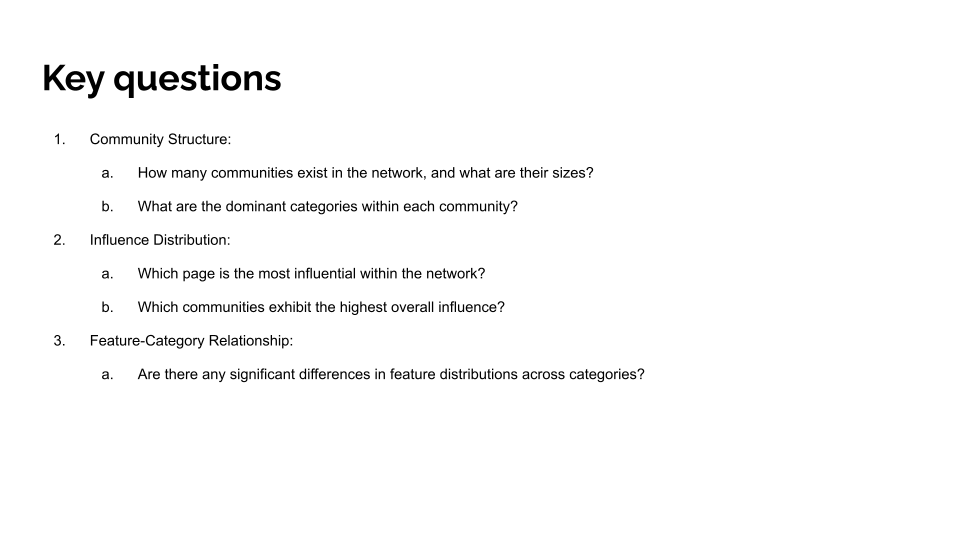
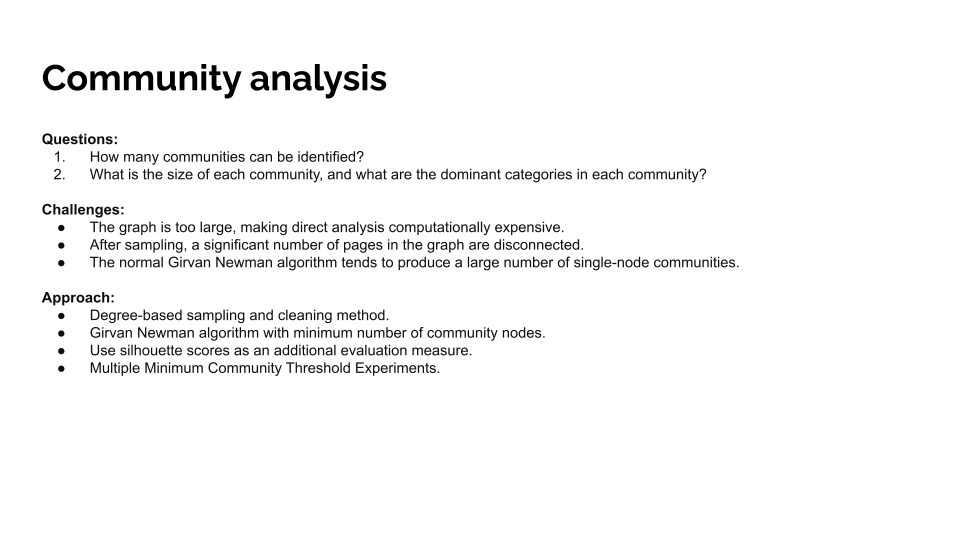
yuqi draft

****

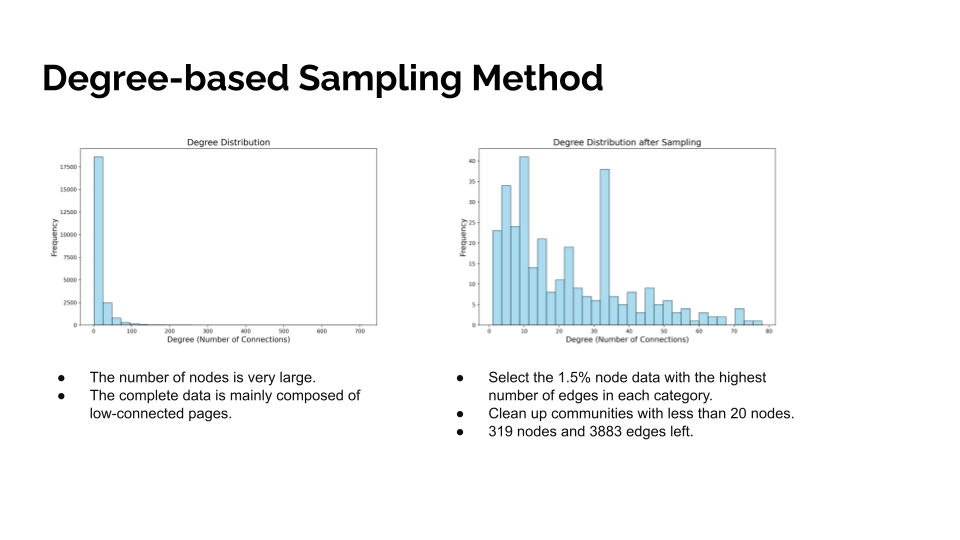
**Slide 4**

By framing the project around three key questions, we clarified the main focus areas, exploring Facebook pages’ community structure, influence distribution, and feature relationships. These questions are significant as they not only unveil the overall network structure but also provide targeted guidance for business decision-making. Analyzing community structure offers insights into how nodes are organized within the network and the distribution patterns of different types of nodes. Influence distribution analysis further identifies key nodes and communities that play critical roles in the network, providing a basis for enterprises to formulate precise marketing and resource allocation strategies. Finally, studying the relationship between features and categories helps distinguish characteristics across different categories, offering data support for market segmentation and personalized marketing strategies. The analysis framework is progressive, with each question building on the previous, ultimately delivering a comprehensive view for business applications.

