

# Network Analysis of Reddit for Egyptians



## Prepared By:

Karim Ahmed Farhat	120200216
Mina Thabet Telmeez	120200163
Adham Khalid Sayed	120190133
Mostafa Mohamed Sayed	120200104
Adham Mohamed Aboeldahab	120200155
Samy Mostafa Samy	120190076
Omar Mamdouh Abdalgayed	120200082

# Agenda



- 01** | **Introduction**
- 02** | **Prepare and Collect our Dataset**
- 03** | **Build the Network**
- 04** | **Analysis which we applied on our Network**
- 05** | **Conclusion**

# 1. Introduction



- What is the Goal of our project ?
  - Our goal is to analyse the network structure and dynamics of **Reddit** users who follow **Egyptian subreddits**, where:
    - Each **node** represents a **User**.
    - Each **edge** represents a **common Subreddit** between two Users.

## 2. Prepare and Collect our Dataset



- The dataset was collected by scraping Reddit using the PRAW library.
- This phase passed through three sub-phases:
  1. First, we collect Egyptian subreddits (105 subreddits).
  2. Then, we collect usernames in each subreddit and put them in separate files (105 files).
  3. Finally, we built our **dataset** which consists of **23185 rows** and **24 columns**:
    - The **first column** contains the usernames of **23185** Reddit users who follow at least one Egyptian subreddit. The other **23 columns** contain the names of the subreddits that each user follows.

## 3. Build the Network



3.1. What we do to build the Network?

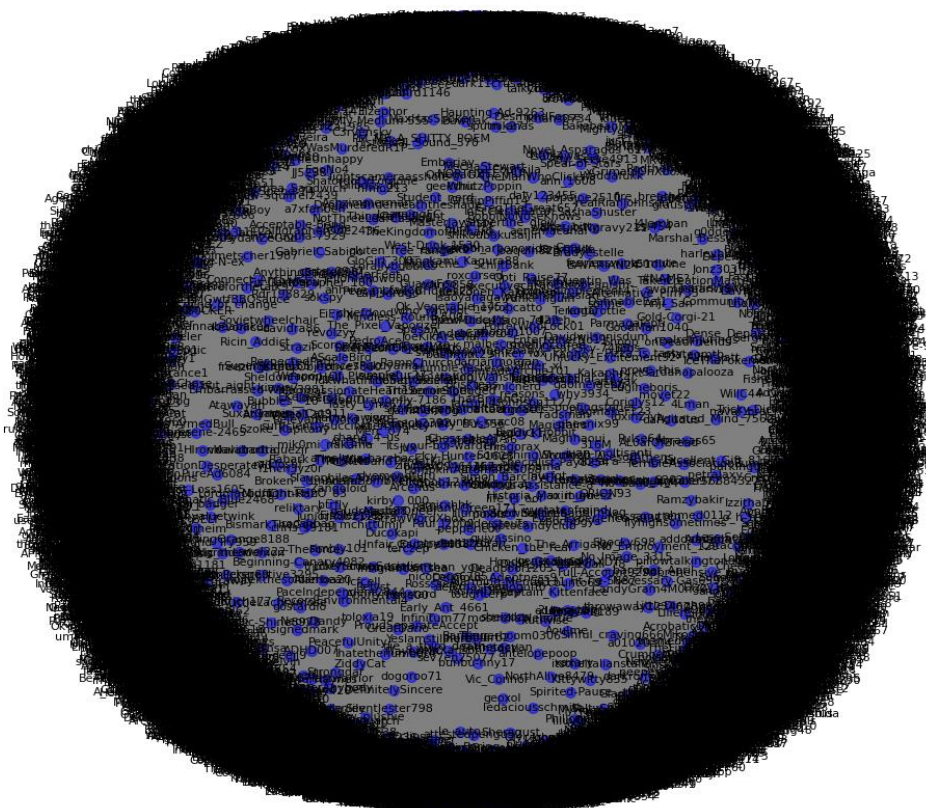
3.2. Display the Network.

## 3.1. What we do to build the Network?

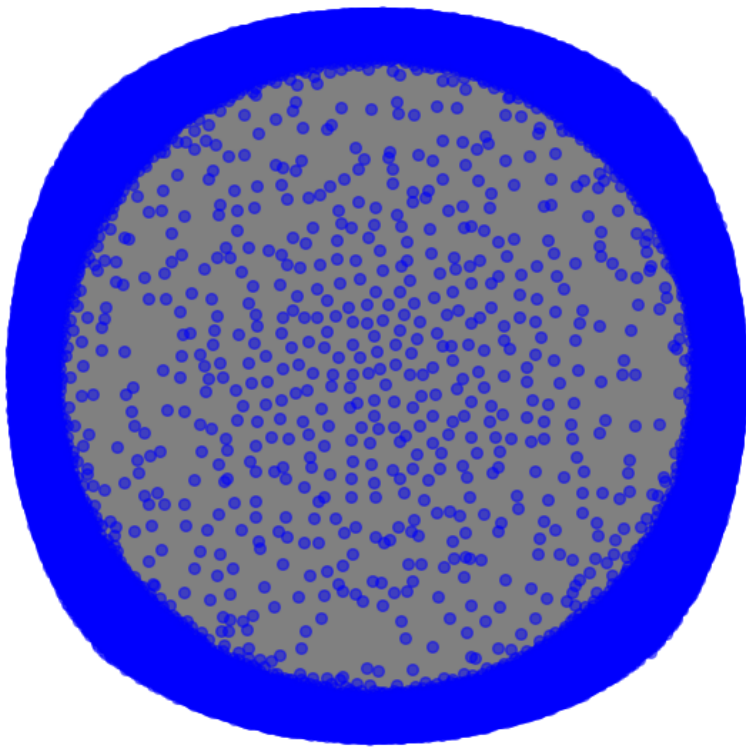


- First, we clearly define our nodes and edges.
- Then, we used these libraries to build the Network:
  - ✓ pandas, networkx, and matplotlib.pyplot
- Finally, we generate our network/graph in 2 formats:
  - ✓ \*.png and \*.graphml formats.
- We observed that:
  - ✓ Number of nodes: **23173**
  - ✓ Number of edges: **6877773**

