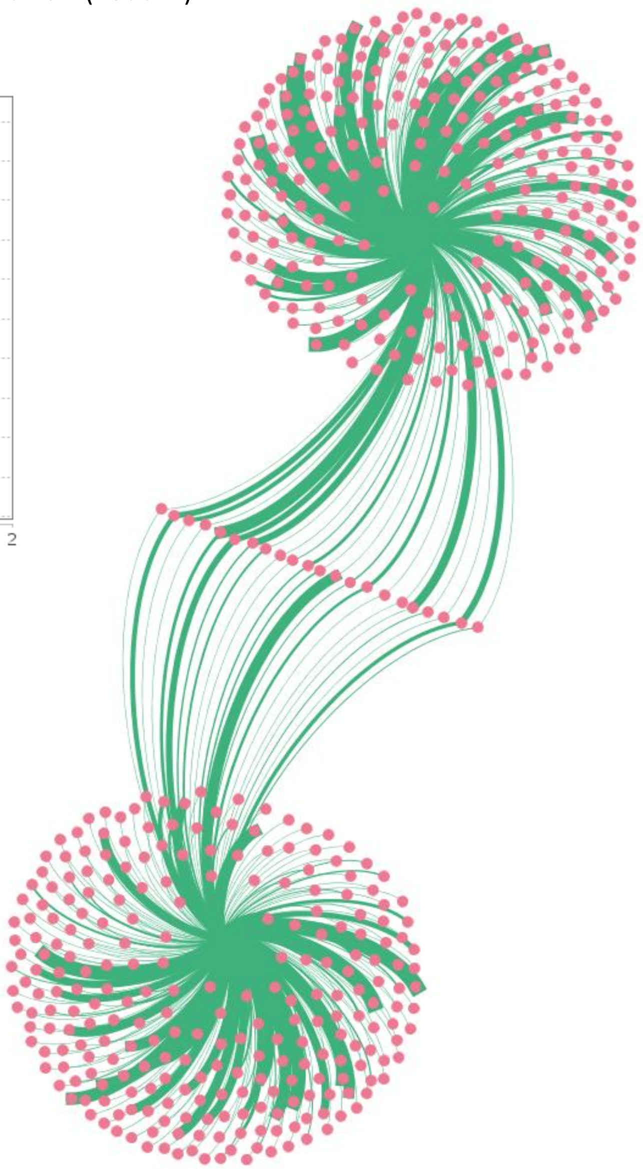
- Pink Nodes: 99.62% is 530/532 nodes are all 530 videos shared from both channels
- Green Nodes: 0.38% is 2/532 nodes are both channels that is Young Turk (node#1) and Fox (node#2)





Closeness Centrality

🟥	0.0	(99.62%)
🟩	1.0	(0.38%)



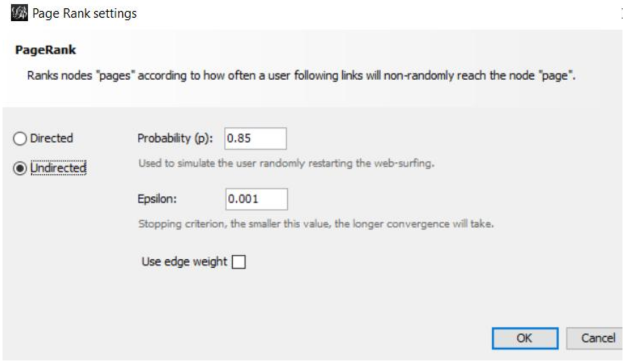
Task 7

Calculate PageRank for your network, a special version of Eigenvector Centrality. Then answer the following questions: (attach screen shot of the network)

1. How many groups of nodes do you get for PageRank?

Solution: both un-directed and directed page rank are shown in **2 columns** side by side for better analysis (two column structure as shown in in-degree out-degree page 11)