**Slide 5**

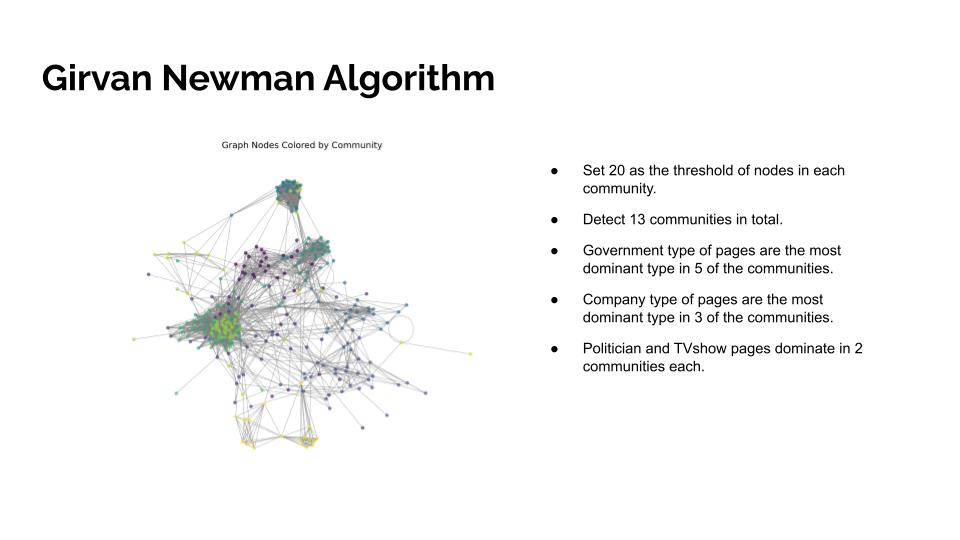
When analyzing the network’s community structure, we faced three major challenges. First, the graph is extremely large, consisting of 22,470 nodes and 171,002 edges, which demands significant computational resources to analyze the entire structure. In fact, we initially attempted to analyze the full graph, but due to the excessive runtime, we had to explore alternative methods. Second, many nodes have low connectivity, which contributes minimally to the community structure analysis and can even distort results by affecting the division of highly connected nodes. Lastly, the traditional Girvan-Newman algorithm tends to produce single-node communities, which contradicts our goal of analyzing overall community characteristics.

To address these challenges, we designed a degree-based sampling method, focusing on high-degree nodes to significantly reduce data size and improve analysis efficiency and effectiveness. Additionally, we modified the Girvan-Newman algorithm by introducing a node count threshold to avoid generating numerous insignificant small communities. We conducted experiments using thresholds of 10, 20, 30, 40, and 50 nodes per community. To ensure robust evaluation, we used both modularity and silhouette scores, as relying solely on modularity may introduce bias. After comparison, a threshold of 20 nodes produced the best overall results for modularity and silhouette scores. This method not only improved computational efficiency but also ensured the reliability and interpretability of the results.

**Slide 6**

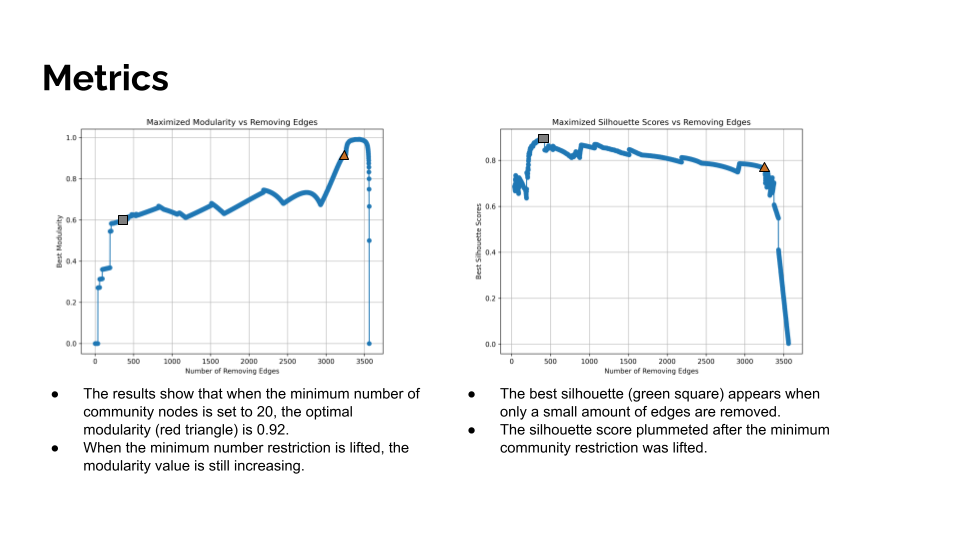
To effectively address the issue of the graph’s large scale, we adopted a degree-based sampling strategy during the data reduction phase, selecting the top 1.5% of nodes with the highest degree in each category. This method focuses on the key nodes that have the greatest impact on the network structure, significantly reducing data complexity while maintaining the original distribution of categories. We then set a minimum community size threshold of 20 nodes to filter out small, insignificant communities.

This process reduced the original dataset from 22,470 nodes and 171,002 edges to 319 nodes and 3,883 edges. It not only decreased the computational burden but also established a solid foundation for subsequent community detection. By preserving the category distribution in the original data and retaining the most critical node information, this sampling approach is particularly suitable for the structural analysis of large-scale networks.

**Slide 7**

To enable the Girvan-Newman algorithm to detect communities with sufficient size and representativeness, we set a minimum threshold of 20 nodes per community. Under this restriction, modularity was used as a metric to identify the optimal community structure. Once no more edges could be removed, the restriction was lifted, but the best community structure was not updated further until all edges were removed.