# 3.2. Display the Network



With Labels



Without Labels

## 4. Analysis which we applied on our Network



4.1. Degree analysis.

4.2. Degree Distribution Analysis.

4.3. Clustering Coefficients.

4.4. Network Type.

4.5. Centrality Analysis.

4.6. Community Discovery.

4.7. Dynamic Community Discovery.


4.8. Connected Component Analysis.

4.9. Density Analysis.

4.10. Path Analysis

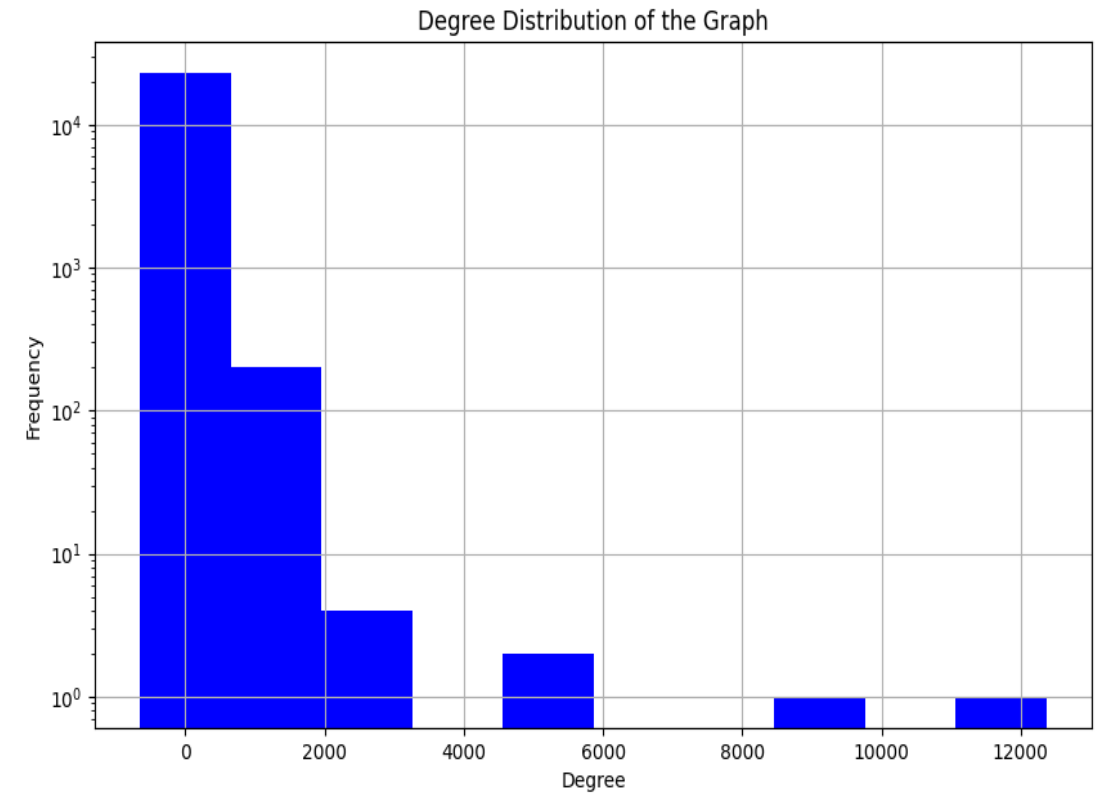


## 4.1. Degree analysis

- 
- The degree of a node in a graph is the number of edges connected to that node.
  - No specific algorithm just a basic computation of the degree and average degree of a graph
  - We observed **4 things**:
    - Minimum degree : 0
    - Maximum degree : 13019
    - Average degree : 593.602.
    - Total number of degrees : 13755546.
  - We will use this data to determine Network Type.

## 4.2. Degree Distribution Analysis

- We use numpy, networkx, matplotlib, scipy libraries
- **Frequency** is the number of nodes that have a certain degree
- We calculated beside average degree, min degree, max degree : **standard deviation, skew degree, kurt degree**



## 4.2. Degree Distribution Analysis

If skew degree  $> 0$ :