Un-directed Page Rank: 5 groups



Page Rank settings

PageRank

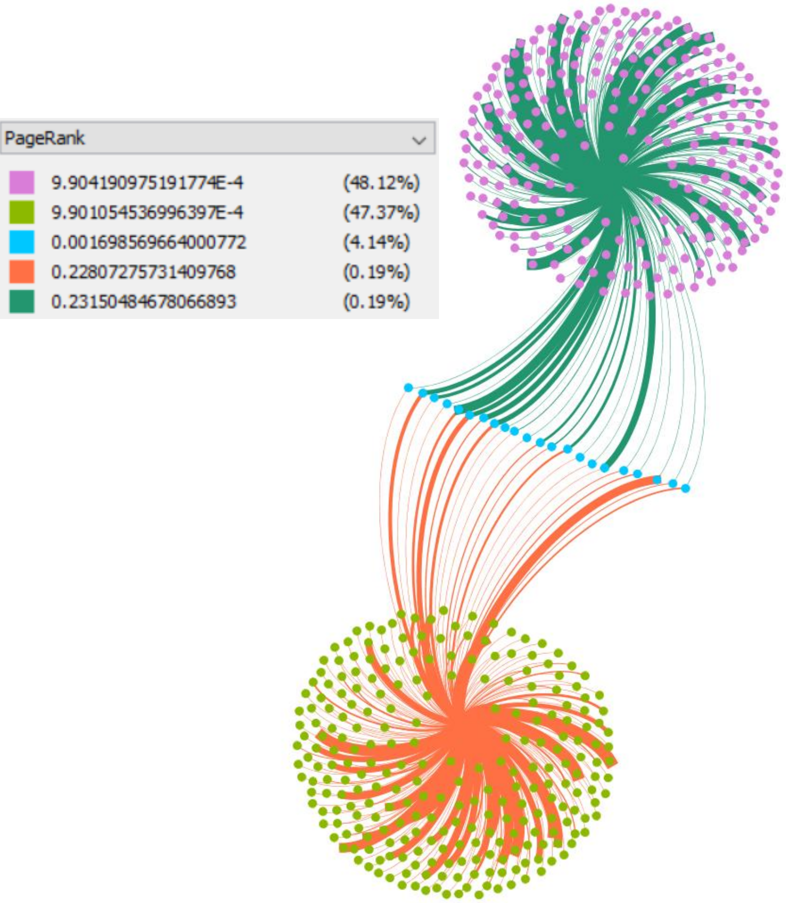
Ranks nodes "pages" according to how often a user following links will non-randomly reach the node "page".

○ Directed
● Undirected

Probability (p): 0.85
Used to simulate the user randomly restarting the web-surfing.

Epsilon: 0.001
Stopping criterion, the smaller this value, the longer convergence will take.

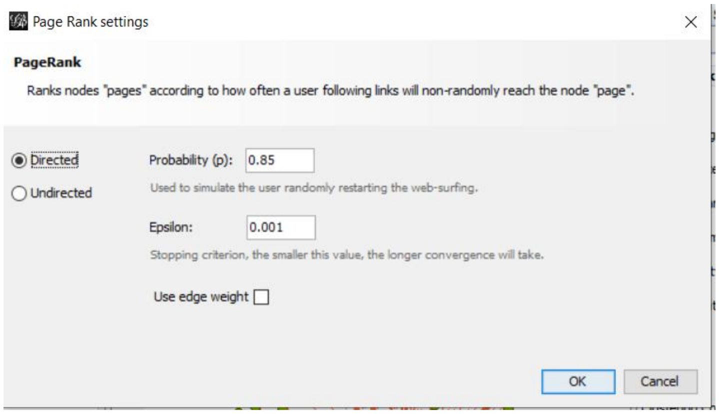
Use edge weight □



PageRank

🟪	9.904190975191774E-4	(48.12%)
🟩	9.901054536996397E-4	(47.37%)
🟦	0.001698569664000772	(4.14%)
🟧	0.22807275731409768	(0.19%)
🟩	0.23150484678066893	(0.19%)