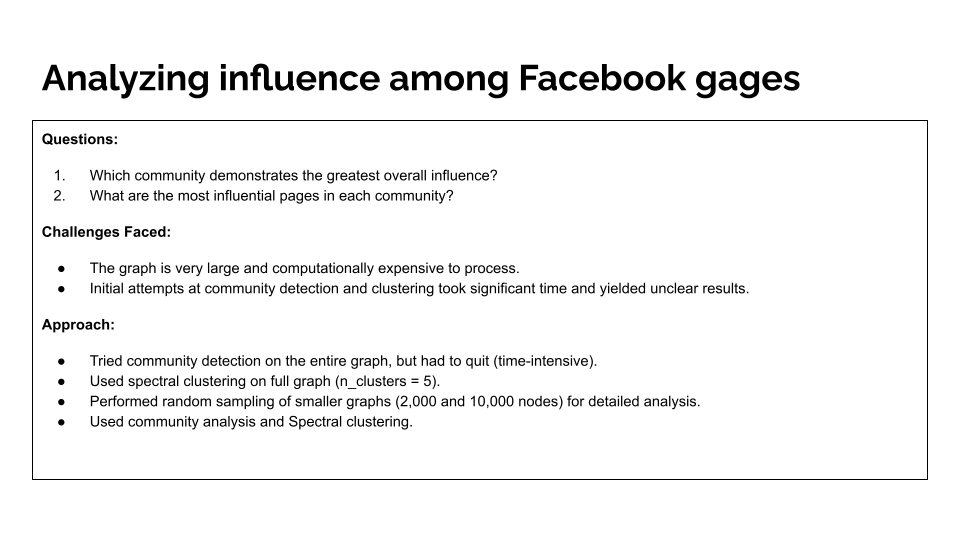
Using this improved Girvan-Newman algorithm, we identified 13 significant communities within the network. Among these, five communities were dominated by government pages, highlighting their prominence; three communities were dominated by company pages, reflecting the importance of business nodes. Other types, such as politicians and TV shows, dominated two communities each. Additionally, the dominant category in each community accounted for more than 50% of its pages. This analysis not only reveals the concentrated distribution of categories within the network and the strong association between communities and categories but also provides critical insights into the interactions and distributions among communities. These findings offer a reliable foundation for further exploration of the relationship patterns and potential value within social networks.

**Slide 8**

To validate the effectiveness of community detection, we introduced two key metrics: modularity and silhouette scores. Modularity measures the compactness and rationality of community divisions, achieving its optimal value of 0.92, represented by the red triangle in the figure. This result indicates a well-structured community division. Silhouette scores reflect the separation between communities and the cohesion within them, achieving the best score when only a small number of edges were removed. Notably, the silhouette score remained relatively high throughout the process.

After removing the node count restriction, modularity initially increased slightly but then dropped rapidly, while the silhouette score plummeted to 0, indicating blurred community boundaries and significantly reduced quality of division. These results demonstrate that setting a reasonable node count threshold not only improves computational efficiency but also ensures the scientific validity and interpretability of community detection. The evaluation results provide strong evidence for the rationality of the algorithm design.

Aish

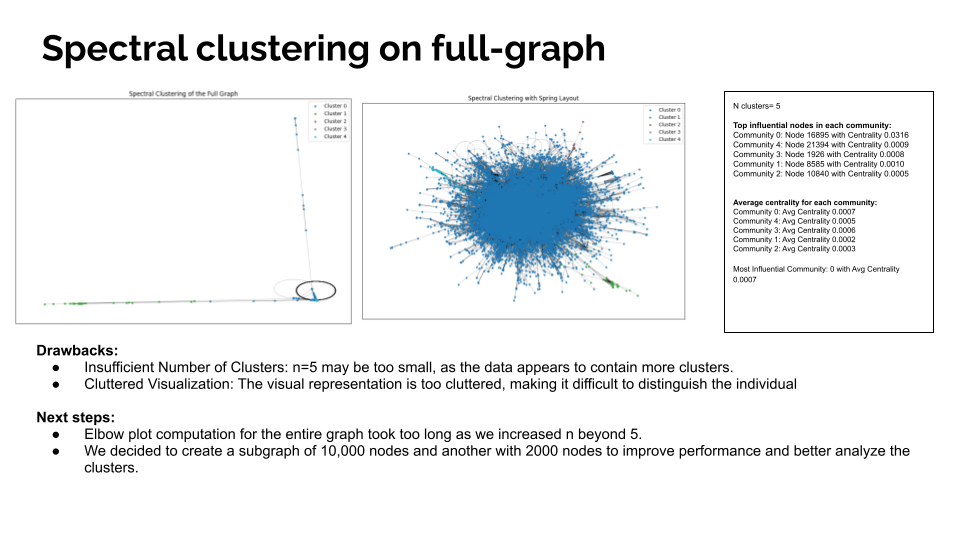
****

**Slide 9: Influence Analysis**

In this segment of the project, we focused on exploring the concept of influence within the Facebook graph network. The analysis was structured around three guiding questions:

1. Which community demonstrates the greatest overall influence?
2. What are the most influential pages within each community?

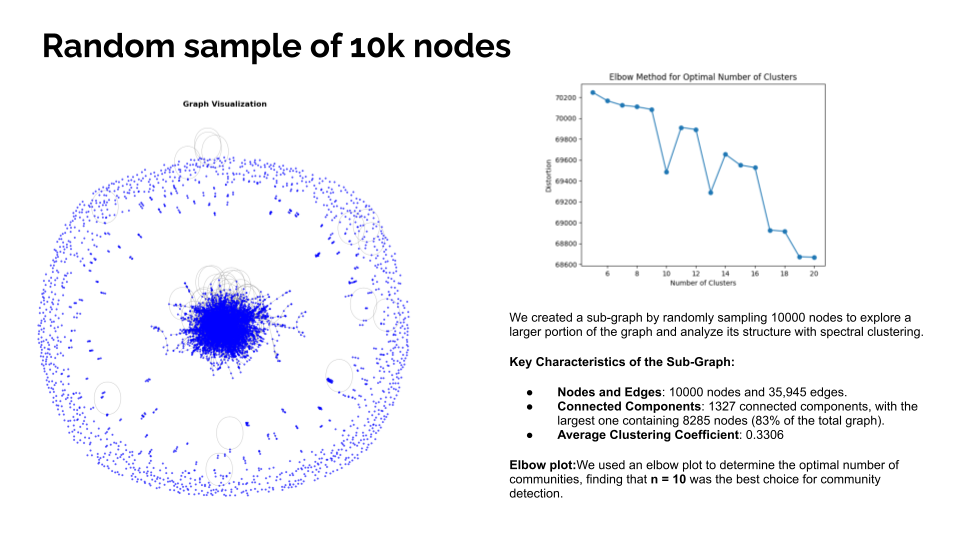
To tackle these questions, we attempted to assess influence across the graph’s structure. However, the analysis was not without challenges. The graph's vast size posed significant computational hurdles. Processing the graph to detect communities and clusters required extensive time and resources, often leading to unclear or suboptimal results in our initial attempts. This necessitated iterative refinement of our methods to manage the computational complexity effectively and derive meaningful insights.