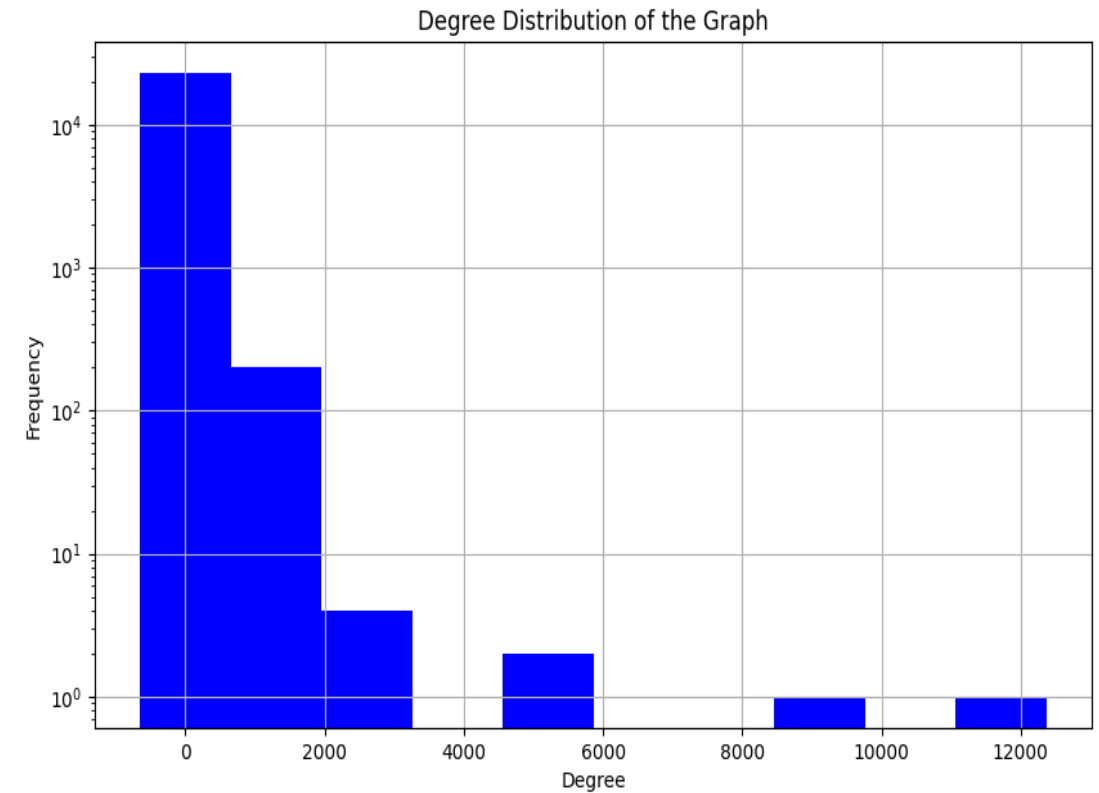
The distribution is **positively skewed** : meaning that most nodes have low degrees, and a few nodes have high degrees

Else If skew degree  $< 0$ :

The distribution is **negatively skewed**: meaning that most nodes have high degrees, and a few nodes have low degrees

Else :

The distribution is **symmetric**, meaning that nodes have similar degrees around the average



## 4.2. Degree Distribution Analysis

If kurt degree  $> 0$ :

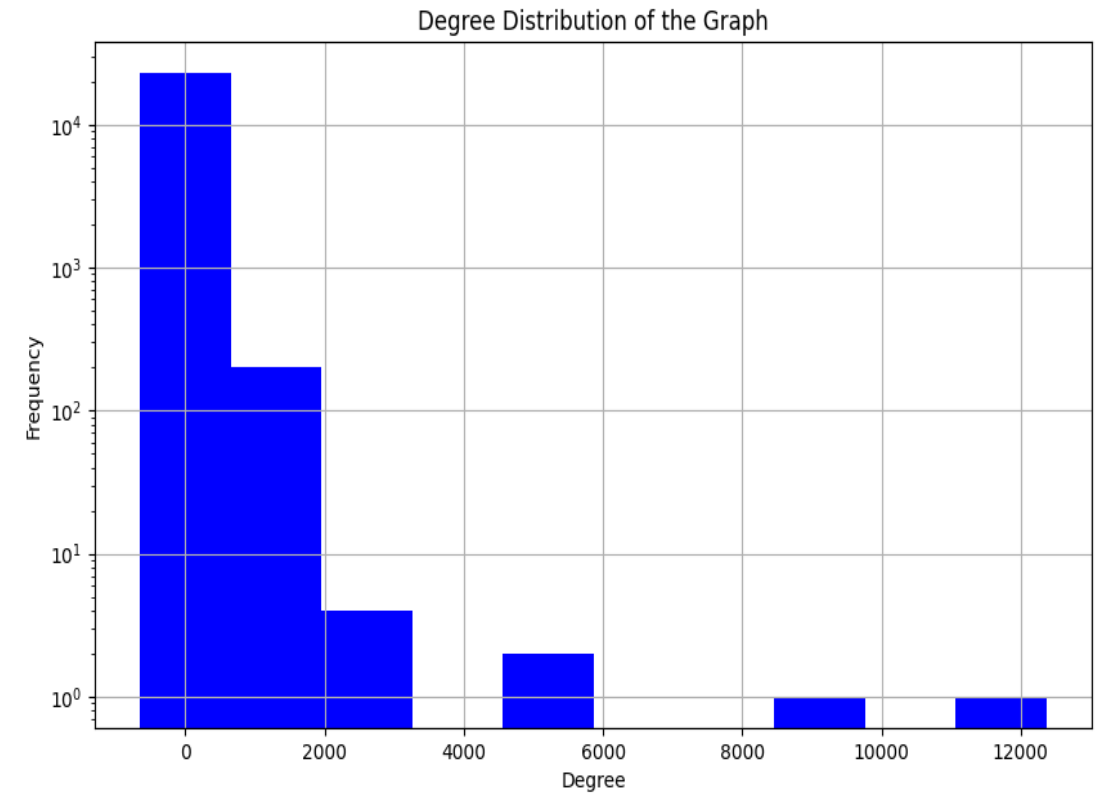
The distribution is **leptokurtic**, meaning that it has a sharp peak and heavy tails.

