**Facebook Network Analysis: A Large-Scale Social Graph Study**

**Author:** Mildred Okonye

**Slack ID:** Mildred O

**Date:** 20th March 2025

**Project:** Facebook Network Analysis

**Tools Used:** Python, NetworkX, Pandas, Matplotlib, Seaborn

**Table of Contents**

1. Title Page
2. Introduction
3. Objective
4. Dataset Overview & Data Cleaning
5. Methodology
6. Network Analysis & Results

* Basic Network Statistics
* Centrality Measures
* Degree Distribution
* Community Detection
* Inter-Community Connectivity
* Most Influential Nodes

1. Visualization & Insights
2. Discussion & Key Findings
3. Conclusion
4. Reference

**2. Introduction**  
  
This work examines the structural properties of a large Facebook network dataset belonging to the musae\_facebookdataset(<https://snap.stanford.edu/data/facebook-large-page-page-network.html>). Applying network science principles, we study the most important factors like degree distribution, network centrality, clustering coefficients, and community structures. We identify prominent nodes, between-community connections, and how the different types of pages affect the building of the network structure in our work. Results indicate that the network is sparsely yet densely connected, with some nodes playing a central role in information propagation. These findings aid in understanding large social graphs and dynamics.

**2.1 .Objectives**  
  
Social networks like Facebook are complex networks of nodes (users, pages) and edges (links) that are linked. Understanding their structure is important for a wide range of applications, including marketing campaigns, information dissemination, and security. This research

• Analyses the connectivity and structure of the Facebook network.

• Identify the most influential pages using centrality measures.

• Detect and examine communities within the network.

• Visualize key insights such as degree distribution, clustering, and network density.

• Explore relationships between page types and their structural importance.

**2.2 Dataset Overview**  
The dataset, downloaded from the Stanford Large Network Dataset (Facebook Large) Collection, contains two main components:  
  
**1.** **Edge List (musae\_facebook\_edges.csv)** – The relationships (friendships) between Facebook pages.  
 **2. Node Attributes (facebook\_large.csv)** – Contains attributes of Facebook pages Includes metadata like ID, Facebook ID, page type and page name.

2.2 Data Cleaning

* No missing values are detected.
* Data was structured with nodes representing pages and edges indicating friendships.
* The dataset was loaded into **NetworkX** graph.

**Dataset Statistics:**

|  |  |
| --- | --- |
| Metrics | Values |
| Total Nodes (Pages) | 22,740 |
| Total Edges (Connections) | 171,002 |
| Average Degree | 15.22 |
| Network Density | 0.000677 (Sparse Network) |
| Largest Connected Component Size | 22,470 (Whole network is connected) |
| Clustering Coefficient | 0.3597 (Moderate local connectivity |

**3. Methodology**

We applied the following network analysis techniques:

* **Graph Construction**: Created a network G from the edge list with NetworkX**.**
* **Basic Statistics:** Computed the number of nodes, edges, average degree, and density.
* **Centrality Measures**: Discovered influential nodes with Degree, Betweenness, and Eigenvector centralities.
* **Community Detection:** Applied **Label Propagation Algorithm (LPA)** for community detection.
* **Network Visualization:** Used **Matplotlib and Seaborn** for plotting degree distributions, connectivity, and community structure**.**

**4. Network Analysis & Results**  
  
4.1 Basic Network Statistics  
  
• The average degree (15.22) shows most pages are moderately connected.  
• The network density (0.00068) is low, which suggests a sparse network where most nodes are not connected.  
• A high clustering coefficient (0.36) suggests that pages tend to cluster locally. **A computer screen shot of a code

AI-generated content may be incorrect.**

**A number and numbers on a white background

AI-generated content may be incorrect.**

4.2 Centrality Analysis

We calculated three key centrality measures:

|  |  |  |
| --- | --- | --- |
| Metric | Top Node ID | Centrality Score |
| Betweenness Centrality | 701 | 0.1194 |
| Eigenvector Centrality | 16895 | 0.1778 |
| Degree Centrality | Multiple Pages | High Connectivity |

* Betweenness Centrality: Node 701 is a significant bridge between different components of the network.
* Eigenvector Centrality: Node 16895 is highly influential, perhaps because it is linked to other influential pages.
* Degree Centrality: Measures the number of direct connections.

