Project ID: P16
Project Title: Facebook Social Circle
Group ID: 1
Team Members with ID numbers:

2022H1030005G	SHUBHAM SAWAIKER
2022H1030007G	TELANAKULA VENKATA SAI ROHITH
2022H1030020G	SAKSHI SINGH TANWAR
2022H1030033G	ADITYA SHRIVASTAVA

1. **Background of the Dataset:**
   This dataset comprises profile and network data sourced from Facebook, specifically from 10 ego networks comprising 193 circles and 4,039 users. A proprietary Facebook application was developed to collect this data, and a survey was conducted among ten users. These participants were tasked with manually identifying all the circles to which their friends on Facebook belonged.

   On average, participants identified approximately 19 circles within their ego networks, each circle comprising around 22 friends. These circles represent various social groups, including university students, sports teams, relatives, etc.

   In summary, this dataset offers insights into the structure of Facebook social networks, capturing user profiles and affiliations with different circles or groups on the platform. It provides an opportunity to analyse social connections, group dynamics, and network characteristics within this online social environment.

2. **Literature Review:**

   TITLE: Learning to Discover Social Circles in Ego Networks
   AUTHORS: Julian McAuley, Jure Leskovec
   AFFILIATION: jmcauley@cs.stanford.edu, jure@cs.stanford.edu
   SUMMARY:
   This paper explores the task of automatically categorising friends and acquaintances in online social networks, such as Facebook, Google+, and Twitter, into "social circles." These circles help users manage information overload, privacy, and sharing preferences.

   Current methods for creating circles are either manual or based on shared attributes, with limitations. The paper proposes a novel unsupervised approach to tackle this problem. It treats it as a clustering task within an individual user's ego network, the network of friends' connections.

The objective is to identify the circles each friend, referred to as "alters," belongs to, including nested and overlapping communities. This approach leverages network edge relationships and shared profile attributes.

The key innovations include predicting hard assignments of circles to multiple circles and allowing circles to form along different social dimensions based on profile similarity. The authors validate their method using a dataset of 1,143 ego-networks, outperforming existing methods.

In summary, this paper presents an innovative, unsupervised method to discover social circles in online social networks automatically. It combines network structure and user profile information to enhance accuracy and adaptability. This research has practical implications for managing online social connections.

3. Experimental Dataset

a. Exploratory Data Analysis:

The insights and conclusions drawn from Exploratory Data Analysis (EDA) on the Facebook social network data are as follows:

Based on Degree:

- The degree of a node is the number of edges incident on it.
- **Top 5 users:** [(107, 1045), (1684, 792), (1912, 755), (3437, 547), (0, 347)]
- Node 107 has 1045 Facebook friends, the most any Facebook user has in this analysis. Followed by nodes 1684 and 1912, having more than 700 friends.

Based on Degree Centrality:

- Degree centrality assigns an importance score based on the number of links each node holds. In this analysis, that means that the higher the degree centrality of a node is, the more edges are connected to the particular node, and thus, the more neighbor nodes (Facebook friends) this node has.
- **Top 5 users:** [(107, 0.258791480931154), (1684, 0.1961367013372957), (1912, 0.18697374938088163), (3437, 0.13546310054482416), (0, 0.08593363051015354)]
- Node 107 has the highest degree centrality with 0.259, meaning that this Facebook user is friends with around 26% of the whole network.
- Similarly, nodes 1684, 1912, 3437, and 0 also have very high degree centralities.
- The majority of users have degree centralities of less than 0.05 since the dataset consists of a friends list of particular nodes. Hence, many nodes have extremely low degree centralities as they are not interconnected in this network.

Based on Betweenness Centrality:

- Betweenness centrality measures the number of times a node lies on the shortest path between other nodes, meaning it acts as a bridge.
- In the Facebook graph, this measure is associated with the user's ability to influence others. A user with a high betweenness centrality acts as a bridge to many users that are not friends and thus can influence them by conveying information (e.g. by posting something or sharing a post) or even connect them via the user's circle (which would reduce the user's betweenness centrality after).
- **Top 5 users:** [(107, 0.4805180785560152), (1684, 0.3377974497301992), (3437, 0.23611535735892905), (1912, 0.2292953395868782), (1085, 0.14901509211665306)]
- Nodes 107, 1684, 3437, and 1912 exhibit both the highest degree and betweenness centralities, making them spotlight nodes in the network. This signifies that these nodes are the most popular entities within the network and can exert significant influence and disseminate information. However, this outcome is unsurprising as these nodes constitute part of the social circles within the network.
- In contrast, Node 1085, while not a spotlight node, possesses one of the highest betweenness centralities but does not have the highest degree centralities. This implies that despite not being the most popular users, such nodes hold considerable sway within the network, especially among the friends of spotlight nodes, concerning the dissemination of information.