Else if kurt degree  $< 0$ :

The distribution is **platykurtic**, meaning that it has a flat peak and light tails.

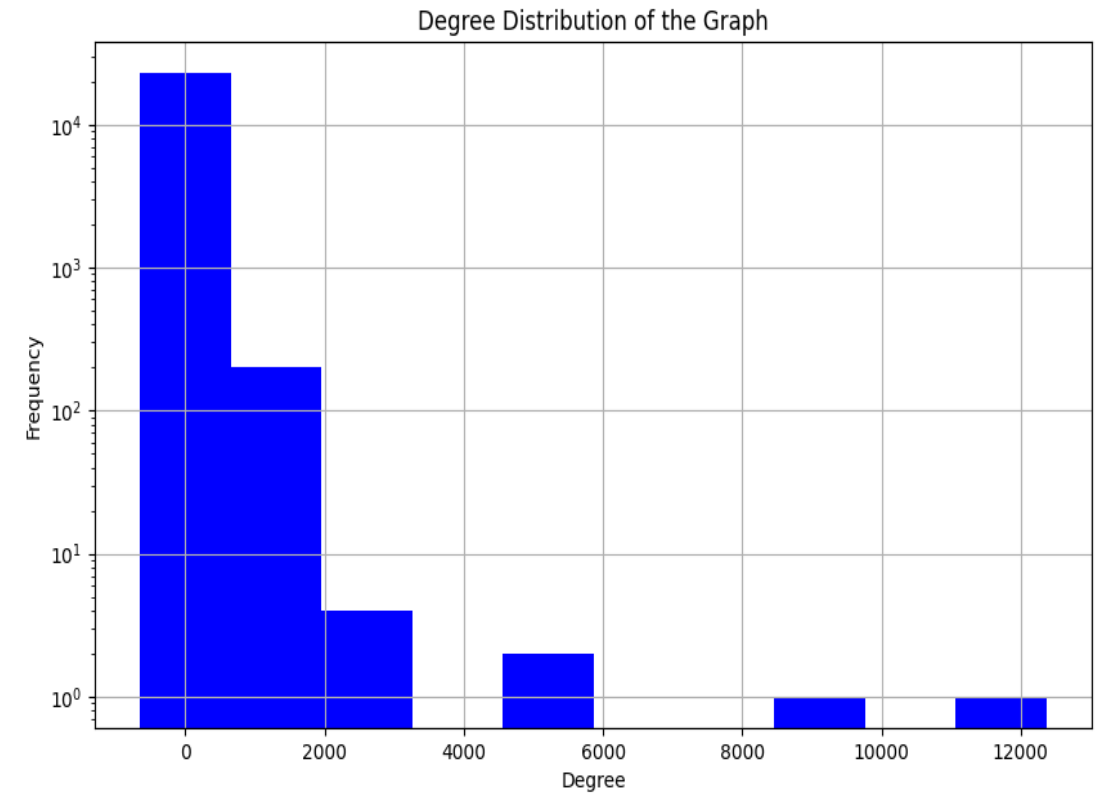
Else:

The distribution is **mesokurtic**, meaning that it has a normal peak and tails.



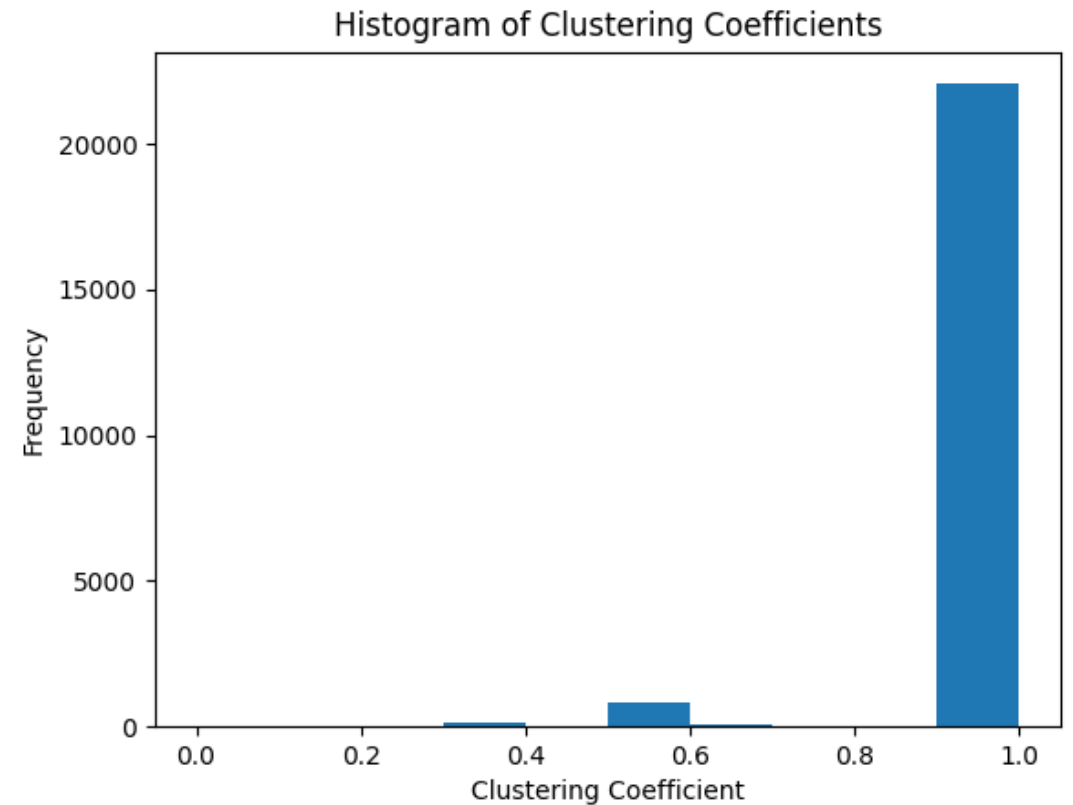
## 4.2. Degree Distribution Analysis

- We observed that :
  1. The degree distribution of the graph ranging **from 0 to 13019**.
  2. The average degree is **593.602**, with a standard deviation of **257.982**.
  3. The distribution is **positively skewed**, meaning that **most nodes** have low degrees, and a few nodes have high degrees.
  4. The distribution is **leptokurtic**, meaning that it has a **sharp peak** and **heavy tails**.