4.3 Degree Distribution  
  
The network has a right-skewed degree distribution with most nodes having low degrees and a few nodes having very high connectivity.  
  
• Visualization: A log-scale histogram of degree distribution is typical of power-law-like behaviour, which shows the presence of hub nodes.

A graph of a graph

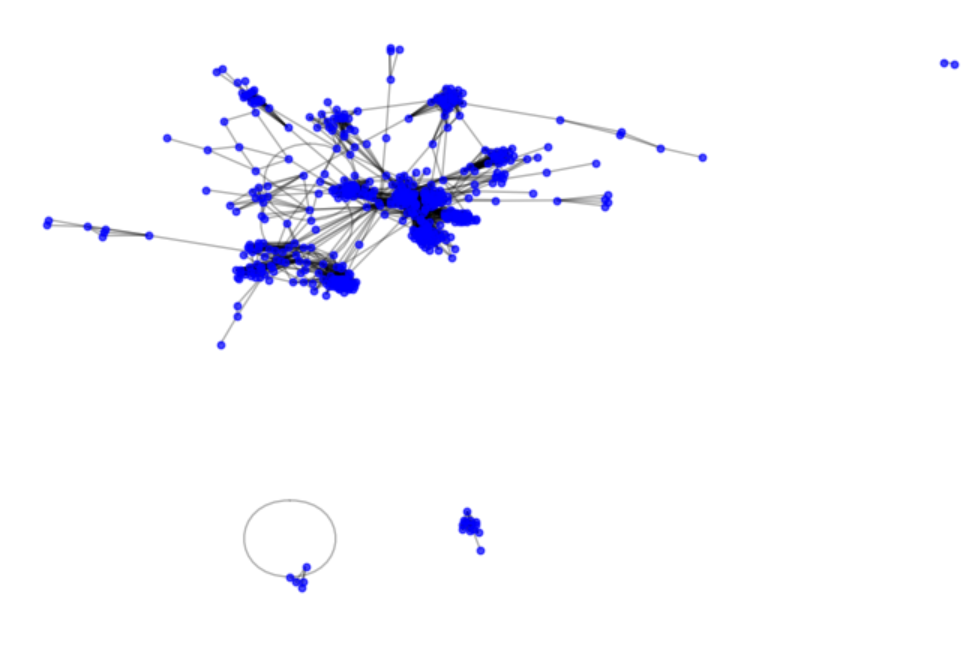
AI-generated content may be incorrect.

A graph of a degree

AI-generated content may be incorrect.

