Social Network Project

By (Team 28)

❖ Dataset & Link

Name: Twitch Social Networks
Reference: https://snap.stanford.edu/data/twitch-social-networks.html

❖ Names & Roles

NO	NAME	ID	ACADEMIC YEAR	ROLES
1	قمر مسلم برازي	22011438	المستوى الثاني	- Analysis Cytoscape - Measurement Details - Description of the dataset
2	زياد محمد الخطيب	22010100	المستوى الثاني	- Analysis Igraph - Igraph implementation - Report format
3	معتز محمد احمد محمد الشيمي	22011453	المستوى الثاني	- Analysis Cytoscape - Measurement Details
4	منه اسامه محمد محمد الوزان	22010399	المستوى الثاني	- Analysis Cytoscape - Measurement Details
5	تبيان أشرف عبد الله يوسف	22011497	المستوى الثاني	- Report format - Analysis Cytoscape
6	احمد عماد عبدالفتاح عبدالغني	22010027	المستوى الثاني	- Analysis Cytoscape - Measurement Details
7	محمد محمود نعيم محمد	22010231	المستوى الثاني	- Analysis Igraph - Measurement Details
8	مايا بسطاوي احمد	22011511	المستوى الثاني	- Analysis Cytoscape
9	مايكل مراد نصر شكرى	22010206	المستوى الثاني	- Analysis Cytoscape
10	ياسمين على عبد الرحمن	22010292	المستوى الثاني	- Analysis Igraph - Measurement Details

❖ Description of the dataset

Twitch is an American video live streaming service that focuses on video game live streaming, including broadcasts of E-sports competitions, in addition to offering music broadcasts, creative content, and "In Real Life" streams. Twitch is operated by Twitch Interactive, a subsidiary of Amazon.com, Inc.

The Twitch Social Networks dataset is a collection of social networks of Twitch users who stream in a certain language. The dataset was collected in May 2018 and is available for download from the Stanford Large Network Dataset Collection.

Each node in the network represents a Twitch user, and each edge represents a friendship between two users. The dataset also includes various node features, such as the user's location, streaming habits, and games played and liked.

The Twitch Social Networks dataset can be used for:

- Studying the structure of social networks.
- Developing new methods for node embedding.
- Analyzing the relationship between streaming habits and social interaction.
- Understanding the role of language in online communities.

In our analysis, we focused on a diverse dataset comprising social networks of Twitch users streaming in six different languages: DE, EN, ES, FR, PT, and RU. However, to avoid redundancy, we chose to concentrate our detailed analysis on the English (EN) language dataset. Given that the characteristics and measures were consistent across all languages, our findings from the English dataset are representative of the broader trends observed in the entire multilingual dataset.

❖ **Cytoscape software & Igraph with R used to make some network measurements:**

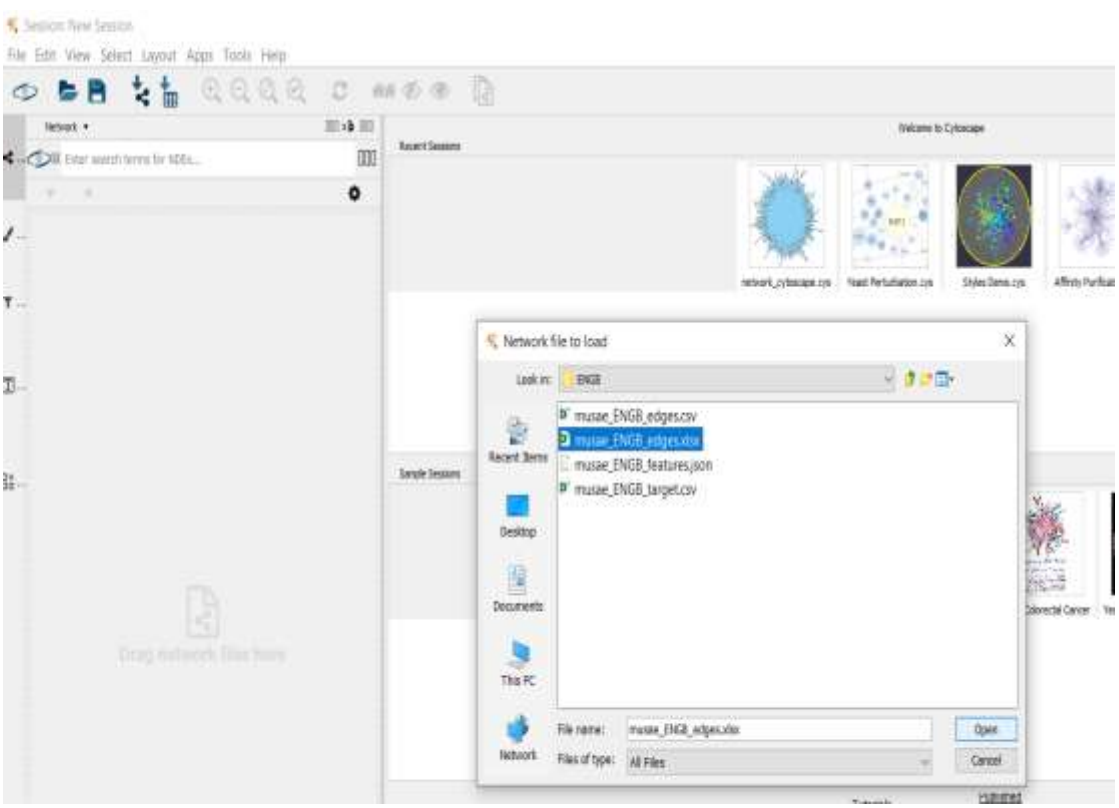
1) What the input to the program will be?

-We import Network from "the musae_ENGB_edges.xlsx" file.