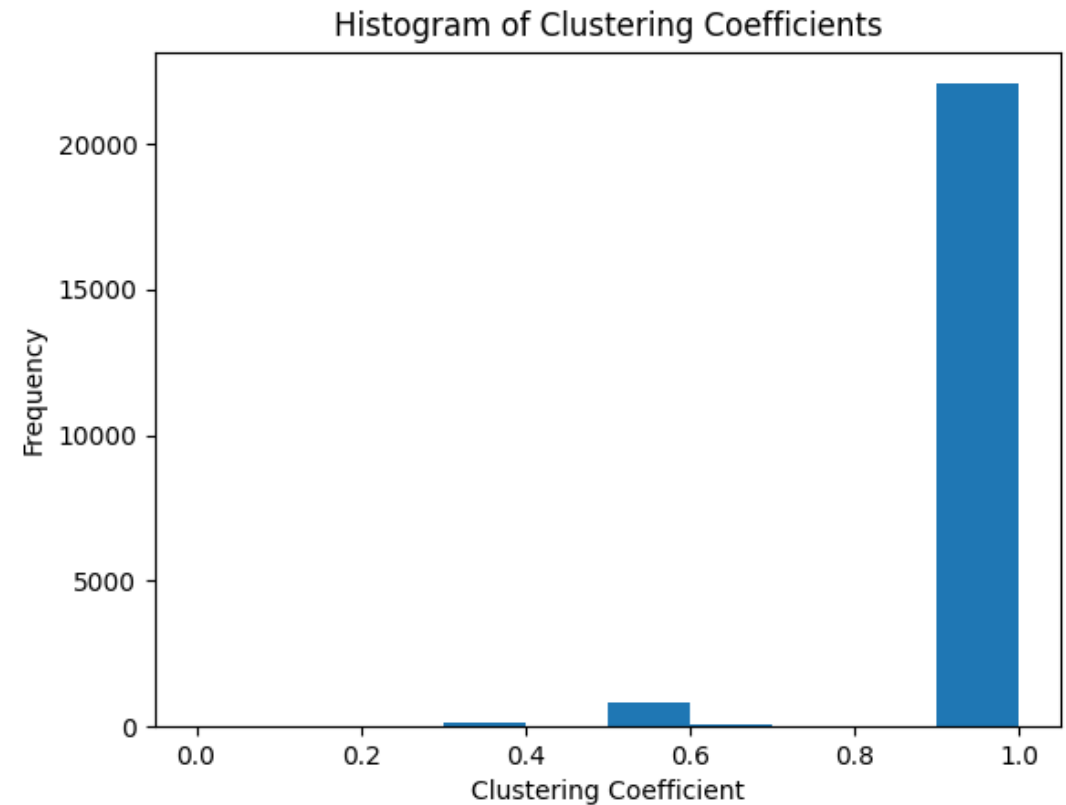
## 4.3. Clustering Coefficients

- The clustering coefficient is a measure of how densely connected a node's neighbors are in a graph
- It is calculated by dividing the number of triangles that a node forms with its neighbors by the maximum possible number of triangles that it could form.