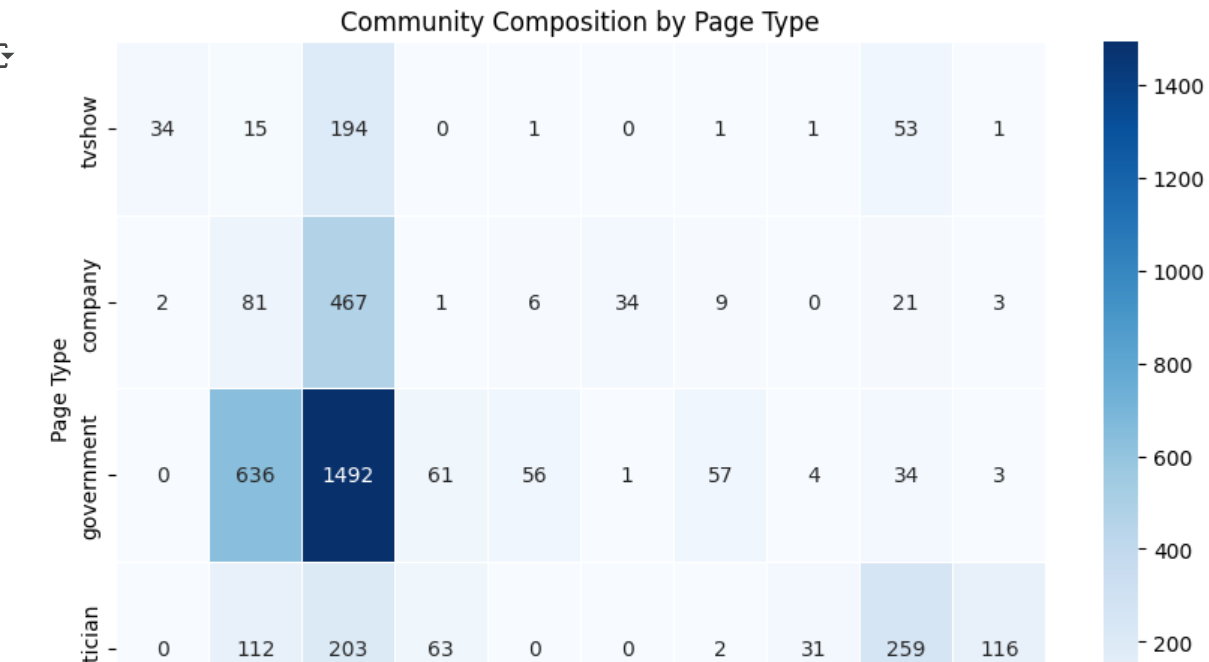
4.3 (b) Network Structure

A subgraph (500 nodes) was visualized using a force-directed layout. The network reveals a highly clustered structure with distinct communities.



4.4 Community Detection  
  
We detected several communities using Label Propagation Algorithm (LPA) and examined the largest ten (10) communities in closer detail.

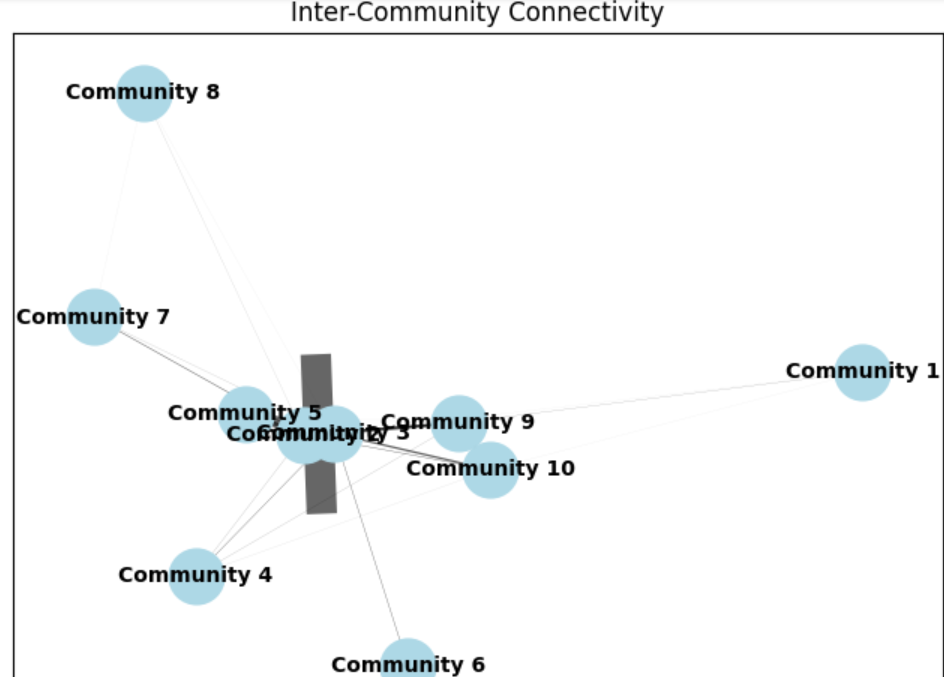
   •   The most prominent community comprises government pages, pointing to a strong inclination of similar entities to cluster in cohesive groups.  
  
  •    Visualization: A heatmap of page type composition of communities illustrates that politician, TV shows, and news pages cluster together.