Based on Closeness Centrality:

- Closeness centrality scores each node based on their 'closeness' to all other nodes in the network. For a node *'v'*, its closeness centrality measures the average fairness to all other nodes. In other words, the higher the closeness centrality of *'v'*, the closer it is to the network's center.
- **Top 5 users:** [(107, 0.45969945355191255), (58, 0.3974018305284913), (428, 0.3948371956585509), (563, 0.3939127889961955), (1684, 0.39360561458231796)]
- Examining users with the highest closeness centralities reveals a relatively small gap between them compared to previous metrics. Notably, nodes 107 and 1684 are the sole spotlight nodes among those with the highest closeness centralities. This suggests that a node with numerous friends may not be positioned near the network's center.

Based on Eigenvector Centrality:

- Eigenvector centrality is the metric that shows how connected a node is to other important nodes in the network. It measures a node's influence based on how well it is connected inside the network, how

many links its connections have, and so on. This measure can identify the nodes with the most influence over the network.

- **Top 5 users:** [(1912, 0.09540696149067629), (2266, 0.08698327767886552), (2206, 0.08605239270584342), (2233, 0.08517340912756598), (2464, 0.08427877475676092)]
- Node 1912 has the highest eigenvector centrality at 0.095, making it a spotlight node and undeniably the most crucial entity in the network in terms of overall influence. This node also holds some of the highest degree and betweenness centralities, indicating its significant popularity and influence over other nodes.
- Nodes 2266, 2206, 2233, and 2464 exhibit high eigenvector centralities while not being spotlight nodes. This is particularly intriguing because these nodes are newly identified, signifying that they do not hold the highest degree, betweenness, or closeness centralities in the graph. This suggests a strong likelihood of these nodes being connected to Node 1912, resulting in their exceptionally high eigenvector centralities.

Based on Bridges:

- An edge joining two nodes, A and B, in the graph is considered a bridge if deleting the edge would cause A and B to lie in two different components.
- The network consists of a total of 75 bridges. So many bridges exist because this network only contains the spotlight nodes and their friends. As a result, some friends of spotlight nodes are only connected to a spotlight node, making that edge a bridge.

b. Statistics and Measures:

DATASET STATISTICS:

1. *facebook_combined.txt*

Number of Nodes (users)	4039
Number of Edges (friendships)	88234
Connected Components	1
Diameter	8
Average Clustering Coefficient	0.607
Average Degree	43.691
Size of largest connected component	4039

Network has Bridges	True
Number of Bridges	75
Number of Local Bridges	78

Statistics for NodeId.edges

2. Node 0

Diameter	Cannot be determined; the graph is disconnected.
Number of Nodes (users)	333
Number of Edges (friendships)	2519
Average Degree	15.129
Average Clustering Coefficient	0.488
Number of Connected Components	5
Size of Largest Connected Component	324
Network has bridges	True
Number of bridges	31
Number of local bridges	80
Number of Circles	23

3. Node 107

Diameter	3
Number of Nodes (users)	10
Number of Edges (friendships)	27
Average Degree	5.4
Average Clustering Coefficient	0.805
Number of Connected Components	1
Size of Largest Connected Component	10
Network has bridges	False

Number of bridges	0
Number of local bridges	0
Number of Circles	8

4. Node 348

Diameter	9
Number of Nodes (users)	224
Number of Edges (friendships)	3192
Average Degree	28.5
Average Clustering Coefficient	0.542
Number of Connected Components	1
Size of Largest Connected Component	224
Network has bridges	True
Number of bridges	8
Number of local bridges	24
Number of Circles	13

5. Node 414

Diameter	Cannot be determined; the graph is disconnected.
Number of Nodes (users)	150
Number of Edges (friendships)	1693
Average Degree	22.57
Average Clustering Coefficient	0.668
Number of Connected Components	2
Size of Largest Connected Component	148
Network has bridges	True
Number of bridges	2