- Igraph import code:

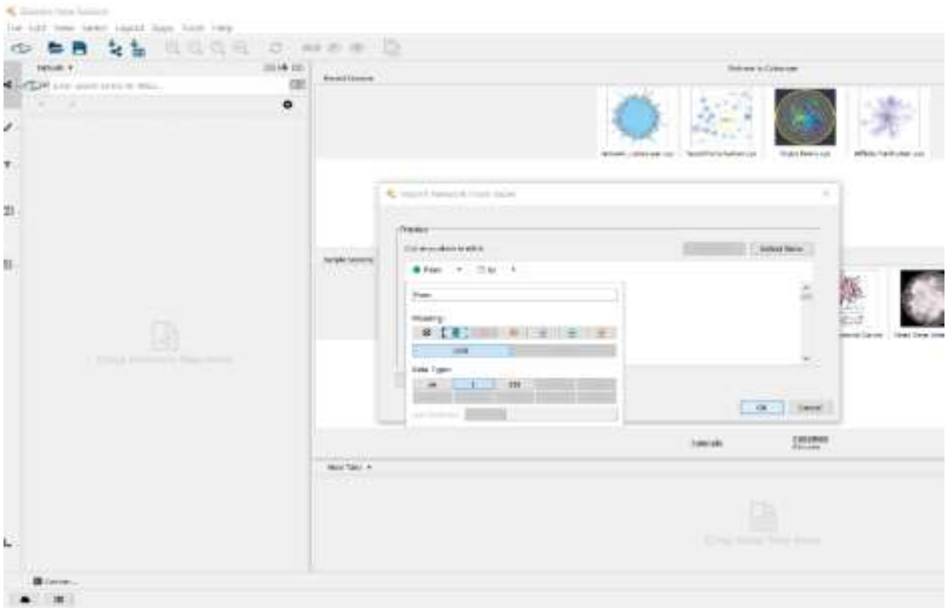
```
require(igraph) #use igraph library to make measurements on graph


Enedges <- read.csv("G:/FCDS/semester 3/Social Network/Final
Project/twitch/ENGB/musae_ENGB_edges.csv")


# create a network graph (connections)

net <- graph.data.frame(Enedges,directed = F)
```

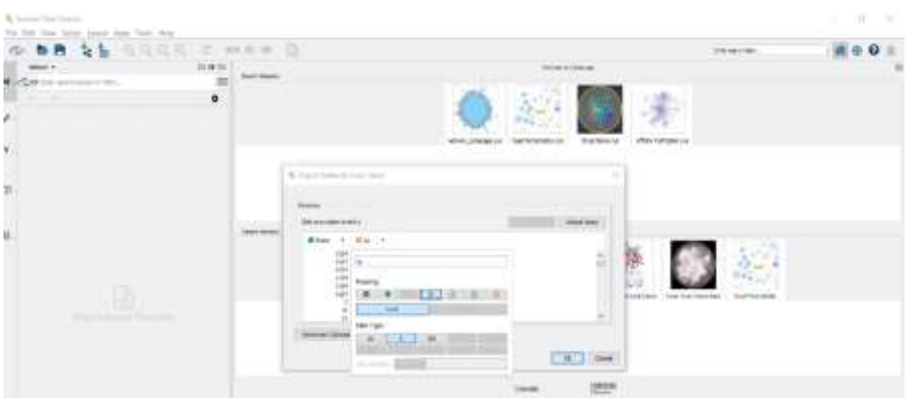
- Cytoscape:



-Then determine the Source Node and the Target Node:

(This is just a normal procedure for using Cytoscape, but will not make any difference in the analysis, because the network is undirected)

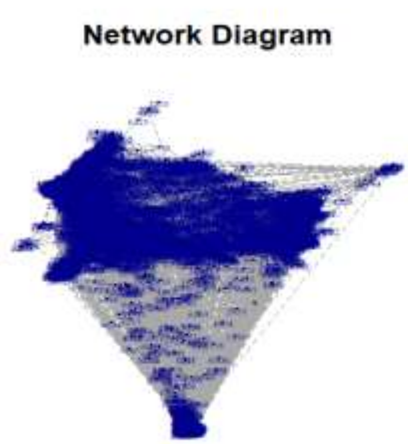




-Now we get the visualization of the Network:

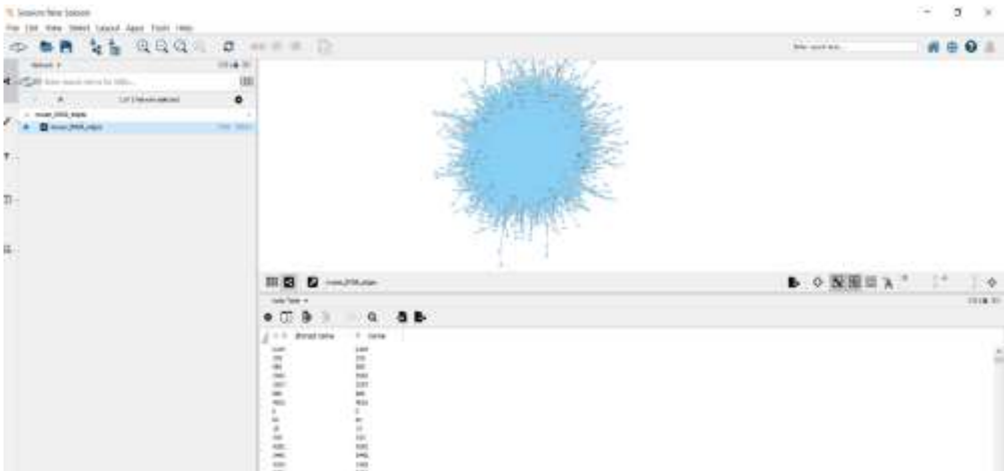
- Igraph code:

```
set.seed(222)
plot(net,
  vertex.color='blue',
  vertex.size= 0.3,
  edge.arrow.size=0.01,
  vertex.label.cex=0.4,
  edge.lty = 2,
  edge.lwd = 0.5,
```



```
main = "Network Diagram")
```

- Cytoscape:

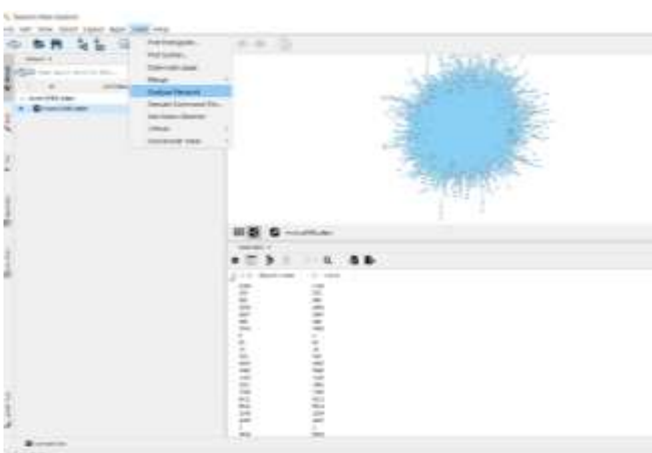


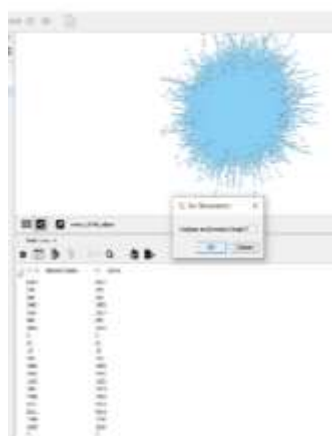
2) What the output from the program will be? (Measurements)

