**Cluster Quest - Mapping Social Cohorts in Facebook Networks**

**Objective:**

In this project, we aim to develop and implement a graph-based unsupervised learning system that analyzes social media networks to discover community structures and cluster users based on their local network properties. By examining the Facebook ego network, we intend to extract meaningful node-level features and apply clustering techniques to identify socially cohesive groups. These insights can drive user segmentation, influencer identification, and community engagement strategies for digital marketing and social research.

This project operates at the intersection of **graph theory, unsupervised learning, and social network analysis**, supporting marketing, platform design, and social behavior understanding. Through clustering, businesses and researchers can gain a deeper understanding of user connectivity and influence dynamics within social media platforms.

**Dataset:**

The dataset is sourced from the **Stanford Network Analysis Project (SNAP)** and consists of the **Facebook Ego Network**. It includes:

* **4039 nodes (users)**
* **88,234 undirected edges**, representing mutual friendships
* No additional node attributes; all relationships are structure-based

Each line in the dataset (facebook\_combined.txt) represents an undirected edge between two users. This simple format supports building an unweighted, undirected graph, which can be directly imported into graph libraries like **NetworkX** or **igraph** for further analysis.

## **Working with Graph Data:**

The dataset will be loaded using NetworkX or igraph:

| import networkx as nx  G = nx.read\_edgelist('facebook\_combined.txt', create\_using=nx.Graph(),nodetype=int) |
| --- |

## Approach and Methodology:

### 1.Feature Engineering (Node-Level Local Features):

### We will derive local graph features for each node that describe its structural role in the network. These include:

* **Degree**: Number of immediate neighbors
* **Clustering Coefficient**: Measure of how interconnected a node’s neighbors are
* **Betweenness Centrality**: Measures node’s role in bridging communities
* **Closeness Centrality**: Inverse of the sum of distances to all reachable nodes
* **Average Neighbor Degree**: Average degree of a node’s neighbors
* **Ego Network Size**: Number of nodes in the 1-hop ego network
* **Local Efficiency**: Efficiency of communication in the ego network

Features will focus on **local structure only**, up to 2–3 hops, in line with the problem constraints.

We will convert the categorical features into embedding space so that we can work on numerical continuos features for clustering.

### 2. Data Preparation:

* All computed features will be compiled into a pandas DataFrame
* Normalization/scaling will be applied (e.g., MinMaxScaler or StandardScaler)
* Any NaN or infinite values will be handled appropriately
* The final dataset will be in structured tabular format suitable for clustering

### 3. Exploratory Data Analysis (EDA):

* Visualize feature distributions using histograms and boxplots
* Correlation matrix to check feature redundancy
* Network visualizations for high-degree nodes and community snapshots

### 4. Clustering:

We will apply and compare multiple clustering algorithms:

* K-Means Clustering
* Agglomerative (Hierarchical) Clustering
* DBSCAN (density-based clustering)
* Optionally: Spectral Clustering using graph Laplacian embeddings

Optimal number of clusters (k) will be chosen using the **elbow method** and **silhouette scores**.

### 5. Evaluation:

We will assess clustering performance using unsupervised metrics:

* Silhouette Score
* Calinski-Harabasz Index
* Davies-Bouldin Index

### 6. Interpretation and Analysis:

* Identify and characterize each cluster: high centrality groups, isolated users, bridge nodes, etc.
* Explore community overlaps, potential influencers, and anomalies
* Visualize clusters using dimensionality reduction (e.g., PCA or t-SNE)

### 7. Refinement and Iteration:

* Revisit feature engineering or clustering algorithm based on evaluation
* Experiment with edge pruning, alternate centrality measures, and subgraph analysis
* Tune clustering parameters for improved cohesion and separation

## Potential Users:

* **Marketers**: Segment audiences and identify target communities
* **Platform designers**: Improve recommendation and engagement algorithms
* **Academic researchers**: Study user behavior and network dynamics
* **Product teams**: Discover user groups for feature rollout strategies

This application fits into the digital marketing and social analytics space. By leveraging graph-based clustering, it enables data-driven segmentation and engagement strategies, uncovers influential nodes, and offers structural insights that businesses and researchers can use to navigate complex social networks effectively.

**Reference:**

McAuley, J., & Leskovec, J. (2012). *Learning to discover social circles in ego networks*. Stanford Network Analysis Project (SNAP). <https://snap.stanford.edu/data/ego-Facebook.html>