****

**Slide 10: Spectral Analysis**

Our initial approach involved performing spectral clustering on the full graph, starting with five clusters. This method provided some preliminary insights into the community structure. However, we quickly recognized that using only five clusters was insufficient to capture the graph's complexity. The resulting visualization was overly cluttered, making it difficult to discern clear patterns or interpret the data meaningfully.

To refine our approach, we attempted to generate an elbow plot for a higher number of clusters. However, this process proved computationally expensive and time-consuming due to the graph's size. Consequently, we shifted our strategy towards creating subgraphs, allowing us to improve computational performance and analyze the clusters more effectively.

****

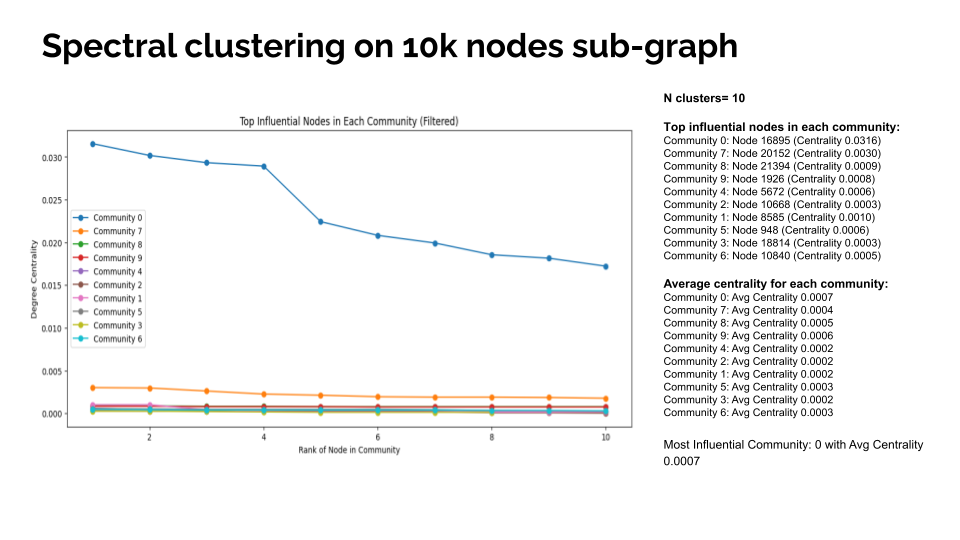
**Slide 11: Subgraph Creation and Cluster Analysis**

To address computational challenges, we created a subgraph by sampling 10,000 nodes from the original graph. This subgraph contained 35,945 edges and 1,327 connected components. While random sampling introduced some disconnected nodes, the subgraph retained enough data to provide meaningful insights while significantly reducing computation time.

Key observations included:

* The largest connected component within the subgraph comprised 8,285 nodes.
* The average clustering coefficient for the subgraph increased to 0.33, reflecting stronger local connectivity compared to the full graph.

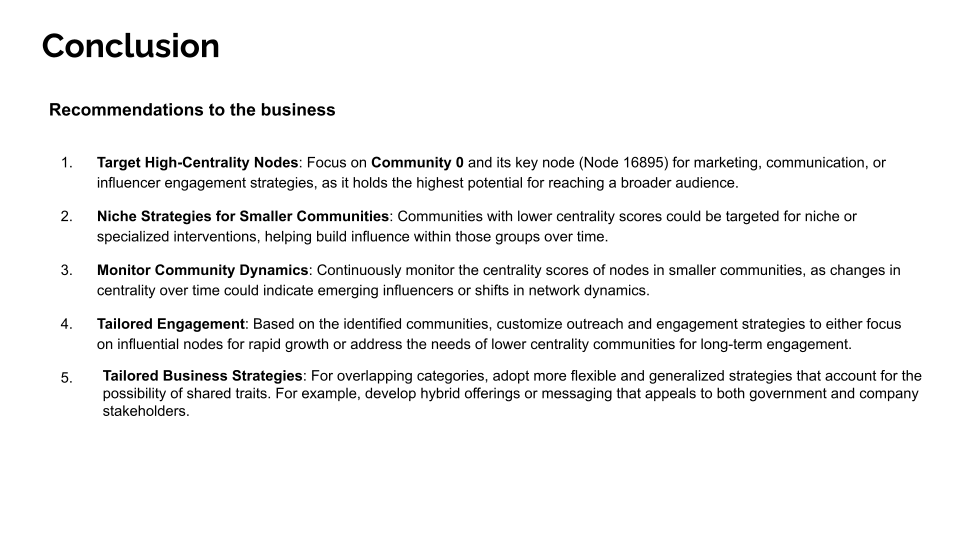
We then generated an elbow graph to determine the optimal number of clusters, testing values from to . The plot revealed a significant drop in distortion values between and . Based on this result, we selected the optimal number of clusters for further analysis.