-To benefit from this network, we will conduct a network analysis:

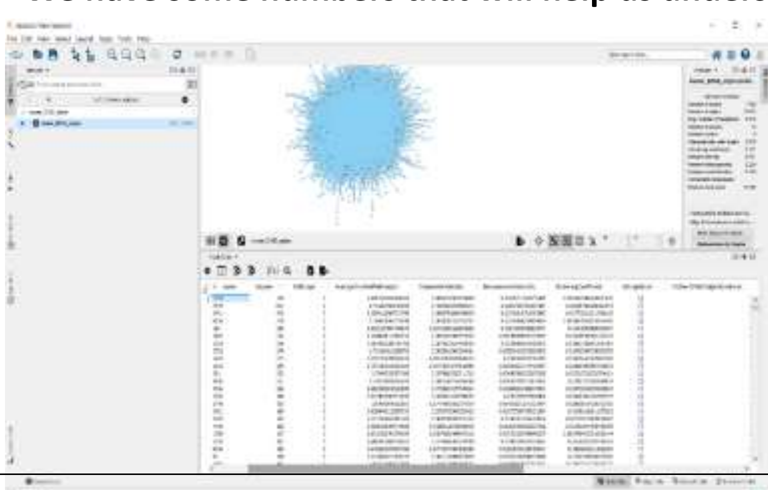
- Cytoscape:

Tools →Analyze the network as undirected.





- We have some numbers that will help us understand the network more:



Description of the columns (measurements) that we will need in our Report:

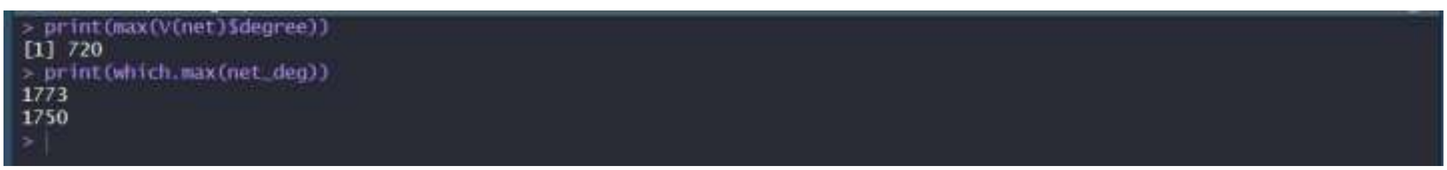
- In our dataset the edge refers to friendship →person 1 follows person 2 & person 2 follows person 1.
- A streamer cannot follow another streamer when the other streamer does not follow him. This is not available in our set because it is undirected.
- Of course, there are no self-loops, because no one can follow themselves, which will be shown in a screenshot.

(Centrality measurements)

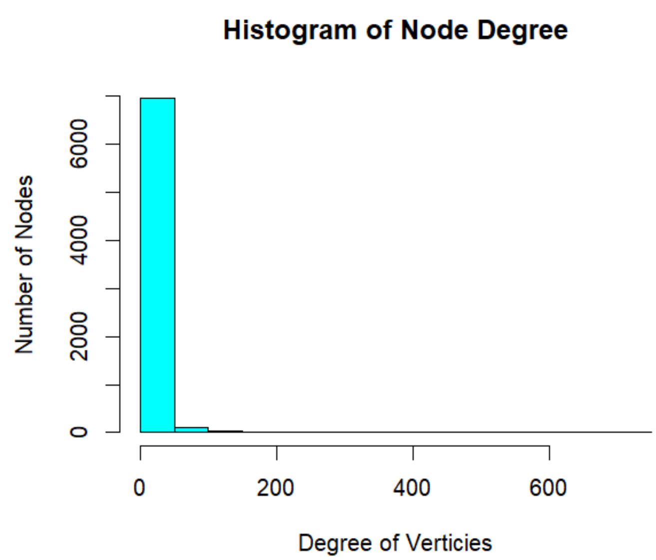
1. Degree Centrality: refers to the number of friends (shows how social a person is).

- Igraph code:

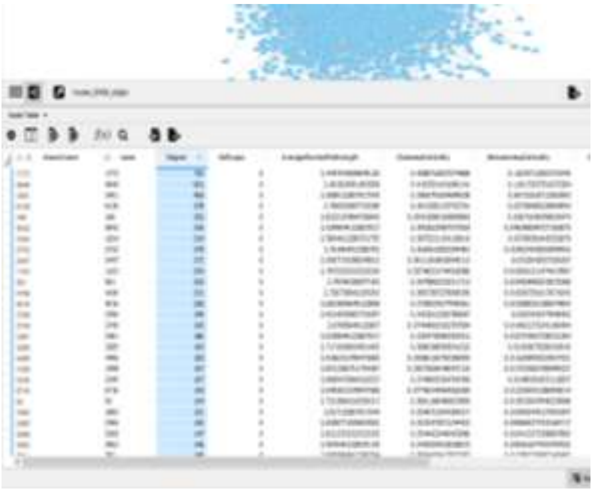
```
# Degree centrality
net_deg <- degree(net)
V(net)$degree <- net_deg
# know the highest degree value
print(max(V(net)$degree) #720

#show vertex & index of the highest degree
print(which.max(net_deg)) # vertex "1773" in index 1750
```




```
#histogram of degrees
hist(V(net)$degree,
     col = 'cyan',
     main = "Histogram of Node Degree",
     ylab = "Number of Nodes",
     xlab = 'Degree of Verticies')
```