Directed Page Rank: 4 groups



Page Rank settings

PageRank

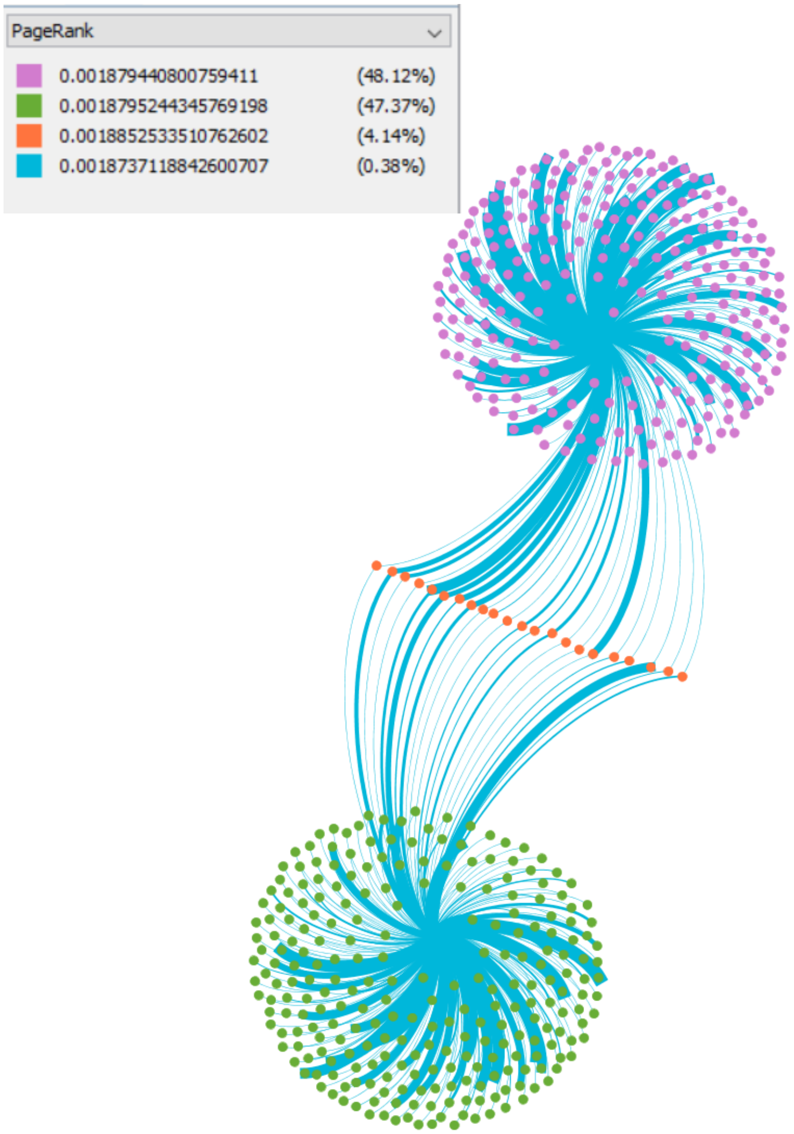
Ranks nodes "pages" according to how often a user following links will non-randomly reach the node "page".

● Directed
○ Undirected

Probability (p): 0.85
Used to simulate the user randomly restarting the web-surfing.

Epsilon: 0.001
Stopping criterion, the smaller this value, the longer convergence will take.

Use edge weight □



PageRank

🟪	0.001879440800759411	(48.12%)
🟩	0.0018795244345769198	(47.37%)
🟧	0.0018852533510762602	(4.14%)
🟦	0.0018737118842600707	(0.38%)

