Project ID: 16
Project Title: Facebook Social Circle
Group ID: 1
Team Members with ID numbers:

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

1. **Introduction:**
   Community detection in social networking refers to the task of identifying groups of individuals within a network who share common patterns of connections or interactions. The goal is to uncover the underlying structure of a social network by **grouping nodes** (representing individuals) in a way that nodes within a group have **more connections** or **similarities** with each other than with nodes in other groups. This process helps reveal communities or subgroups of individuals with stronger social ties, **common interests**, or **similar behaviour**, providing valuable insights into the organization and dynamics of the social network. Community detection has applications in various fields, including sociology, marketing, and epidemiology, where understanding the relationships between individuals can lead to a better understanding of social phenomena and more effective interventions.

2. **Motivation:**
   The motivation behind the community detection problem in social networking arises from the need to understand the complex structure and dynamics of social systems. Here are some key motivations:

   - **Identifying Subgroups**: Social networks are often large and intricate, making it challenging to discern meaningful patterns. Community detection helps in identifying subgroups or communities within the network, revealing clusters of individuals who share common interests, activities, or characteristics.

   - **Understanding Social Structure**: Communities represent a form of social structure. Detecting these communities helps researchers and analysts understand how individuals are organized within a network, uncovering hidden relationships and hierarchies.

   - **Enhancing Network Visualization**: Community detection aids in effectively visualizing large networks. By grouping nodes into communities, complex networks become more manageable, and the visual representation becomes more interpretable, facilitating insights into network structure.

   - **Targeted Marketing and Recommendations**: In social networks, understanding community structure is crucial for targeted marketing and recommendations. If users within a community exhibit similar preferences, behaviors, or purchasing patterns, targeted interventions or recommendations can be tailored to specific communities.

   - **Detecting Anomalies or Outliers:** Communities can help identify anomalies or outliers in a network. Deviations from the expected community structure may indicate events or behaviors that require attention, such as the spread of misinformation or the emergence of new trends.

- **Social Network Analysis for Research:** Researchers in fields like sociology, anthropology, and psychology use community detection to analyze social networks. It helps them study social phenomena, influence patterns, and the impact of individuals or groups on the overall network.

- **Epidemiology and Disease Spread:** In the context of epidemic modeling, understanding community structure is vital for predicting the spread of diseases. Communities may represent groups with higher interaction rates, aiding in the development of targeted intervention strategies.

- **Optimizing Communication and Connectivity:** Companies and organizations can benefit from community detection by optimizing communication strategies within and between communities. Understanding how information flows within communities helps in designing more efficient communication networks.

In essence, community detection is motivated by the desire to unravel the underlying structures of social networks, enabling a deeper understanding of human interactions, facilitating targeted interventions, and improving the efficiency of various applications in social sciences, marketing, public health, and more.

3. **Related work**:

- **Learning to Discover Social Circles in Ego Networks** (2012) by Julian McAuley and Jure Leskovec.
  AFFILIATION: Stanford, USA
  The authors tackle the issue of organizing personal social networks by introducing a machine learning task to identify users' social circles. They formulate the problem as a node clustering task within a user's ego-network, leveraging both network structure and user profile information. Through experiments on Facebook, Google+, and Twitter datasets with hand-labeled ground-truth, their model proves effective in accurately detecting and characterizing social circles, including overlapping and hierarchically nested ones.

- **Community detection and mining in social media** (2010) by Lei Tang and Huan Liu.
  AFFILIATION: Yahoo Labs, USA and Arizona State University, USA
  The paper introduces a data mining lecture on social media, emphasizing its transformative impact on collaboration and communication. The lecture covers key aspects, including characteristics of social media, computing tasks, challenges, and graph-based community detection techniques. It is designed for students, researchers, and practitioners, providing accessible insights into understanding and harnessing social media data for various applications.

- **Finding and evaluating community structure in networks** (2004) by M. E. J. Newman and M. Girvan.
  AFFILIATION: University of Michigan, USA and Cornell University, USA
  While not specific to Facebook, this seminal work on community detection methods is often referenced in the literature. It introduces the modularity metric, which measures the strength of community structure in networks and has been applied to Facebook data.