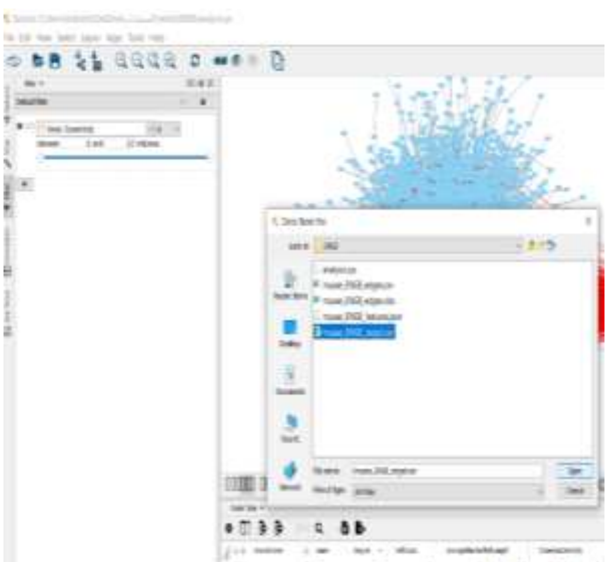
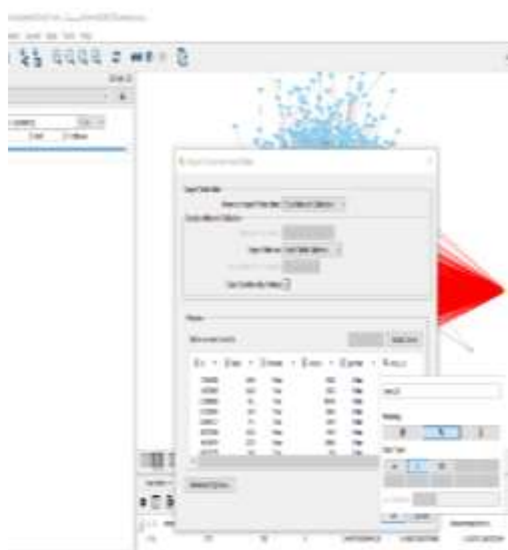
- Cytoscape:



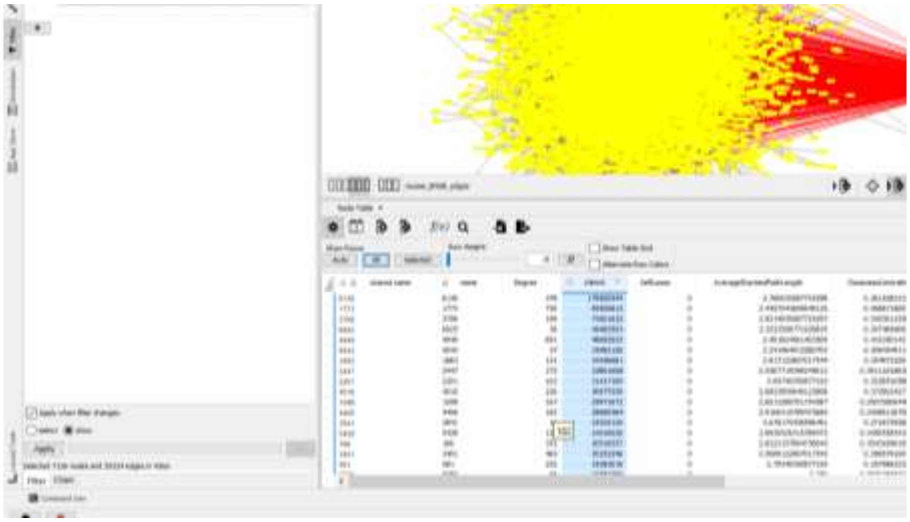

As shown in the screenshots, **the largest degree is 720**, which is for the node with the **name 1773.**

Of course, the smallest degree will be **1** because there are no isolated nodes (i.e. nodes that do not follow anyone and no one follows them).

-To get some additional information about people on Twitch, we will merge the target file with the current file (Edges).

This shows that the person with the most friends on Twitch does not necessarily have the highest views, as the highest degree **was 720 for Node 1773**, and here the highest views are **for Node 6136 with a degree of 378**.



2. Average shortest Path length:

The maximum length of the path in short indicates that the network is well connected, which means that any user can be reached by any other user easily. In contrast, a higher maximum path length indicates that the network is less interconnected, meaning that reaching some users may be difficult.

The maximum reduced path length indicates that information can be transmitted quickly through the network. Conversely, a higher maximum path length indicates that information propagation may be slower.

The average maximum path length can be used to identify influential users in the network. For example, users who have few but strong connections with other users may be the most influential.

- Igraph code:

```
#AVG Shortest Path
average_shortest_path <- average.path.length(net)
print(average_shortest_path)
```

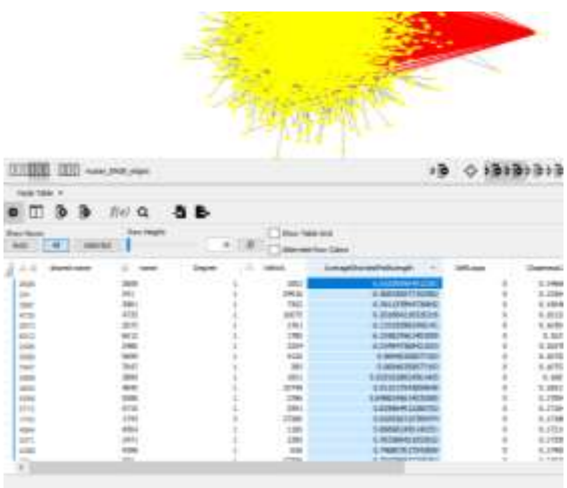
- Cytoscape:

As shown in the picture, **Node 4949** is well connected to the rest of the nodes (because it has the smallest shortest path), this means that we can easily reach **Node 4949** from any other node in the network.

This node may be important as it can access a large number of nodes:

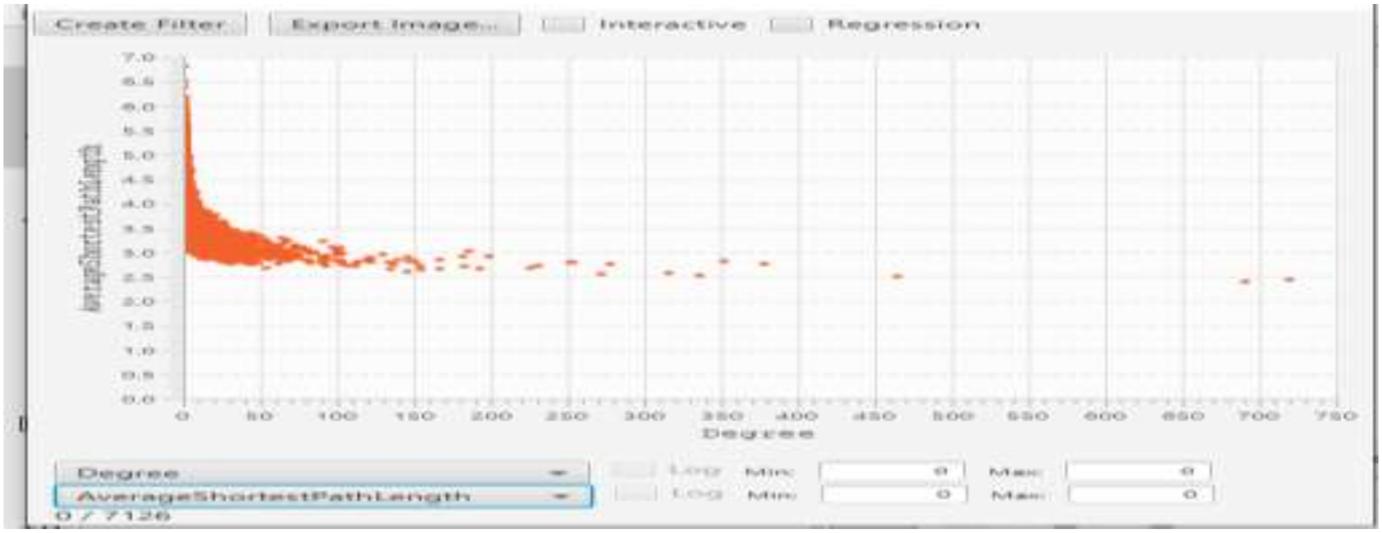
The second largest user in terms of degree (number of friends):

Also, **Node 4949** has a large number of views (Fifth place in terms of views):



The largest shortest path for node 2608 with degree **1** → means that **node 2608** is not important in our network. In addition, it only has one friend.

Relationships between **Degree and Average Shortest Path Length**: there is a negative relationship between Degree and Average Shortest Path Length. This means that nodes with high scores often have a longer Average Shortest Path Length.



We can explain the negativity:

Nodes with high degrees are often the central nodes in the network. These nodes connect many other nodes.

Paths that pass through central nodes typically require a larger number of nodes to get from one node to another.

3. Eigenvector Centrality:

Having more friends does not by itself guarantee that someone is more important. Having more important friends provides a stronger signal. Eigenvector Centrality incorporates the importance of the neighbors.

It measures a node’s influence based on the number of links it has to other nodes in the network.

- Igraph code:

```
net_eig <- evcent(net)$vector

V(net)$Eigen<-net_eig


#know the highest eigenvector value

Print(max(V(net)$Eigen))

#show index & vertex of the highest eigenvector

Print(which.max(net_eig)) #vertex "4949" in index 4417
```

```
> print(max(V(net)$Eigen))
[1] 1
> print(which.max(net_eig))
4949
4417
>
```

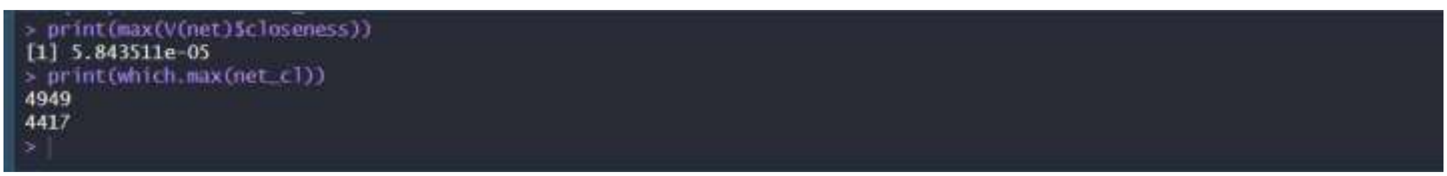
4. Closeness Centrality:

A high value of closeness centrality indicates that a user can reach all other users in the network after only a few hops. This means that this user has a significant influence on the network, as he can easily spread information and influence the behavior of other users.

A high value of closeness centrality also indicates that the user acts as an intermediary between different groups of users. This means that this user plays an important role in connecting the various parts of the network. It also indicates that the user has better access to the information available in the network. This means that this user has a greater chance of knowing what is happening in the network.

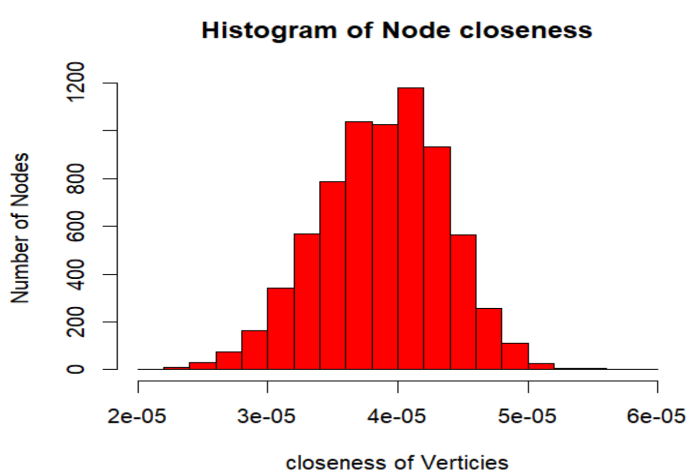
- Igraph code:

```
net_cl<-closeness(net)
V(net)$closeness<-net_cl
print(V(net)$closeness)

#know the highest closeness value

max(V(net)$closeness)
min(V(net)$closeness)

#show index & vertex of the highest closeness
which.max(net_cl) #vertex "4949" in index 4417
```

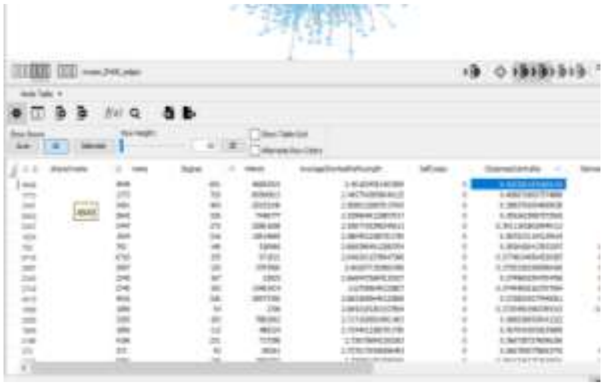


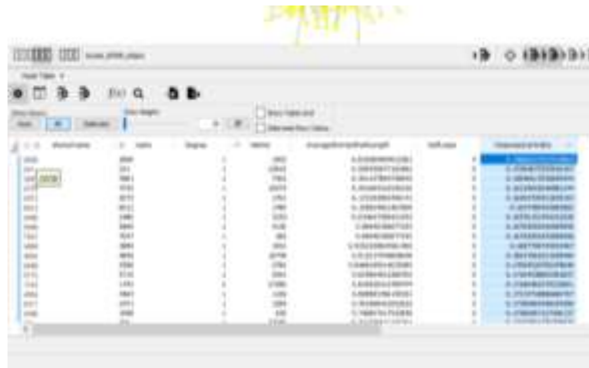
```
> print(max(V(net)$closeness))
[1] 5.843511e-05
> print(which.max(net_cl))
4949
4417
>
```

```
#histogram of closeness
hist(V(net)$closeness,
     col = 'red',
     main = "Histogram of Node closeness",
     ylab = "Number of Nodes",
     xlab = 'closeness of Verticies')
```



- Cytoscape:





As shown, the largest closeness centrality was **Node 4949**. These nodes can influence the trends of the network. This means that the behavior of this user can affect the behavior of other users on the network.