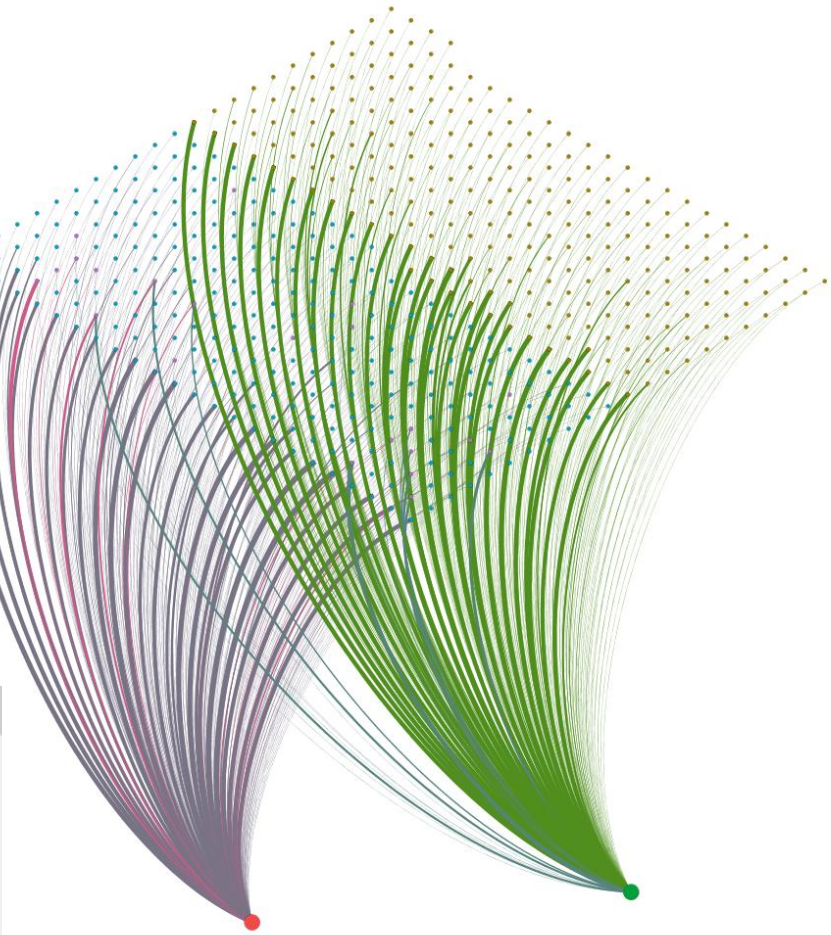
2. What do they measure?

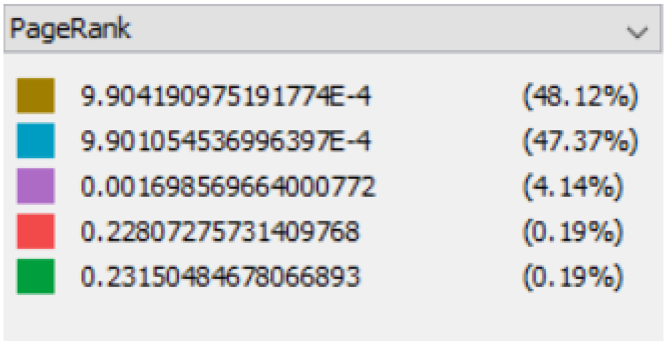
   Solution: The underlying assumption in page rank is that more important websites/nodes are likely to receive more links from other websites/nodes.

   We checked for self-loop and found none and we sum the page rank calculations and found it to be 1 (proving no self-loop)

For undirected Page rank has 5 groups

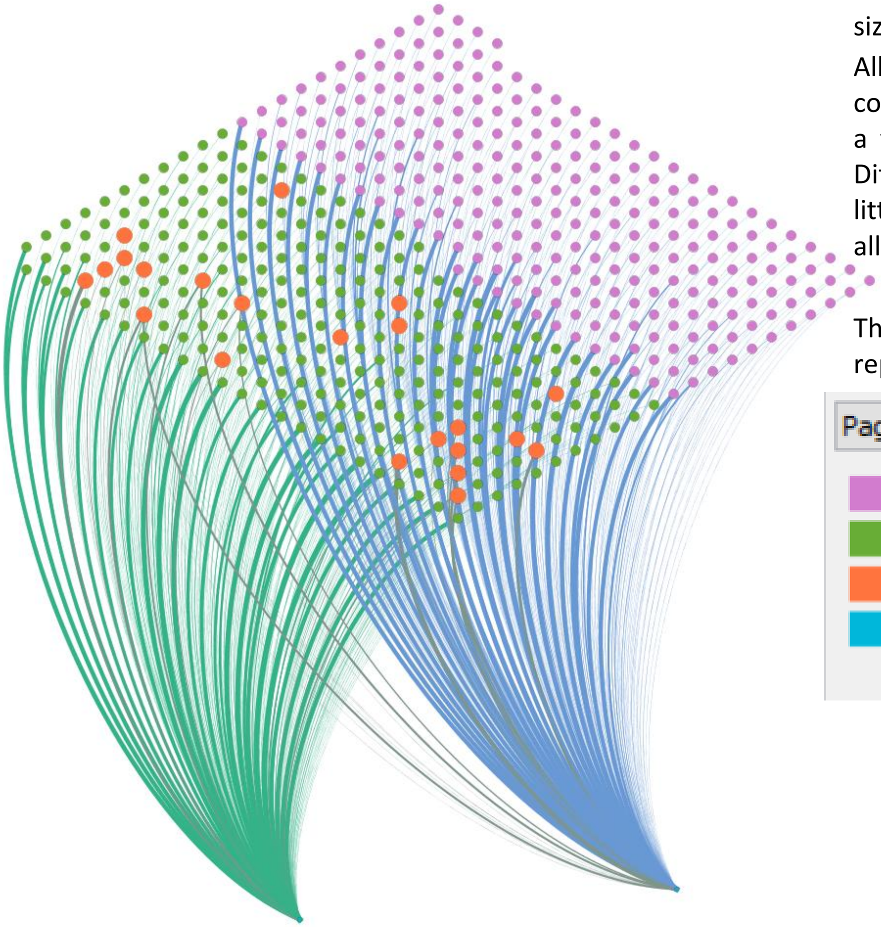
1. Green Node has the highest page rank value of 0.2315: represents channel # 1 Young Turk that shared 274 videos. The node size is also bigger than the rest.
2. Red Node has 2nd highest page rank value of 0.2280: represents channel # 2 Fox Channel that shared 258 videos. The node size is also bigger than the rest.
3. Purple Node has 3rd highest page rank value of 0.0016: represents the 22 videos shared by both channels
4. Blue Node and Sea Green has 4th& 5th highest page rank value respectively considerably less than 0 representing all the 257+253 = 510 videos shared by respective channels (**257**+22+**253**=532)