4. **Problem statement**:
   a. **Community Detection**: elaborates which clusters or communities of nodes form naturally within the network. Community detection, also known as **clustering or partitioning**, is a fundamental problem in network analysis. The goal is to identify groups of nodes in a network that are more **densely connected** to each other than to the rest of the network. These groups are referred to as "communities" or "**clusters**," and their detection provides insights into the underlying structure and organization of the network.
   In social networks, for example, communities might represent groups of individuals with stronger connections or shared interests. In citation networks, communities could correspond to thematic clusters of research papers. The problem of community detection can be formulated as follows:
   Given a network represented by a graph, where nodes represent entities (e.g., individuals, papers, websites) and edges represent relationships or interactions between entities, the task is to partition the nodes into groups in such a way that nodes within the same group have more connections between them than with nodes in other groups.
   b. The problem of community detection can be formulated mathematically using graph theory. Let's denote a graph as **G = (V, E)**, where V is the set of vertices and E is the set of edges. The objective is to partition the set of nodes V into k communities, denoted as C_1, C_2,…, C_k, such that the modularity of the partition is maximized.

   Modularity comes from the word 'module', a network-centric metric to determine the quality of a community structure.

   The modularity denoted by Q, for a partition is given as,

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

   Our goal is to find the partition that maximizes the modularity, which can be expressed as the optimization problem, $\max_{C1,\ C2,\dots,\ Ck} Q$. A detailed explanation is given in the next section.

5. **Hypothesis:**
   a. **Null Hypothesis (H_0)**: There is no significant structure or clustering in the network, and the observed connections between nodes are random.
   Hypothesis (H_1): There exists a meaningful structure in the network, and nodes form distinct communities or clusters.
   b. **Approach:**
   Mathematical Representation:
   Let G = (V, E) be the graph representing the network.
   C1, C2,..., Ck are the communities in the network.

   Null Model:
   Define a null model G_{null} where edges between nodes are distributed randomly, preserving the degree distribution of the original graph.
   Modularity Maximization:
   Formulate modularity **Q** as the objective function to maximize:

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

   Maximize Q over different community assignments as positive modularity shows the presence of a strong community structure.
   Ci is the identifier of the community to which node i belongs.

Ci, Cj results to value 1 represents that i and j belong to the same community and 0 represents otherwise.

Hypothesis Testing:

We conducted the statistical tests to compare the observed network's modularity with the null model's modularity. If H_0 is rejected, it supports the existence of meaningful communities.

Validation: If Q_{H1} is significantly greater than Q_{H0}, it indicates the presence of meaningful communities where Q represents the modularity.

Community Detection Algorithm*:*

We have utilize a community detection algorithm (e.g., Louvain method, spectral clustering) to find the partition that maximizes modularity. The algorithm aims to identify groups of nodes with higher internal connectivity compared to random connectivity.

The *Greedy algorithm* is build communities by greedily adding or removing nodes to optimize a certain objective function, often modularity. The process continues until no further improvement can be achieved.

The *Louvain method*, also known Louvain modularity optimization algorithm, is a community detection algorithm used in network analysis. It's an **iterative, bottom-up approach** that optimizes modularity by iteratively moving nodes between communities to find a partition of the network that maximizes the modularity score, indicating the strength of the community structure.

6. Solution Approach:

a. We are performing community detection using two algorithms, one is the greedy algorithm and another is the Louvain algorithm.

b. Greedy Algorithm:

i. **Mathematical Notation**:

Given a graph G = (V, E) and a set of communities $C_1, C_2, \ldots, C_k$.

The modularity gain ΔQ for moving a node i to a different community C_j is calculated as:

$$\Delta Q = \frac{1}{2m}\left[Q(i, C_j) - Q(i, C_i)\right]$$

The greedy algorithm iteratively moves nodes to the community that provides the maximum modularity gain until no further improvement is possible.

ii. **Code Mapping:**

In the code, the **greedy_modularity_communitiy_detection** function iteratively considers each node and evaluates the modularity gain for moving it to a different community.

Steps Involved:

➢ ***Initialization:***

1. Start with each node in its own community.
2. Calculate the initial modularity Q.

➢ ***Iteration:***

For each node: Calculate the modularity gain ΔQ for moving the node to different communities. Then, move the node to the community that maximizes ΔQ.

Subsequently, update the modularity Q accordingly. Lastly, we repeat the process until no further improvement is possible.