****

**Slide 12: Spectral Clustering Results**

Using spectral clustering on the largest connected component of the subgraph, we successfully identified 10 clusters. Among these, Community 0 emerged as the most influential, with an average centrality score of 0.0007. The most influential node within Community 0 exhibited a centrality of 0.0316.

The analysis revealed significant variations in centrality scores across the identified communities. Some clusters exhibited very low centrality, indicating minimal influence, while others, like Community 0, demonstrated high centrality values, signifying greater influence. This variation underscores the heterogeneous nature of influence within the Facebook graph network and highlights the importance of tailored community-specific strategies in network analysis.



**Slide 17:**

Based on our analysis, we recommend the following business strategies:

**Target High-Centrality Nodes**: Focus on Community 0 and its key node (Node 16895) for marketing and influencer engagement, as it has the highest potential to reach a broad audience.

**Niche Strategies for Smaller Communities**: Communities with lower centrality scores could be targeted for specialized interventions to gradually build influence.

**Monitor Community Dynamics**: Track changes in centrality scores to identify emerging influencers and shifts in community influence.

**Tailored Engagement**: Customize outreach strategies for each community based on their centrality and influence levels to optimize engagement.

This study demonstrates the power of graph analysis in uncovering community structures, influence distributions, and feature relationships. By addressing computational challenges and leveraging advanced techniques, we provide actionable insights that can inform practical applications in social network analysis.

Shreyas draft

## 

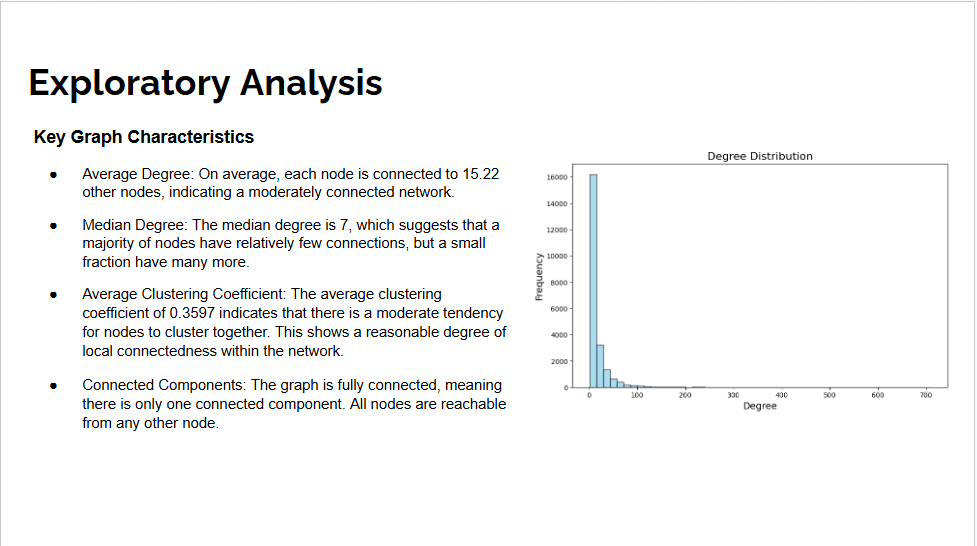
## 

## **1. Introduction**

The analysis of social networks provides valuable insights into community interactions and influence distributions. This study examines a publicly available graph of verified Facebook pages, where nodes represent pages, and edges represent mutual likes. Our goal is to analyze the graph’s structure and interactions to extract feature embeddings for machine learning tasks such as clustering and dimensionality reduction. By addressing key challenges, including computational complexity and low-connected nodes, we aim to provide a comprehensive understanding of the graph’s properties and potential applications.

## **2. Dataset Description**

The dataset consists of 22,470 nodes and 171,002 edges. Each node represents a verified Facebook page, and each edge represents a mutual “like” relationship. The graph’s fully connected nature ensures that all nodes are reachable from any other node, making it suitable for comprehensive analysis. The dataset also includes category labels for pages, enabling feature-category relationship analysis.



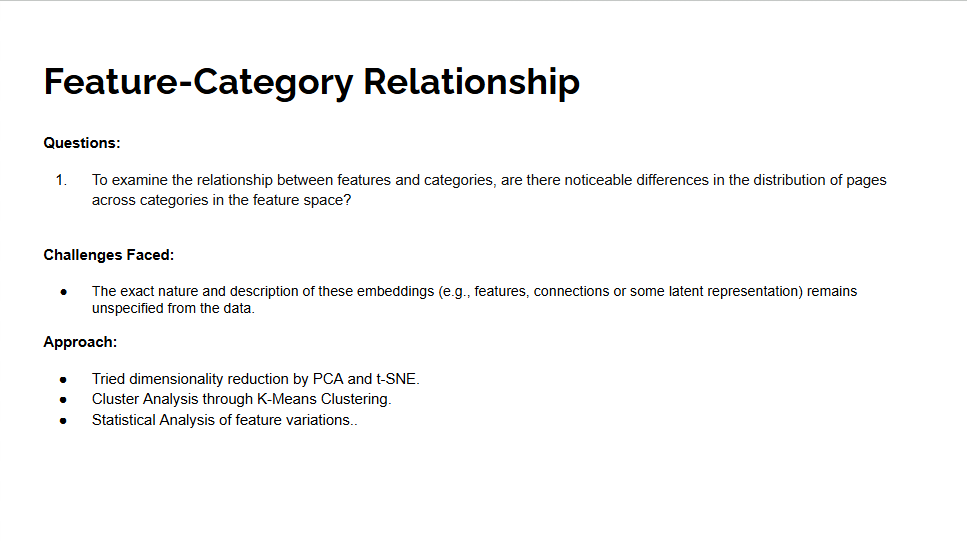
## 

## 

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

We performed an exploratory data analysis (EDA) to understand the graph's basic properties. The average node degree is 15.22, suggesting moderate connectivity, while the median degree of 7 indicates that most nodes have fewer connections. The average clustering coefficient of 0.36 shows that there is moderate local connectivity, and importantly, the graph is fully connected, meaning all nodes are reachable from any other node.