PageRank

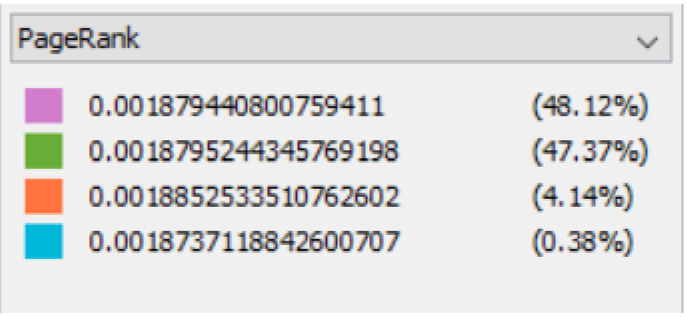
🟫	9.904190975191774E-4	(48.12%)
🟦	9.901054536996397E-4	(47.37%)
🟪	0.001698569664000772	(4.14%)
🟥	0.22807275731409768	(0.19%)
🟩	0.23150484678066893	(0.19%)



For directed Page rank: has 4 groups (same node size in all cases as same value 0.0018)

All nodes have similar values and based on the color we can see a few nodes but we cannot make a valid assumption by using directed page rank. Different colors assigned to nodes based on very little variation in the calculated rank. That shows all nodes are of equal importance in this case.

The 22 nodes are clearly shown in red here that represents the videos shared by both channels.



PageRank

🟪	0.001879440800759411	(48.12%)
🟩	0.0018795244345769198	(47.37%)
🟧	0.0018852533510762602	(4.14%)
🟦	0.0018737118842600707	(0.38%)