Closeness centrality can be used to identify users who can help promote content on Twitch. For example, marketers can target users who have a high value of closeness centrality to reach a broader audience (like **Node 4949**).

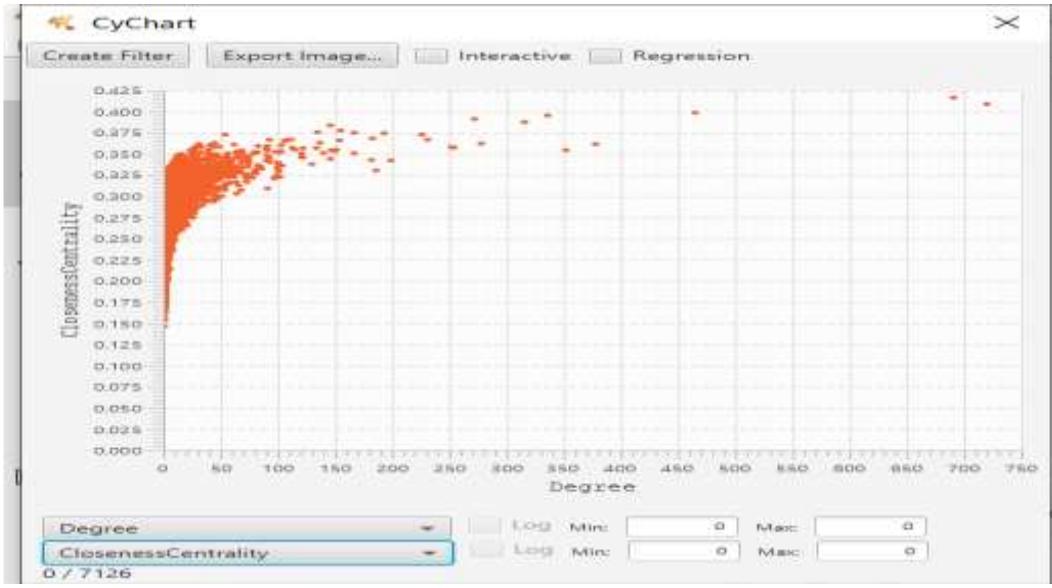
The smallest Closeness Centrality of a particular node means that the node is weakly connected to the rest of the nodes in the network. This means that the node cannot be easily reached from any other node in the network, here we have **Node 2608.**

Node 2608 has the Smallest Closeness Centrality and the largest shortest path with one friend (degree).

-Relationships between **Degree** and **Closeness centrality**: there is a negative relationship between degree and Closeness centrality. This means that nodes with high scores often have low Closeness Centrality.

This means that nodes with high degrees are often less important in Closeness centrality. This is because the central nodes are located far from many other nodes in the network.

However, there is no deterministic relationship between degree and Closeness centrality. Some high-degree nodes have high Closeness centrality as well. This can happen if the central node is connected to many nodes that are in turn connected to many other nodes.



5. Betweenness Centrality:

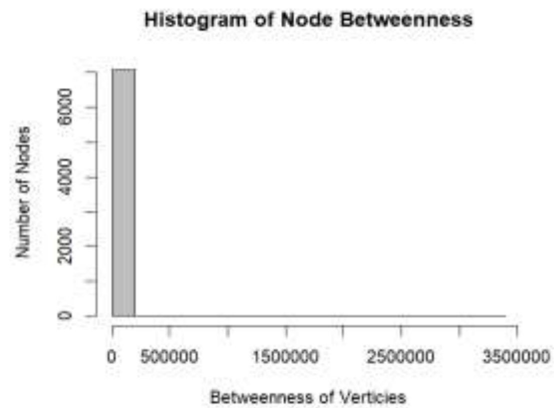
Betweenness Centrality measures how important a particular user is in connecting other users to each other. The more times a given user is on the shortest path between two other users, the greater his betweenness centrality.

A high Betweenness Centrality indicates that the user can help spread information and ideas across the Twitch community. Since the user is on the shortest path between many other users, he can reach a large audience.

- Igraph code:

```
net_bw<-betweenness(net, directed = FALSE)

V(net)$betweenness<-net_bw

print(V(net)$betweenness)

#know the highest betweenness value

max(V(net)$betweenness)

#show index & vertex of the highest betweenness

which.max(net_bw) #vertex "1773" in index 1750
```

```
#histogram of Betweenness
hist(V(net)$betweenness,
     col = 'red',
     main = "Histogram of Node Betweenness",
     ylab = "Number of Nodes",
     xlab = 'Betweenness of Verticies')
```

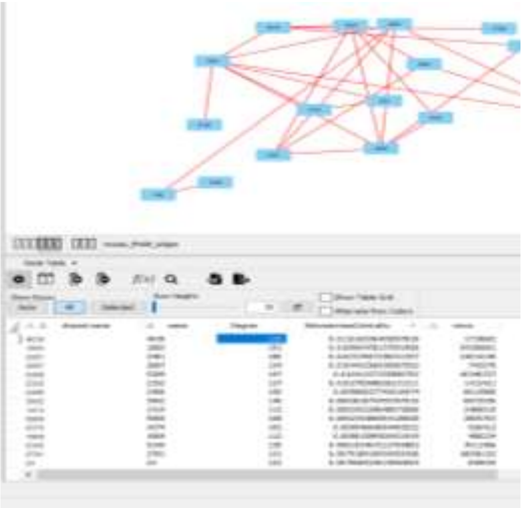


- Cytoscape:

As shown in the picture, the largest node in terms of Betweenness Centrality **is 1773**, which is the largest node in terms of degree, followed by **4949**, which are considered important nodes in the network.