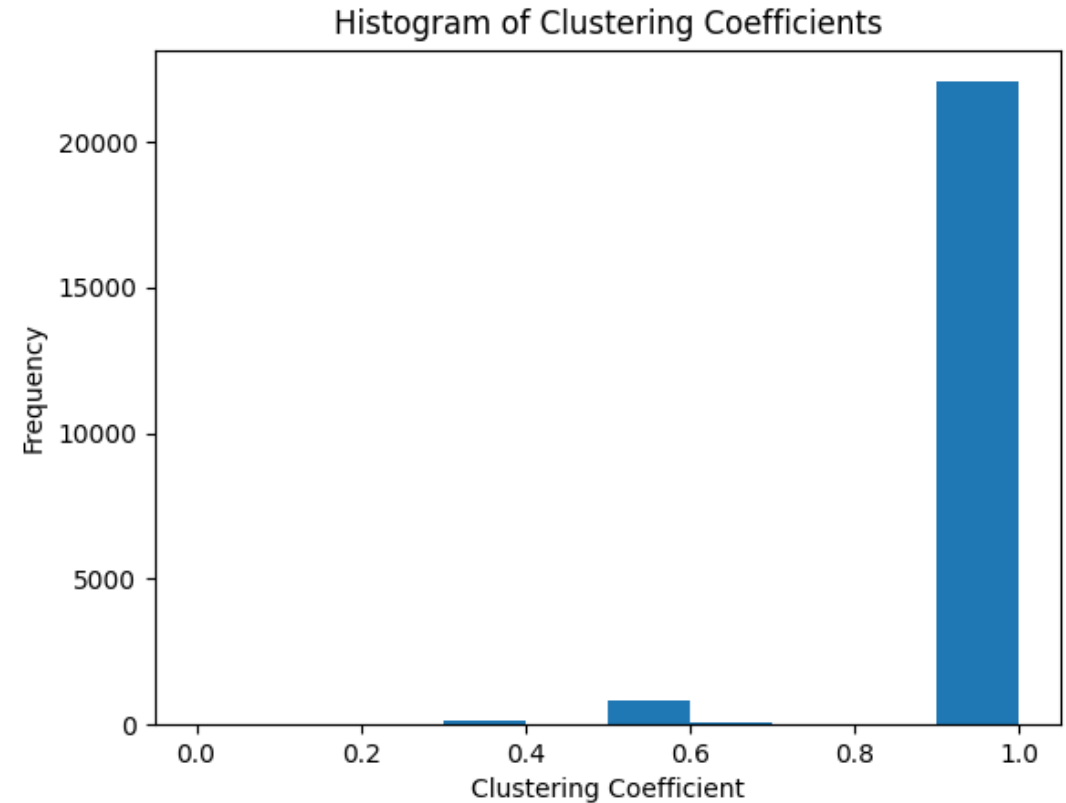
## 4.3. Clustering Coefficients

- The clustering coefficient ranges from 0 to 1, where 0 means that none of the node's neighbors are connected to each other and 1 means that all of the node's neighbors are connected to each other
- The average clustering coefficient ranges from 0 to 1, where 0 means that there is no clustering at all and 1 means that there is perfect clustering.
- We will use average clustering with average degree to determine Network Type.



## 4.3. Clustering Coefficients

- We calculated the average clustering coefficient : 0.976
- We observed that:
  - There is very high clustering in the graph.
  - Most of the nodes in the graph form triangles with their neighbors.
  - There are many cliques or communities in the graph
  - The graph is not very random or sparse, and that there are many common friends among the nodes.






## 4.4. Network Type



- When we search on this topic, we found that there are 4 Types:
  1. **Small World Network:** **High** clustering and **high** connectivity. Tightly-knit clusters of nodes separated by relatively few long-range connections. Found in **social networks**.
  2. **Random Network:** **High** connectivity but **low** clustering. Nodes are connected randomly without any preference for forming clusters or communities. Used as **null models in network analysis**.
  3. **Clustered Network:** **Low** connectivity but **high** clustering. Nodes form tight-knit groups or communities that are relatively isolated from each other. Found in **biological or ecological systems**.
  4. **Sparse Network:** **Low** connectivity and **low** clustering. Nodes are weakly connected and there are few or no tight-knit clusters or communities. Found in **systems where interactions between nodes are rare or unimportant**.

## 4.4. Network Type

- 
- We observed from average Clustering Coefficients and Average degree that :
    - Our Network is **High** clustering and **High** connectivity
    - Which means:
      - The Type of our Network is: **Small World Network.**

## 4.5. Centrality Analysis



- We found that there are 4 properties of Centrality:
  - **Degree Centrality** measures the number of connections a node has in a network
  - **Betweenness Centrality** measures the extent to which a node lies on the shortest paths between other nodes
  - **Closeness Centrality** measures how quickly a node can reach all other nodes in a network
  - **Eigenvector Centrality** measures the influence of a node based on the influence of its neighbors.
- Each of these measures can be used to identify important nodes in a network and can give insight into the structure and connectivity of the overall network.

## 4.5. Centrality Analysis

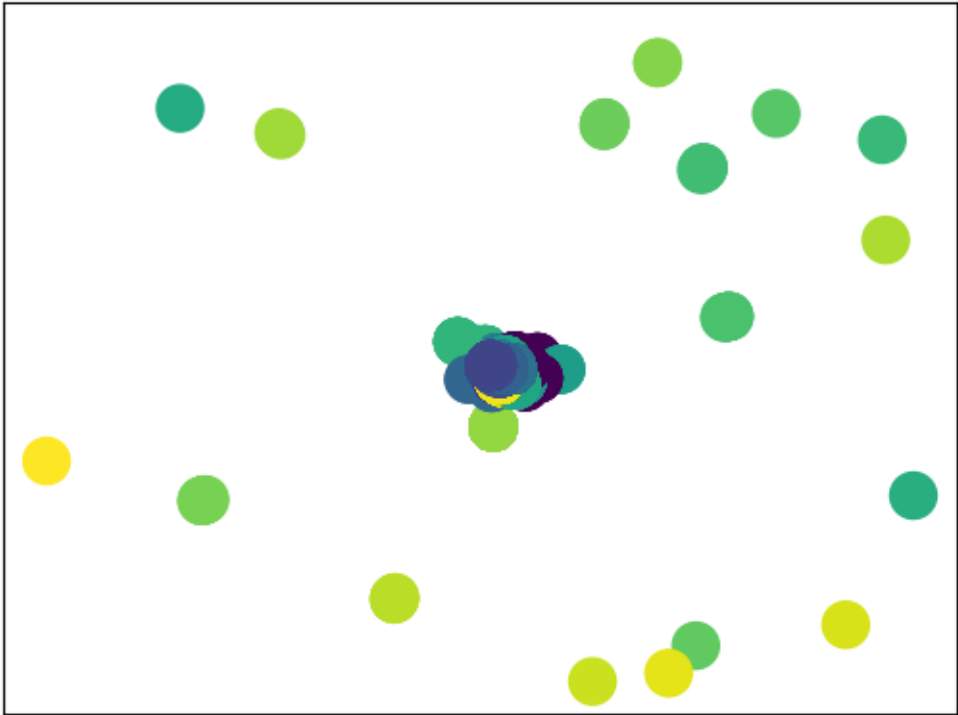


- We observed that there is a user who follows 23 Subreddits out of 105:
  - His **Degree Centrality** : 0.561842.
  - His **Betweenness Centrality** : 0.269488.
  - His **Closeness Centrality** : 0.692209.
  - His **Eigenvector Centrality** : 0.036501.
  - Name of the user: "Wil".
  - Number of the row in our dataset: 19206.
- We found that this user is highly **influential** and **well-connected** in the network.
- This user will affect on the results of our Connected Components.