### 

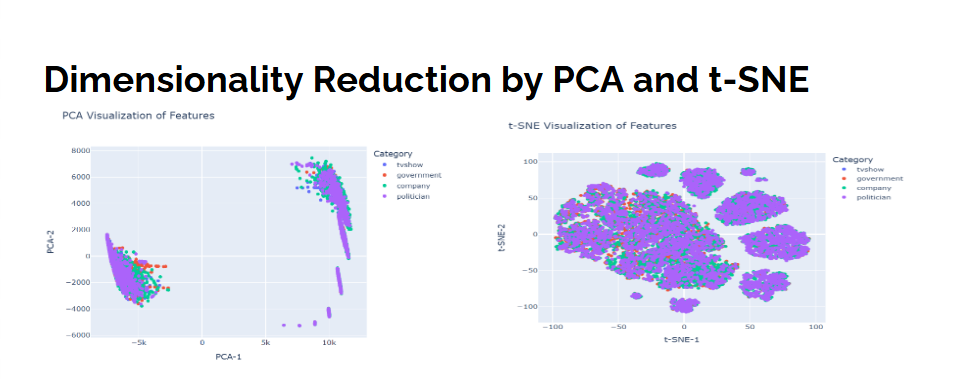
****

**4.**  **Feature-Category Relationship**

We examined feature-category analysis by addressing one key question:

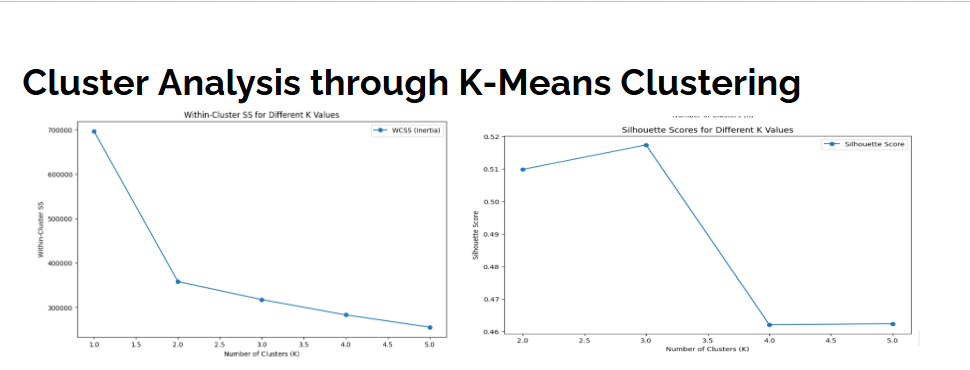
1. To examine the relationship between features and categories, are there noticeable differences in the distribution of pages across categories in the feature space?

The main challenge in this analysis lies in the lack of clarity regarding the exact nature and interpretation of these embeddings, which could be specific features, relationships, or latent representations they capture from the data.

****

**5.** **Dimensionality Reduction by PCA and t-SNE**

We focussed on implementing PCA and t-SNE plots with a number of components equal to 2. From the PCA visualization, it appears that the clusters of data points overlap significantly across categories. This is because PCA may fail to capture non-linear relationships in the data, leading to overlapping clusters when the features have complex dependencies. The t-SNE plot shows tighter and more distinct clusters compared to PCA. From the t-SNE plot, the purple category(Politician) seems to form tightest clusters. This suggests that its features are well-represented and distinct, with less overlap with other categories. Other categories, such as red category (government) and green category (company), seem more dispersed and intermixed and overlap significantly in the feature space.

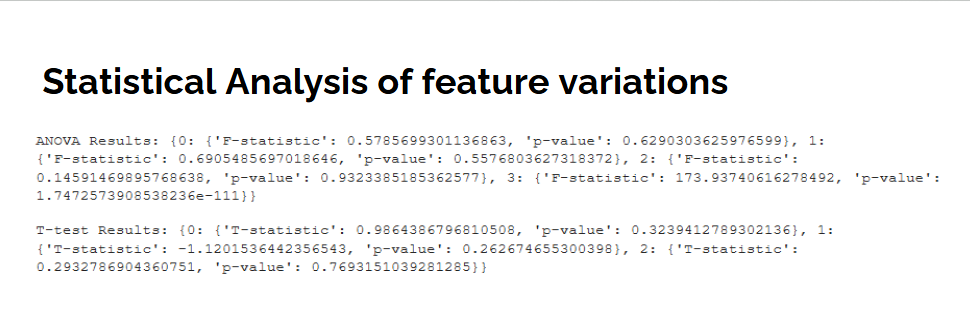
****

**6.** **Cluster Analysis through K-Means Clustering**

We also implement K-Means clustering with the number of clusters equal to 5 and max number of iterations equal to 100. We compute within cluster sum of squares and silhouette scores for each iteration and also implement their plots.

The "Within-Cluster SS (WCSS)" plot is used to identify the optimal number of clusters (K). Through the elbow method, we see a sharp decline in WCSS between K=1 and K=2, followed by a more gradual reduction, indicating that three clusters may effectively capture the structure of the data. This suggests meaningful groupings in the feature space and also potentially representing distinct category patterns.

Similarly, the “Silhouette” plot evaluates the quality of clustering. From the Silhouette plot, we find that K=3 achieved the highest score, indicating well-defined clusters. Beyond K=3, the drop in the score suggests that adding more clusters results in overlapping or less distinct groupings.

****

**7.** **Statistical Analysis of feature variations**

We perform ANOVA and T-tests between two categories: TV-Show and Government.

From the ANOVA test, features with high F-statistics and low p-value indicates significant differences in the distribution of the feature across categories such as Feature 3, which has a F-statistic of 173.93 and p-value of 1.74\*10^(-111). From the T-tests, features with high T-statistics and low p-value indicate a large difference between two groups for a given feature. From the above figure, we have no features which have high T-statistic and low p-value.