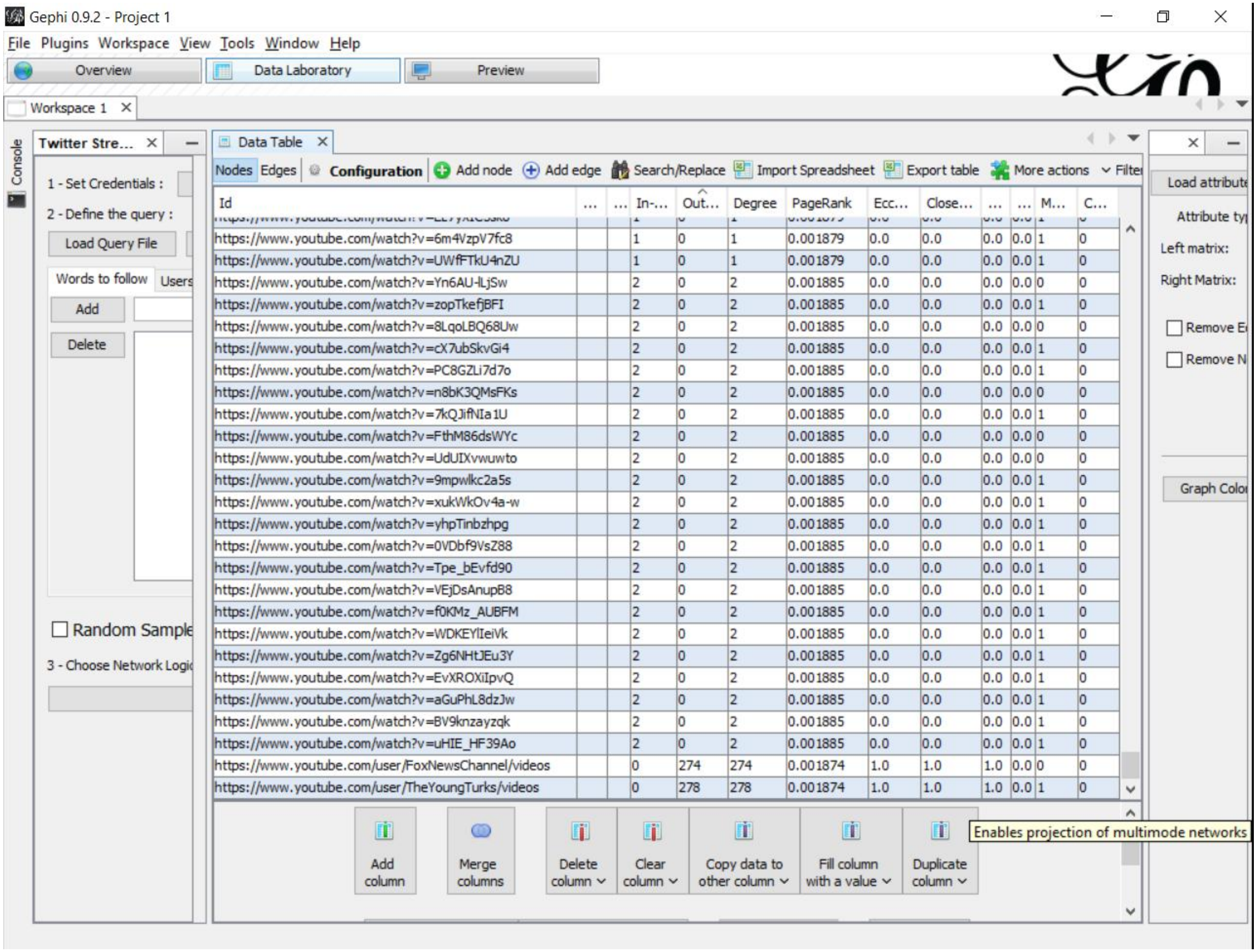
3. Is this useful?
   Solution: in case of un-directed network we can see that the important nodes with highest page rank (Red & Green color nodes in our case referring to question 2 task7 undirected case) can be clearly seen based on size of the node.
   Whereas in directed network case all 532 nodes are almost of the same size, with color we can make sense if we already know the network but we cannot clearly sate which node has more value based on directed page rank.

Attach Screen shot of each if any changes done in network

Task 8

Please attach a screenshot of your "Data Laboratory" tab, now at the end, after you have done the preceding analysis.
Go to "Data Table > Nodes" (not Edges) and make sure that the "Id" column is completely readable (not cut off to its right):

1. Take a screenshot of the full size "Data Laboratory" window (not just the part shown in this excerpt screen shot).
   Solution: Excel file attached with the zip file (named: task8_file.csv)

