final version

**Project Abstract: Exploring the Page-Page Graph of Verified Facebook Sites**

For this project, we propose to explore a publicly available dataset representing a page-page graph of verified Facebook sites. This dataset provides a unique opportunity to analyze the structure and dynamics of mutual connections among officially endorsed pages on one of the world's largest social platforms.

**Dataset Description**The key components and variables:

* **Nodes**: Each node represents a verified Facebook page.
* **Edges**: Edges signify mutual “like” relationships between two nodes.
* **id\_1**: Denotes one verified Facebook page in a mutual connection.
* **id\_2**: Represents the corresponding page that forms a connection with id\_1.

Each row thus represents an undirected edge between two nodes. The dataset includes approximately **22,470 nodes** and **171,002 edges**, offering a rich network to study. In addition, there is a JSON file named ‘**musae\_facebook\_features.json’** that contains feature embeddings. This file maps numeric identifiers to corresponding lists of numeric values, which serve as typical embeddings in machine learning, representing items or nodes in a multidimensional feature space. This is helpful in implementing clustering and PCA.

**Goal:** Perform EDA on Facebook graph data to mine meaningful insights into the community, categories and feature relationships.

**Q1.** To understand the structural features of Facebook pages. How many communities can be identified? What is the size of each community, and what are the dominant categories in each community?

**Application:** Reveal network structural characteristics by identifying communities and their dominant categories to support targeted marketing and collaboration strategies.

**Procedures:** Community detection using modularity maximization, Categorization within the community, Evaluate community significance using graph metrics.

**Q2.** To analyze the distribution of influence among Facebook pages, which page has the highest influence within each category? Which community demonstrates the greatest overall influence? What are the most influential pages in each community?

**Application:** Identify influential pages and communities, leveraging network structures to target them for marketing or partnerships to maximize engagement.

**Procedures:** Centrality Analysis by degree, betweenness and closeness, Aggregated influence at the community level, Validate influence using Cross-category edges.

**Q3.** To examine the relationship between features and categories, are there noticeable differences in the distribution of pages across categories in the feature space? Which category forms tighter clusters, and which shows greater variation?

**Application:** Enhance understanding of category distributions through feature pattern analysis, which can help the business tailor their marketing or partnership strategies based on the category's connectivity patterns and utilize page features that occur frequently across categories.

**Procedures:** Dimensionality reduction by PCA or t-SNE, Cluster analysis, Statistical analysis of feature variations.

EDA

**Goal:** Perform exploratory data analysis (EDA) on Facebook graph data to uncover meaningful insights into the relationships between verified Facebook pages and their categories: politicians, governmental organizations, television shows, and companies.

1. Which communities exhibit higher overall influence?  
Which pages have the highest influence within each category?  
Is the distribution of page categories consistent across different communities?

**Procedures/Techniques:** Community detection, Graph centrality measures (Degree, Betweenness, Closeness, PageRank), Modularity analysis for community strength, Network visualization with force-directed layouts, Graph partitioning algorithms (spectral clustering)

**Applications:** Target influential pages and communities for marketing or partnerships to maximize engagement.

2. What factors contribute to the categorization of pages as politicians, companies, or TV shows based on their network and features?

**Procedure/Techniques:** Clustering (k-means or hierarchical clustering), Dimensionality reduction (PCA, t-SNE), Community detection

**Applications:** Enable predictive categorization and marketing strategies tailored to successful characteristics of influential pages.

3. How do the size and connectivity of pages in different categories vary?

What are the key differences in the types of interactions (edges) between pages from different categories?

**Procedure/Techniques:** Graph Centrality Measures (Degree, Betweenness, Closeness), Descriptive Statistics, t-SNE for Visualization

**Applications:** Understanding the strengths of different categories in terms of reach and connectivity, helping the business tailor their marketing or partnership strategies based on the category's connectivity patterns.

EDA\_V2

**Goal:** Exploratory data analysis (EDA) is performed on Facebook graph data to mine meaningful insights into the community relationships formed by different Facebook pages and the relationships between their categories (politicians, government organizations, TV shows, companies).

**Q1.** To understand the structural features of Facebook pages, how many communities can be identified? What is the size of each community, and what are the dominant categories in each community?

**Application:** Reveal network structural characteristics by identifying communities and their dominant categories to support targeted marketing and collaboration strategies.

**Procedures**

1. Community detection using modularity maximization.
2. Categorization within the community.
3. Evaluate community significance using graph metrics.

**Q2.** To analyze the distribution of influence among Facebook pages, which page has the highest influence within each category? Which community demonstrates the greatest overall influence? What are the most influential pages in each community?

**Application:** Identify influential pages and communities, leveraging network structures to optimize engagement strategies and resource allocation.

**Procedures**

1. Centrality Analysis by degree, betweenness and closeness.
2. Aggregated influence at the community level.
3. Validate influence using Cross-category edges.

**Q3.** To examine the relationship between features and categories, are there noticeable differences in the distribution of pages across categories in the feature space? Which category forms tighter clusters, and which shows greater variation?

**Application:** Enhance understanding of category distributions through feature pattern analysis, providing a basis for unsupervised page classification.

**Procedures**

1. Dimensionality reduction by PCA or t-SNE.
2. Cluster analysis.
3. Statistical analysis of feature variations.

**Q4.** To assess whether existing tags sufficiently generalize page characteristics, how do classification results based on tags compare to clustering results based on features? Do tag-based classifications better distinguish pages than feature-based clustering?

**Application:** Validate the applicability of tags through feature clustering, refine classification standards, and uncover overlooked page categories.

**Procedures**

1. Classification by given labels.
2. Unsupervised Clustering.
3. Comparison Metrics.

Data Description

**Project Abstract: Exploring the Page-Page Graph of Verified Facebook Sites**

For this project, we propose to explore a publicly available dataset representing a page-page graph of verified Facebook sites. This dataset provides a unique opportunity to analyze the structure and dynamics of mutual connections among officially endorsed pages on one of the world's largest social platforms.

**Dataset Description**The dataset comprises two key components:

* **Nodes**: Each node represents a verified Facebook page.
* **Edges**: Edges signify mutual “like” relationships between two nodes.

The data is structured as a table with two columns:

1. **id\_1**: Denotes one verified Facebook page in a mutual connection.
2. **id\_2**: Represents the corresponding page that forms a connection with id\_1.

Each row thus represents an undirected edge between two nodes. The dataset includes approximately **22,470 nodes** and **171,002 edges**, offering a rich network to study. In addition, there is a JSON file named ‘**musae\_facebook\_features.json’** that contains feature embeddings. This file maps numeric identifiers to corresponding lists of numeric values, which serve as typical embeddings in machine learning, representing items or nodes in a multidimensional feature space. This is helpful in implementing clustering and PCA.