## 

## 

## 

## 

## 

## 

## 

## 

final version

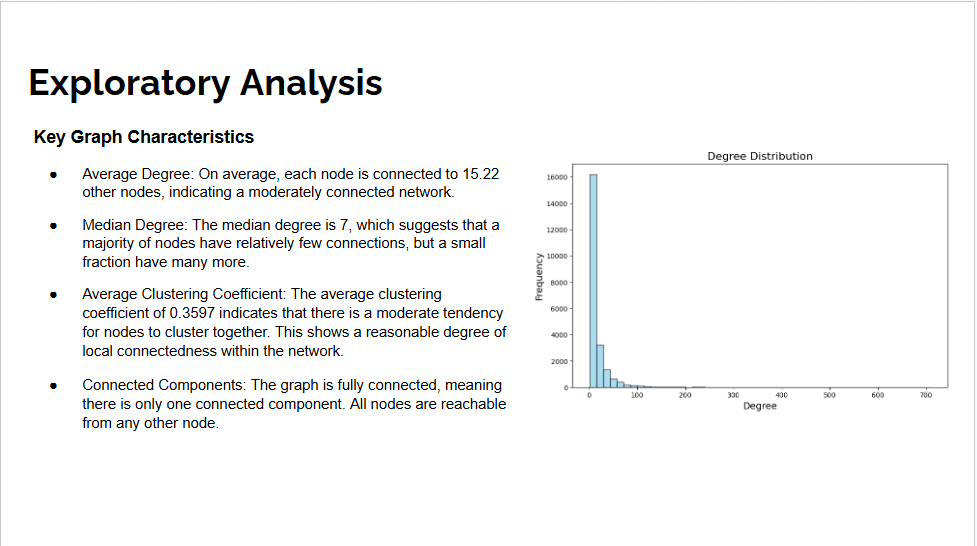
## **Project Report**

## 

## **Introduction & Dataset Description**

The analysis of social networks provides valuable insights into community interactions and influence distributions. This study examines a publicly available graph of verified Facebook pages, where nodes represent pages, and edges represent mutual likes. Our goal is to analyze the graph’s structure and interactions to extract feature embeddings for machine learning tasks such as clustering and dimensionality reduction. By addressing key challenges, including computational complexity and low-connected nodes, we aim to provide a comprehensive understanding of the graph’s properties and potential applications.

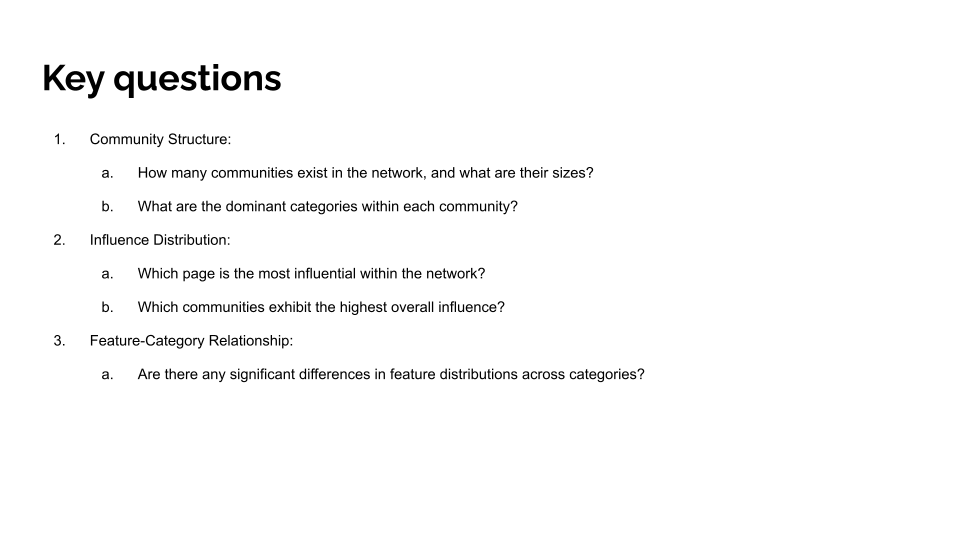
The dataset consists of 22,470 nodes and 171,002 edges. Each node represents a verified Facebook page, and each edge represents a mutual “like” relationship. The graph’s fully connected nature ensures that all nodes are reachable from any other node, making it suitable for comprehensive analysis. The dataset also includes category labels for pages, enabling feature-category relationship analysis.



### **Exploratory Data Analysis (EDA)**

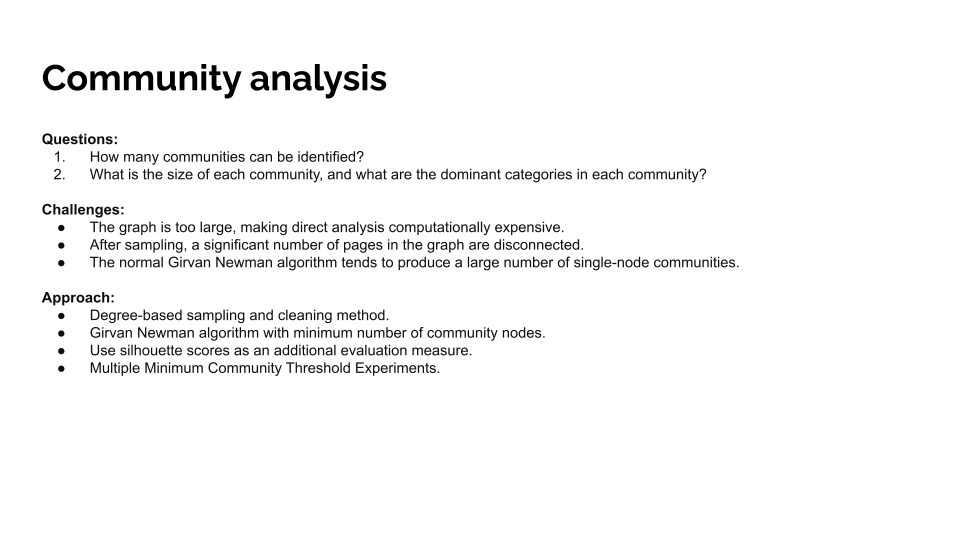
We performed an exploratory data analysis (EDA) to understand the graph's basic properties. The average node degree is 15.22, suggesting moderate connectivity, while the median degree of 7 indicates that most nodes have fewer connections. The average clustering coefficient of 0.36 shows that there is moderate local connectivity, and importantly, the graph is fully connected, meaning all nodes are reachable from any other node.