# 4.6. Community Discovery


Static Community	Number of elements
0	217
1	758
2	512
3	717
4	916
5	800
6	617
7	416
8	785
9	625
10	2228
11	777
12	1316
13	638
14	491
15	728
16	1267
17	399
18	732
19	766
20	809
21	952
22	751
23	746
24	522

Static Community	Number of elements
25	559
26	709
27	368
28	751
29	509
30	1
31	1
32	24
33	1
34	4
35	9
36	1
37	1
38	5
39	5
40	2
41	88
42	3
43	1
44	5
45	1
46	1
47	1
48	637
49	1



There are 50 (Static) Community.

## 4.6. Community Discovery

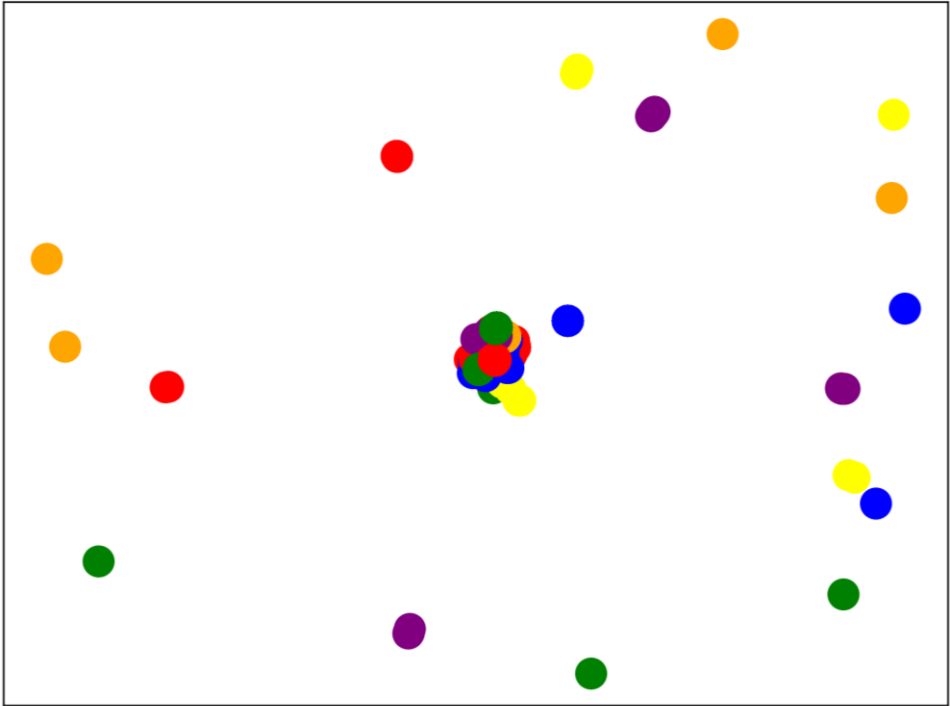
- 
- **The Louvain algorithm** is a widely used **community detection algorithm** that aims to optimize the modularity measure to identify communities in a network.
  - It is an **iterative algorithm** that partitions the network into communities by **maximizing the modularity**, which **quantifies the quality of the division of a network** into communities

# 4.7. Dynamic Community Discovery




Dynamic Community	Number of Elements
1	13468
2	8189
3	368
4	3
5	1
6	66
7	28
8	2
9	1
10	24
11	1
12	4
13	9
14	1
15	1
16	5
17	208
18	5
19	2
20	88

Dynamic Community	Number of Elements
21	16
22	3
23	1
24	5
25	5
26	3
27	1
28	9
29	1
30	2
31	15
32	2
33	9
34	1
35	153
36	243
37	12
38	1
39	217



There are 39 (Dynamic) Community.

## 4.7. Dynamic Community Discovery

- 
- The Label Propagation algorithm is an **efficient algorithm** for **community detection** in graphs.
  - It is a **semi-supervised** algorithm that assigns labels (community assignments) to the nodes based on the network structure and the labels of their neighbors.
  - The algorithm propagates labels through the network until a stable state is reached, where each node is assigned a label that maximizes its agreement with its neighbors.