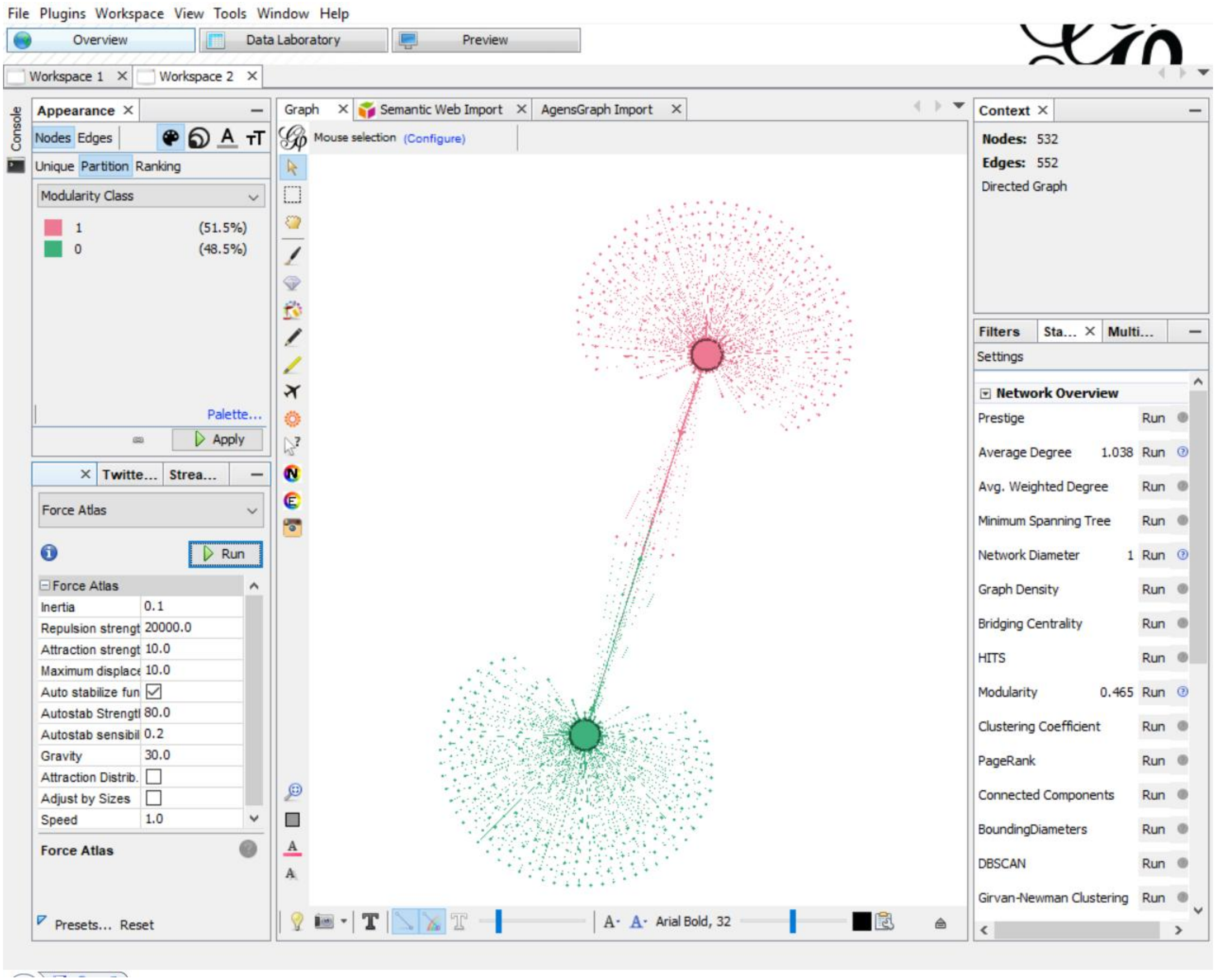


Id	...	...	In-...	Out...	Degree	PageRank	Ecc...	Close...	...	...	M...	C...
https://www.youtube.com/watch?v=6m4VzpV7fc8			1	0	1	0.001879	0.0	0.0	0.0	0.0	1	0
https://www.youtube.com/watch?v=UWfFTkU4nZU			1	0	1	0.001879	0.0	0.0	0.0	0.0	1	0
https://www.youtube.com/watch?v=Yn6AU-lLjSw			2	0	2	0.001885	0.0	0.0	0.0	0.0	0	0
https://www.youtube.com/watch?v=zopTkefjBFI			2	0	2	0.001885	0.0	0.0	0.0	0.0	1	0
https://www.youtube.com/watch?v=8LqoLBQ68Uw			2	0	2	0.001885	0.0	0.0	0.0	0.0	0	0
https://www.youtube.com/watch?v=cX7ubSkvGi4			2	0	2	0.001885	0.0	0.0	0.0	0.0	1	0
https://www.youtube.com/watch?v=PC8GZLi7d7o			2	0	2	0.001885	0.0	0.0	0.0	0.0	1	0
https://www.youtube.com/watch?v=n8bK3QMsFKs			2	0	2	0.001885	0.0	0.0	0.0	0.0	0	0
https://www.youtube.com/watch?v=7kQJifNIa1U			2	0	2	0.001885	0.0	0.0	0.0	0.0	1	0
https://www.youtube.com/watch?v=FthM86dsWYc			2	0	2	0.001885	0.0	0.0	0.0	0.0	0	0
https://www.youtube.com/watch?v=UdUIXvwuwto			2	0	2	0.001885	0.0	0.0	0.0	0.0	0	0
https://www.youtube.com/watch?v=9mpwlkc2a5s			2	0	2	0.001885	0.0	0.0	0.0	0.0	1	0
https://www.youtube.com/watch?v=xukWkOv4a-w			2	0	2	0.001885	0.0	0.0	0.0	0.0	1	0
https://www.youtube.com/watch?v=yhpTinbzhpg			2	0	2	0.001885	0.0	0.0	0.0	0.0	1	0
https://www.youtube.com/watch?v=0VDbf9VsZ88			2	0	2	0.001885	0.0	0.0	0.0	0.0	1	0
https://www.youtube.com/watch?v=Tpe_bEvfd90			2	0	2	0.001885	0.0	0.0	0.0	0.0	1	0
https://www.youtube.com/watch?v=VEjDsAnupB8			2	0	2	0.001885	0.0	0.0	0.0	0.0	1	0
https://www.youtube.com/watch?v=f0KMz_AUBFM			2	0	2	0.001885	0.0	0.0	0.0	0.0	1	0
https://www.youtube.com/watch?v=WDKEYlIeiVk			2	0	2	0.001885	0.0	0.0	0.0	0.0	1	0
https://www.youtube.com/watch?v=Zg6NHtJEu3Y			2	0	2	0.001885	0.0	0.0	0.0	0.0	1	0
https://www.youtube.com/watch?v=EvXROXiIpvQ			2	0	2	0.001885	0.0	0.0	0.0	0.0	1	0