A screenshot of a graph

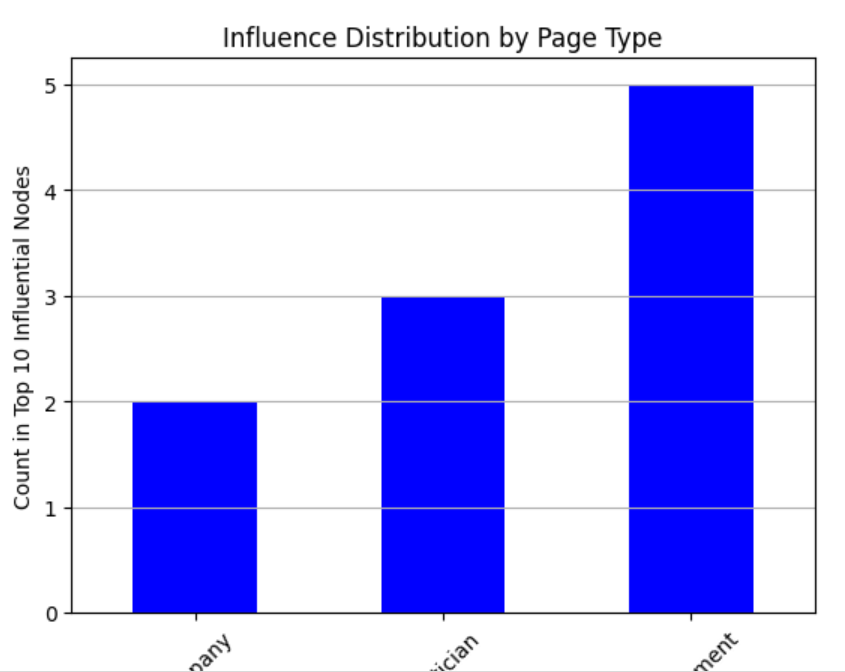
AI-generated content may be incorrect.

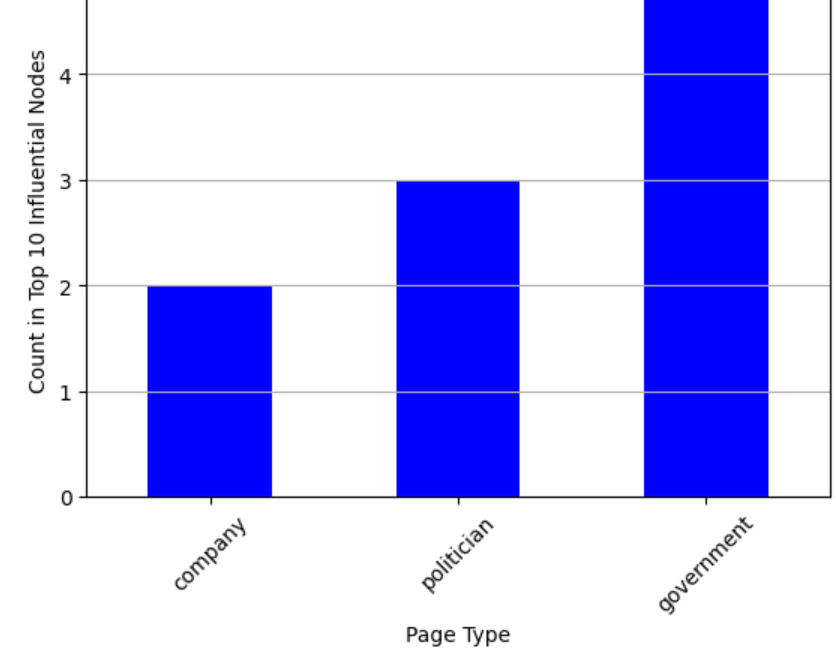
4.5 Inter-Community Connectivity  
To observe how different communities are connected to one another, we formed an inter-community network.  
  
• Weakly connected communities but bridging communities facilitate information sharing.  
  
• Visualization: A force-directed graph of inter-community edges displays sparse but vital edges.



4.6 Most Influential Page Types

Based on centrality measures, the analysis shows that government pages are among the most influential nodes in the network, especially in terms of betweenness centrality. TV shows and entertainment-related pages also appear as influential in terms of eigenvector centrality.