## 4.8. Connected Component Analysis

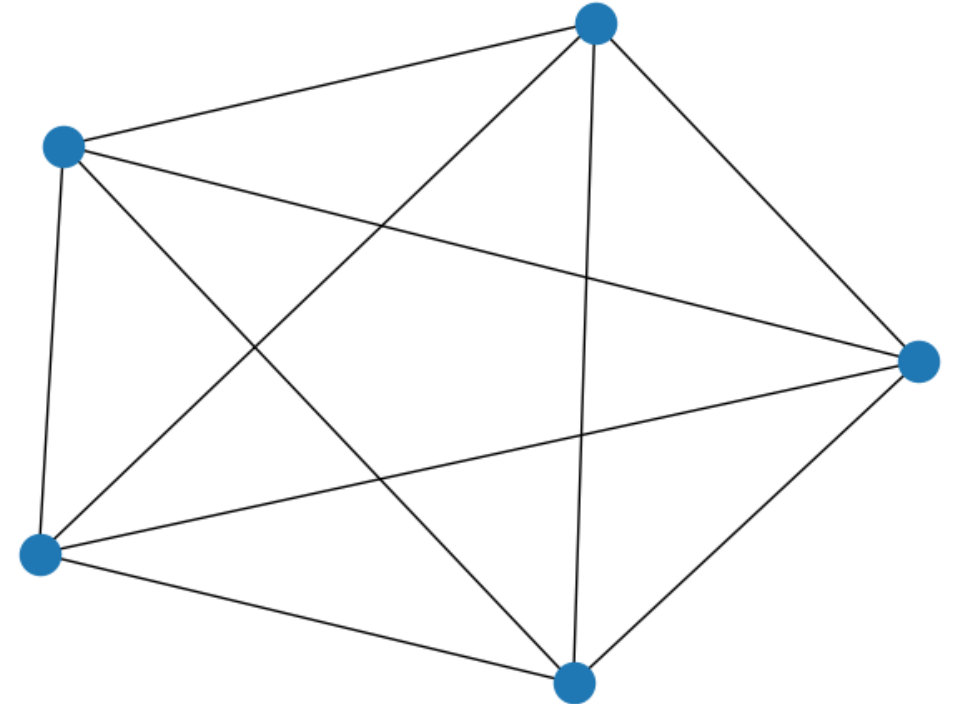
- Our graph generates 19 Connected Components , which means that our graph is divided into 19 distinct groups of nodes, where each group forms a connected component. These components are separate and do not have direct connections between them.
- The maximum number of elements in a component is 23042, and it belongs to Component 1, and they are not in same subreddit.
- There are 10 components with only one element and 9 components have more than one element
- This observation will affect on our Density ratio

Component	Size
Component 1	23042
Component 2	1
Component 3	1
Component 4	1
Component 5	4
Component 6	9
Component 7	1
Component 8	1
Component 9	5
Component 10	5
Component 11	2
Component 12	88
Component 13	3
Component 14	1
Component 15	5
Component 16	1
Component 17	1
Component 18	1
Component 19	1

## 4.8. Connected Component Analysis



- Example of Connected component:
  - Component **10** represent network of subreddit called **“EgyptianHistoryMemes”** has **5 members.**




## 4.9. Density Analysis



- Density is a measure of how connected a graph is, calculated as the ratio of the number of edges to the number of possible edges in a graph.
- We observed that the graph is **relatively sparse**, and this makes sense because no. of not connected components in the graph are **more than** connected components.
- Density = 
$$\frac{2 * \text{no. of edges}}{\text{no. of nodes} * (\text{no. of nodes} - 1)}$$
- Number of Nodes: 23173
- Number of Edges: 6877773
- The number of possible edges in a graph : 268482378
- Density Ratio: 0.025617

## 4.10. Path Analysis

- 
- During our research of this analysis, we found several techniques of path analysis, but we found that among these techniques the most important and popular one in all of them is short path analysis.
  - So, we choose the shortest path analysis and by using Connected Components output file and randomly choose two different name from each component.
  - We ignored the 10 components which have only one element and focused on 9 other components which have more than one component.

## 4.10. Path Analysis



Component	Source Node	Target Node	Shortest Path	Shortest Path Length	Subreddit Name	Number of elements in the Component
1	gqn	cLoUt_diDDit	['gqn', 'ency', 'cLoUt_diDDit']	2	reddit.com , assassinscreed	23042
5	kendralinnette	odedi1	['kendralinnette', 'odedi1']	1	EgyptianFood	4
6	happyboy13	ProgaPanda_	['happyboy13', 'ProgaPanda_']	1	EgyptianGamingSociety	9
9	fatma_ezzouhry	AEssam	['fatma_ezzouhry', 'AEssam']	1	EgyptianGeeks	5
10	IacobusCaesar	Memetaro_Kujo	['IacobusCaesar', 'Memetaro_Kujo']	1	EgyptianHistoryMemes	5
11	InTheKurry	RealHistoryMashup	['InTheKurry', 'RealHistoryMashup']	1	egyptianlanguage	2
12	dishonoredgraves	LimeAndTacos	['dishonoredgraves', 'LimeAndTacos']	1	egyptianmau	88
13	CASCADE_999	Trainer_Opposite	['CASCADE_9W', 'Trainer_Opposite']	1	EgyptianMentalHealth	3
15	muffled_savior	3pr7	['muffled_savior', '3pr7']	1	EgyptianShitposting	5

- We observed that all Connected Components represent the same Subreddit except **Component 1** represent more than Subreddit.
- We also observed any two nodes in the same component its shortest path length will be **equal 1**.
- **Except** component 1 which may **be more than 1**, due to **different Subreddits**

## 5. Conclusion



- Number of nodes: 23173.
- Number of edges: 6877773.
- Average degree : 593.602.
- Total number of degrees : 13755546.
- The distribution is **positively skewed**.
- The distribution is **leptokurtic**.
- The average clustering coefficient : 0.976 (i.e., very high clustering and Most of the nodes in the graph form triangles with their neighbors)
- Network Type: **Small World Network**.
- There are 50 (Static Community).
- There are 39 (Dynamic Community).
- There are 10 components with only one element and 9 components have more than one element.
- Density Ratio: 0.025617 (**relatively sparse**).
- Any two nodes in the **same component** its shortest path length will **be equal 1**, except **component 1** which may be **more than 1**