Number of local bridges	9
Number of Circles	6

6. Node 686

Diameter	6
Number of Nodes (users)	168
Number of Edges (friendships)	1656
Average Degree	19.714
Average Clustering Coefficient	0.564
Number of Connected Components	1
Size of Largest Connected Component	168
Network has bridges	True
Number of bridges	8
Number of local bridges	19
Number of Circles	13

7. Node 698

Diameter	Cannot be determined; the graph is disconnected.
Number of Nodes (users)	61
Number of Edges (friendships)	270
Average Degree	8.852
Average Clustering Coefficient	0.729
Number of Connected Components	3
Size of Largest Connected Component	40
Network has bridges	True
Number of bridges	3
Number of local bridges	3

Number of Circles	12

8. Node 1912

Diameter	Cannot be determined; the graph is disconnected.
Number of Nodes (users)	747
Number of Edges (friendships)	30025
Average Degree	80.388
Average Clustering Coefficient	0.636
Number of Connected Components	2
Size of Largest Connected Component	744
Network has bridges	True
Number of bridges	5
Number of local bridges	62
Number of Circles	45

9. Node 1684

Diameter	Cannot be determined; the graph is disconnected.
Number of Nodes (users)	786
Number of Edges (friendships)	14024
Average Degree	35.684
Average Clustering Coefficient	0.474
Number of Connected Components	4
Size of Largest Connected Component	775
Network has bridges	True
Number of bridges	12
Number of local bridges	213

Number of Circles	16

10. Node 3437

Diameter	Cannot be determined; the graph is disconnected.
Number of Nodes (users)	534
Number of Edges (friendships)	4813
Average Degree	18.026
Average Clustering Coefficient	0.544
Number of Connected Components	2
Size of Largest Connected Component	532
Network has bridges	True
Number of bridges	15
Number of local bridges	122
Number of Circles	31

11. Node 3980

Diameter	Cannot be determined; the graph is disconnected.
Number of Nodes (users)	52
Number of Edges (friendships)	146
Average Degree	5.615
Average Clustering Coefficient	0.452
Number of Connected Components	4
Size of Largest Connected Component	44
Network has bridges	True
Number of bridges	8
Number of local bridges	13

Number of Circles	16

c. **Suitable Network Types:**
The dataset illustrates an undirected homogeneous network, wherein every node symbolizes a user and the edges indicate friendships between pairs of users.

4. **Problem Statements:**
   a. Community Detection: Which clusters or communities of nodes form naturally within the network, and what are the characteristics of these communities?
   b. Central Node Identification: Which nodes are central or most influential in the network? Identifying these can provide insights into key players or hubs.
   c. Network Resilience: How does the removal of certain nodes or edges affect the overall structure and connectivity of the network?
   d. Justification: Community detection can help understand the inherent structure and segmentation of the network. Central node identification is crucial for understanding points of influence or vulnerability. Network resilience is essential for understanding the robustness of the network against failures or attacks.
5. **Solution Approach:**
   a. For Community Detection: Use algorithms like the Louvain method or Girvan-Newman algorithm to detect communities within the network.
   b. For Central Node Identification: Compute centrality measures such as degree centrality, betweenness centrality, and eigenvector centrality to identify central nodes.
   c. For Network Resilience: Conduct network robustness simulations by iteratively removing nodes (or edges) and observing the effect on network connectivity. Use metrics like the size of the largest connected component and average path length.

6. **Takeaways:**
   - Node with maximum betweenness centrality, 107: 0.4805180785560152
   - Node with maximum closeness centrality, 107: 0.45969945355191255
   - Node with maximum degree_centrality, 107: 0.258791480931154
   - When a single node has maximum values for all three of these centrality measures, it signifies that this node is not only highly connected (degree centrality), but it also serves as a critical bridge or intermediary between various parts of the network (betweenness centrality), and it is exceptionally accessible to other nodes in terms of communication or influence (closeness centrality). It's a central and influential node in multiple aspects of the network. Such a node could be crucial for information flow, control, or influence in the network.

7. **References:**

A. http://i.stanford.edu/~julian/pdfs/nips2012.pdf
B. Networkx library[1]

[1] https://networkx.org/documentation/latest/