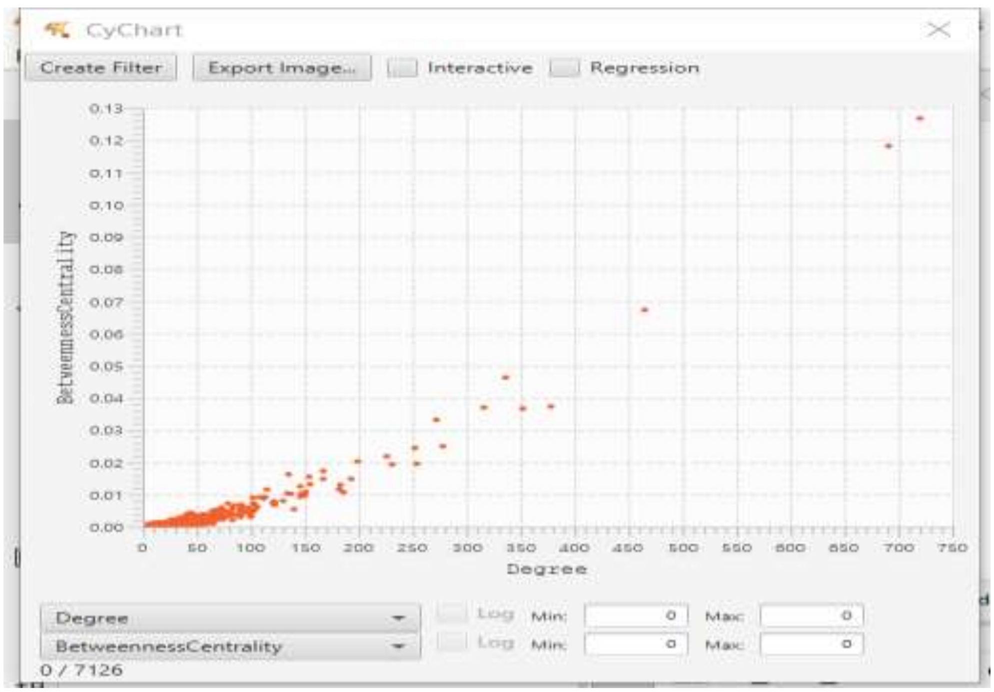
High-degree users often have a high degree of betweenness as well. This is because high-degree users have strong connections with many other users. This makes them more likely to be in the shortest paths between other users.

There are also some cases where high-degree users do not have a high degree of betweenness. For example, users with a high degree may have strong connections with only a few other users. In this case, they are unlikely to be on the shortest paths between other users.



Here: the degree of Node **4639** is lower than those after it, it is greater than them in Betweenness.

-Relationships between **Degrees** and **Betweenness**: there is a positive relationship between Degrees and Betweenness, this means that nodes with high scores often have high Betweenness as well.



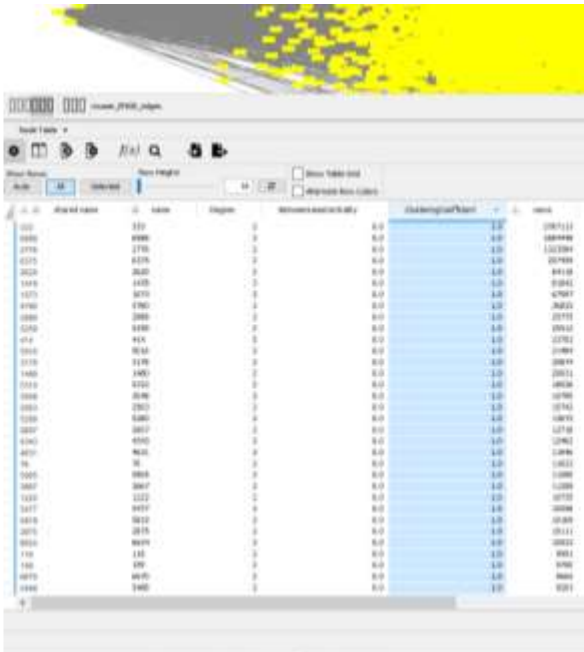
6. Clustering Coefficient:

A high Clustering Coefficient indicates that the user belongs to a community of well-connected users. Community members follow each other and interact with each other regularly.

A high Clustering Coefficient also indicates that information can spread quickly across a given user community because its neighbors follow each other, and information can reach many people quickly.

The Clustering Coefficient can be used in conjunction with other network metrics, such as betweenness centrality, to discover communities within the Twitch network.

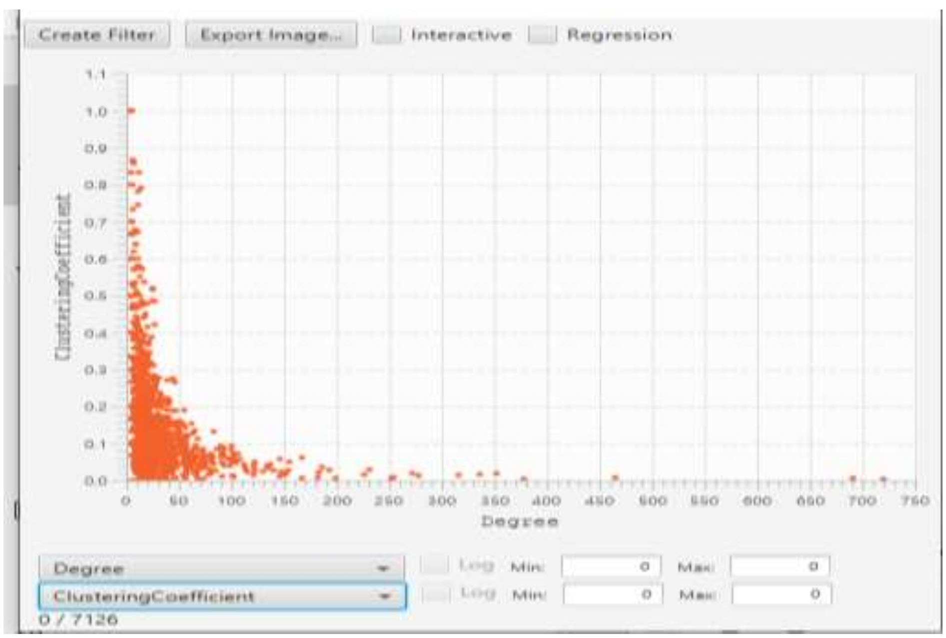
- Cytoscape:



We can explain the positivity:

-Clustering coefficient distribution: most of the nodes in the network have a Clustering coefficient of **less than 0.2**. This means that most nodes in the network are not connected to many other nodes that are connected to each other. There are a few nodes with a high Clustering coefficient, which indicates that there are some nodes that tend to be connected to other nodes that are connected to each other.