➢ ***Result:***

The final partition with the optimized modularity is the output of the program.

c. Louvain Algorithm:

i. <u>Mathematical Notation</u>:

Given a graph G = (V, E) and a set of communities $C_1, C_2, ..., C_k$. The modularity of the current partition is given by:

$$Q = \sum_i \left[\frac{L_i}{m} - \left(\frac{d_i}{2m} \right)^2 \right]$$

where L_i is the number of edges within community C_i, d_i is the sum of degrees of nodes in C_i, and m is the total number of edges in the graph.

The Louvain algorithm iteratively optimizes modularity by moving nodes between communities and merging communities.

ii. <u>Code Mapping</u>:

In the code, the **louvain_algorithm** function iteratively optimizes the modularity by moving nodes between communities. It considers the modularity gain for moving each node and merges communities when it maximizes the overall modularity.

Steps Involved:

➢ **Initialization:**

1. Start with each node in its own community.
2. Calculate the initial modularity Q.

➢ **Iteration (Phase 1 - Local Optimization):**

For each node:

1. Evaluate the modularity gain by moving it to neighboring communities.
2. Move the node to the community that maximizes modularity gain.
3. Repeat until no further improvement is possible.

Iteration (Phase 2 - Global Optimization):

1. Treat each community as a single node and construct a new network.
2. Apply Phase 1 to optimize modularity for the new network.
3. Repeat until no further improvement is possible.

➢ **Result:**

The final partition with optimized modularity is the program's output.

7. Results:

a. To validate our hypothesis, we calculated and found the modularity values for both H_0 (Null Modes) and H_1 (Original Graph) and got the following results:
Modularity of Random Graph: 0.11670217095924265
Modularity of Original Graph: 0.8348811587172857
Therefore, these results indicate the presence of meaningful communities in our original network.

b. We observed that there were a total of **17 communities** with the largest community (community 4) consisting of **535 nodes** (users) and the smallest communities (community 8 and 9) consisting of only **19 nodes** (users) each.

COMMUNITY	NODES
0	350
1	430
2	446
3	423
4	535
5	322
6	48
7	71
8	19
9	19
10	73
11	237
12	25
13	61
14	206
15	548
16	226

Table 7.1 Table represents each community to the number of node it is connected to.

c. Further, we used the following measures:
   i. **Internal Density:** Internal density measures how well-connected the nodes are within a community. It is the ratio of the number of actual edges within the community to the total possible number of edges within that community. A higher internal density indicates a more tightly-knit and interconnected community.
   ii. **Cut Ratio:** Cut ratio quantifies how many edges need to be removed to disconnect a community from the rest of the network. A lower cut ratio implies a more connected community. It is the ratio of the number of edges leaving the community to the number of internal edges within it. It suggests that the community is relatively isolated from the rest of the network.

8. **Insights:**
   a. Based on **Internal Density**: The top 5 communities were:

COMMUNITY	INTERNAL DENSITY
12	0.8666666666666667
8	0.7953216374269005
9	0.7543859649122807
11	0.5915397268111278

10	0.5654490106544902

Table 8.1 Represents top 5 communities based on Internal Density

b. Based on **Cut Ratio**: The top 5 communities were:

COMMUNITY	CUT RATIO
14	0.0008825012607160867
15	0.0021004480955937265
0	0.0045694200351493845
13	0.006067961165048544
16	0.013125

Table 8.2 Represents top 5 communities based on Cut Ratio

c. Even though communities 8 and 9 were the smallest in size, their internal density is relatively high compared to the rest of the communities, and these communities are tightly knit and cohesive.

9. References:

a. https://www.youtube.com/watch?v=QfTxqAxJp0U&t=702s
b. https://www.youtube.com/watch?v=F4RVBAGJcFY
c. https://www.youtube.com/watch?v=0zuiLBOIcsw&t=3s
d. https://domino.ai/blog/social-network-analysis-with-networkx
e. https://networkx.org/nx-guides/content/exploratory_notebooks/facebook_notebook.html
f. https://papers.nips.cc/paper_files/paper/2012/hash/7a614fd06c325499f1680b9896beedeb-Abstract.html
g. https://link.springer.com/book/10.1007/978-3-031-01900-5
h. https://journals.aps.org/pre/abstract/10.1103/hysRevE.69.026113