https://www.youtube.com/watch?v=aGuPhL8dzJw			2	0	2	0.001885	0.0	0.0	0.0	0.0	1	0
https://www.youtube.com/watch?v=BV9knzayzqk			2	0	2	0.001885	0.0	0.0	0.0	0.0	1	0
https://www.youtube.com/watch?v=uHIE_HF39Ao			2	0	2	0.001885	0.0	0.0	0.0	0.0	1	0
https://www.youtube.com/user/FoxNewsChannel/videos			0	274	274	0.001874	1.0	1.0	1.0	0.0	0	0
https://www.youtube.com/user/TheYoungTurks/videos			0	278	278	0.001874	1.0	1.0	1.0	0.0	1	0

2. What does it mean by weight in Data Table in edges Tab?
   Solution: weight in Edge tab in Data Table means how much weight is assigned to that particular edge. As we can see from Task 6 Question 01 that Green nodes represents the 22 videos shared by both channels (Fox and Young Turk). As we can see that some of the nodes have normal width colored connections (be it blue or purple) while a few connections are considerably wider. These wider connections are due to high weight values.

Task 9

Attach a screenshot of your "Overview" tab in Gephi, which shows your network after you ran the "Force Atlas" Layout algorithm with repulsion strength of 20000.



Modularity Class

🟥	1	(51.5%)
🟩	0	(48.5%)

Force Atlas

Inertia	0.1
Repulsion strengt	20000.0
Attraction strengt	10.0
Maximum displace	10.0
Auto stabilize fun	☑
Autostab Strengtl	80.0
Autostab sensibil	0.2
Gravity	30.0
Attraction Distrib.	□
Adjust by Sizes	□
Speed	1.0

Force Atlas

Nodes: 532
Edges: 552
Directed Graph

Network Overview

Prestige		Run
Average Degree	1.038	Run
Avg. Weighted Degree		Run
Minimum Spanning Tree		Run
Network Diameter	1	Run
Graph Density		Run
Bridging Centrality		Run
HITS		Run
Modularity	0.465	Run
Clustering Coefficient		Run
PageRank		Run
Connected Components		Run
BoundingDiameters		Run
DBSCAN		Run
Girvan-Newman Clustering		Run

THE END