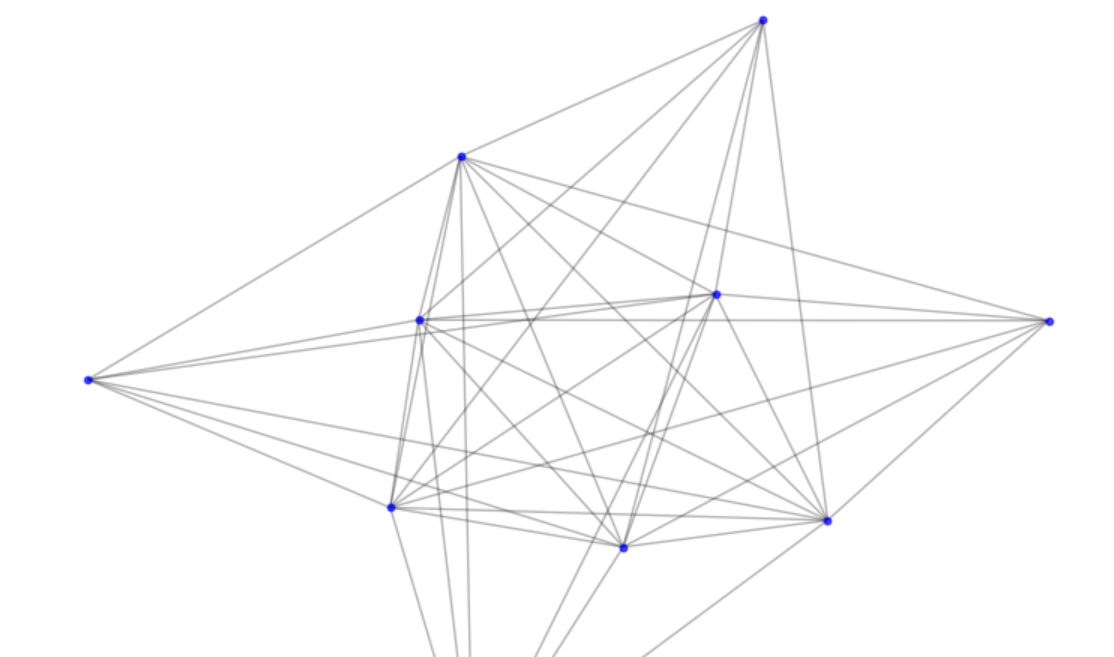




Connectivity of Central Modes

A subgraph of the top 10 central nodes was analyzed.

* These nodes are **highly interconnected**.
* They act as **key influencers** in the Facebook network.



**5. Visualizations & Insights**

1. Degree Distribution: A log-scale histogram shows the skewed distribution, with most pages having a low degree and a few hubs having many connections.

2. Largest Connected Component: A visualization of the largest connected component reveals how the entire network is interconnected.

3. Community Composition: A heatmap shows how different page types are distributed across the detected communities, providing insights into network segmentation.

4. Inter-Community Connectivity: A graph of inter-community connections illustrates how isolated or connected different parts of the network are.

5. Most Central Nodes: A subgraph of the central nodes based on degree centrality highlights the most influential pages in the network.

**6. Discussion & Key Findings**

1. Sparse but Connected Network: Despite the low density, the network is fully connected, with a well-established structure.
2. Influential Nodes: A few pages act as key connectors and bridges between different parts of the network, influencing the flow of information.
3. Community Structure: Communities tend to form around specific themes or page types, such as politics, government, or news.
4. Hubs Drive Connectivity: Certain nodes act as critical bridges between different network regions.
5. Potential Real-World Applications:

• Marketing: Identifying influential nodes for targeted advertising.

• Information Diffusion: Leveraging central nodes for the rapid spread of information.

1. **Conclusion**

This research provides valuable insights into the structure of large-social networks. Key findings highlight the role of influential nodes and community structures in shaping information diffusion.

**8. References**

1. SNAP: Stanford Large Dataset Collection

<https://snap.stanford.edu/data/facebook-large-page-page-network.html>

1. Facebook network analysis (Mildred O)

<https://colab.research.google.com/drive/12JVSSLFgHJ5oIODRvNKJQvpTwSs_fvO?usp=drive_link>