**College of Engineering Guindy**

**Anna University**

**Department of CSE**

**DATA MINING PROJECT**

**PROPOSAL**

**Community detection and influence analysis in social networks**

**Submitted By :**

**KISHORE K (2023103533)**

**HARI S (2023103551)**

**CSE - Q**

**Problem Statement:**

Social media platforms generate massive interaction data every second, forming complex user networks. Understanding how users group into communities and who holds the most influence is crucial for applications such as targeted marketing, trend prediction, and misinformation control.  
However, manually identifying communities or key influencers is infeasible at scale due to the graph’s size and complexity.  
This project aims to detect communities and rank influential users in a large-scale social media graph using graph-based models, providing insights into information flow and network structure.

**Solution Methods**

Our approach involves:

1. **Data Representation:**
   * Construct a social graph with nodes representing users and edges representing interactions (friendship, mentions, retweets).
2. **Graph Preprocessing:**
   * Remove isolated nodes, normalize edge weights, and create the largest connected component.
3. **Community Detection:**
   * Apply multiple graph mining algorithms to identify densely connected subgroups.
4. **Influence Analysis:**
   * Compute centrality measures and node rankings to identify opinion leaders.
5. **Visualization & Interpretation:**
   * Use graph visualization tools to display communities, highlight influencers, and interpret results.

**Models Applied**

To ensure robustness, at least three models will be applied and compared.

| **Category** | **Model** | **Purpose** |
| --- | --- | --- |
| **Classical Community Detection** | **Louvain Algorithm** | Optimizes modularity to find communities quickly and efficiently. |
|  | **Infomap Algorithm** | Uses information-theoretic approach to detect flow-based communities. |
| **Node Embedding + Clustering** | **node2vec + KMeans** | Generates node embeddings capturing structural similarity, then clusters users. |
| **Influence Analysis** | **PageRank** | Identifies most influential users based on connectivity. |
|  | **Eigenvector / Betweenness Centrality** | Cross-verifies influence ranking by measuring reach and network control. |

These models will be compared on modularity, conductance, runtime, and interpretability.

**Dataset Used**

We will use SNAP datasets (Stanford Network Analysis Project):

* Facebook Social Circles dataset – undirected edges representing friendships.
* Twitter Social Graph dataset – directed edges representing follower relationships.

Each dataset is pre-cleaned and widely used in research, ensuring reproducibility and scalability.

**Visualization Tools**

* Gephi – interactive visualization and modularity coloring for communities.
* NetworkX + Matplotlib – Python-based plots for quick analysis.
* PyVis / Plotly – interactive web-based visualization of large networks.

Communities will be color-coded, node sizes will represent influence scores, and top influencers will be highlighted.

**Experimental Pipeline**

**Step 1: Data Collection & Preprocessing**

* Load SNAP dataset (Facebook/Twitter graph).
* Build edge list → remove duplicates → construct graph (NetworkX/iGraph).
* Filter isolated nodes and generate largest connected component.

**Step 2: Community Detection**

* Run Louvain, Infomap, and node2vec + KMeans.
* Store community labels for each node.

**Step 3: Influence Analysis**

* Compute PageRank, Eigenvector centrality, and Betweenness centrality.
* Rank nodes by influence score and pick top-k influencers.

**Step 4: Evaluation & Comparison**

* Compute **Modularity**, **Conductance**, and **Cluster Size Distribution**.
* Compare runtime and scalability of each model.
* Validate influencer rankings using correlation with degree distribution.

**Step 5: Visualization**

* Visualize community structures (color-coded clusters).
* Highlight top influencers in network diagrams.
* Generate plots (community size distribution, top influencer list).

**Step 6: Reporting & Insights**

* Summarize key findings:
  + Community patterns (number, size distribution).
  + Influencer lists & role in information flow.
* Discuss real-world applications (marketing, fake news containment).

**Conclusion**

This project will reveal how users form communities and which individuals drive influence within social media networks. Using SNAP data, we apply Louvain, Infomap, and node2vec + KMeans for community detection, and centrality measures for influencer ranking. The results, supported by clear visualizations, will offer actionable insights into network structure and information flow, with applications in marketing, recommendation systems, and misinformation control.