-Relationships between **Degree** and **Clustering Coefficient**: there is a positive relationship between Degree and Clustering Coefficient. This means that nodes with high scores often have a high Clustering coefficient as well.



We can explain the positivity:

Nodes with high degrees are often the central nodes in the network. These nodes connect many other nodes.

Central nodes are likely to be connected to other nodes that are connected. This is because central nodes are located in the middle of the network, where other nodes have a greater chance of being connected.

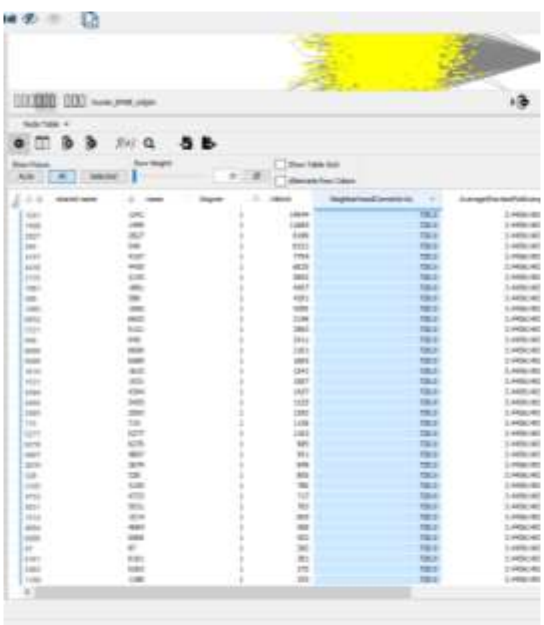
7. Neighborhood Connectivity:

Neighborhood Connectivity is a measure of how well the neighbors of a node are connected to each other. When a node has the highest Neighborhood Connectivity, it often implies that the node is playing a significant role in facilitating communication or interactions between its neighboring nodes.

If we find that users with high Neighborhood Connectivity are the users with the largest number of followers, this means that they are the users who interact most with the Twitch community.

We can also use Neighborhood Connectivity to compare users from different groups. For example, we can compare Neighborhood Connectivity for users who stream different games. If we find that users who stream a certain game have higher Neighborhood Connectivity than users who stream another game, it means that users who stream the first game are more engaged with each other.

- Cytoscape:



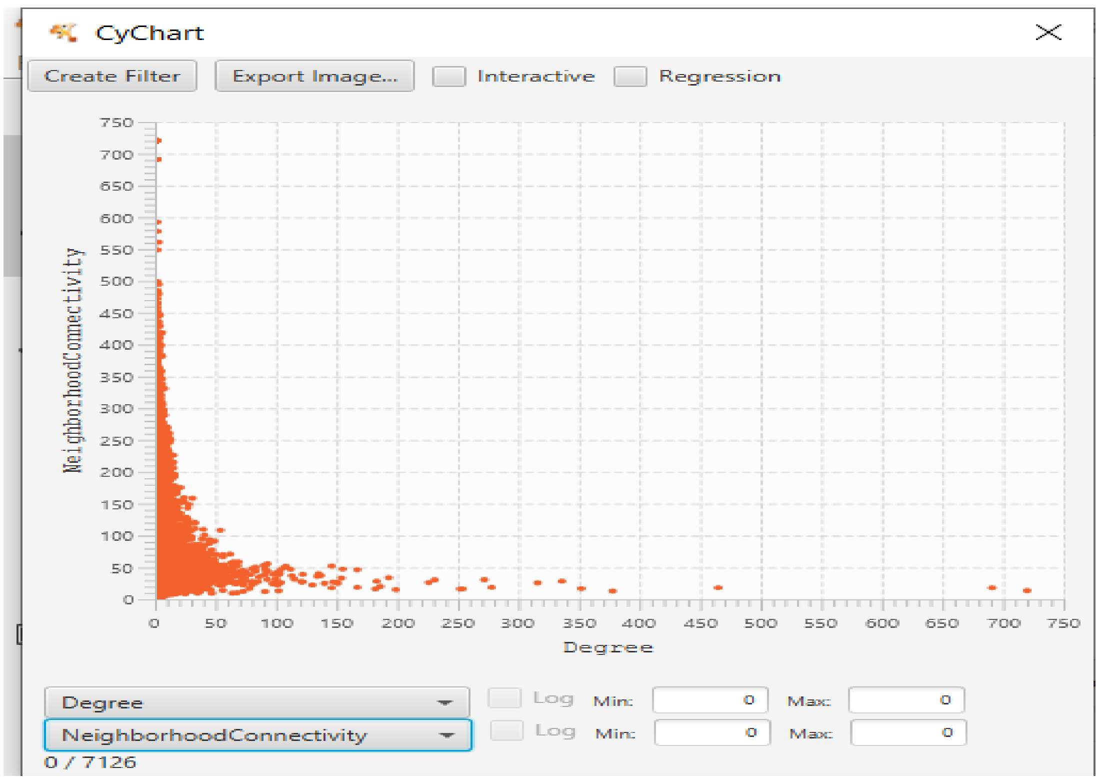
-All Nodes that have the max Neighborhood Connectivity have degree **1**

Single-degree users (users with degree 1) are more likely to be active on Twitch than users with a higher degree. This is because they only follow one user, which means they are more likely to engage with content created by that user.

Single-degree users also are more likely to spread content across Twitch than users with higher degrees. This is because they only have a direct path to one user, which means they are more likely to post the content they view.

-**Neighborhood Connectivity Distribution**: most of the nodes in the network have Neighborhood Connectivity less than 0.5. This means that most nodes in the network are not connected to many other nodes in the node's neighborhood. There are a few nodes with high Neighborhood Connectivity, which indicates that there are some nodes that tend to connect to other nodes in the node's neighborhood.

-**Relationships between Degree and Neighborhood connectivity**: there is a positive relationship between Degree and Neighborhood Connectivity. This means that nodes with high scores often have high Neighborhood Connectivity as well.



8. Adjacency Matrix:

- Igraph code:

```
adj_matrix <-get.adjacency(net)

print(adj_matrix)
```

For more details about R code: https://github.com/ZizoElkhateeb/R_code