### 

****

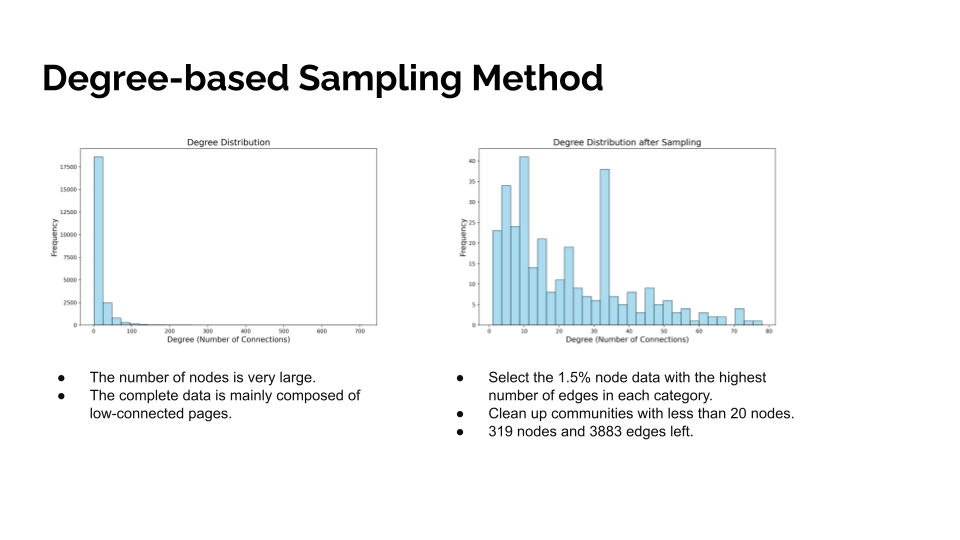
**Key Questions**

By framing the project around three key questions, we clarified the main focus areas, exploring Facebook pages’ community structure, influence distribution, and feature relationships. These questions are significant as they not only unveil the overall network structure but also provide targeted guidance for business decision-making. Analyzing community structure offers insights into how nodes are organized within the network and the distribution patterns of different types of nodes. Influence distribution analysis further identifies key nodes and communities that play critical roles in the network, providing a basis for enterprises to formulate precise marketing and resource allocation strategies. Finally, studying the relationship between features and categories helps distinguish characteristics across different categories, offering data support for market segmentation and personalized marketing strategies. The analysis framework is progressive, with each question building on the previous, ultimately delivering a comprehensive view for business applications.

**Community Detection Question**

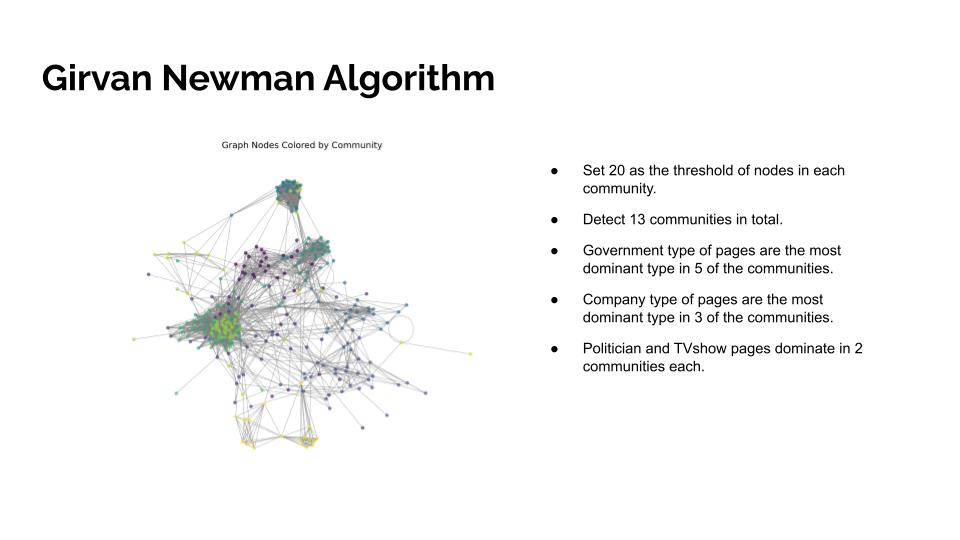
When analyzing the network’s community structure, we faced three major challenges. First, the graph is extremely large, consisting of 22,470 nodes and 171,002 edges, which demands significant computational resources to analyze the entire structure. In fact, we initially attempted to analyze the full graph, but due to the excessive runtime, we had to explore alternative methods. Second, many nodes have low connectivity, which contributes minimally to the community structure analysis and can even distort results by affecting the division of highly connected nodes. Lastly, the traditional Girvan-Newman algorithm tends to produce single-node communities, which contradicts our goal of analyzing overall community characteristics.

To address these challenges, we designed a degree-based sampling method, focusing on high-degree nodes to significantly reduce data size and improve analysis efficiency and effectiveness. Additionally, we modified the Girvan-Newman algorithm by introducing a node count threshold to avoid generating numerous insignificant small communities. We conducted experiments using thresholds of 10, 20, 30, 40, and 50 nodes per community. To ensure robust evaluation, we used both modularity and silhouette scores, as relying solely on modularity may introduce bias. After comparison, a threshold of 20 nodes produced the best overall results for modularity and silhouette scores. This method not only improved computational efficiency but also ensured the reliability and interpretability of the results.

**Degree-based Sampling Method**

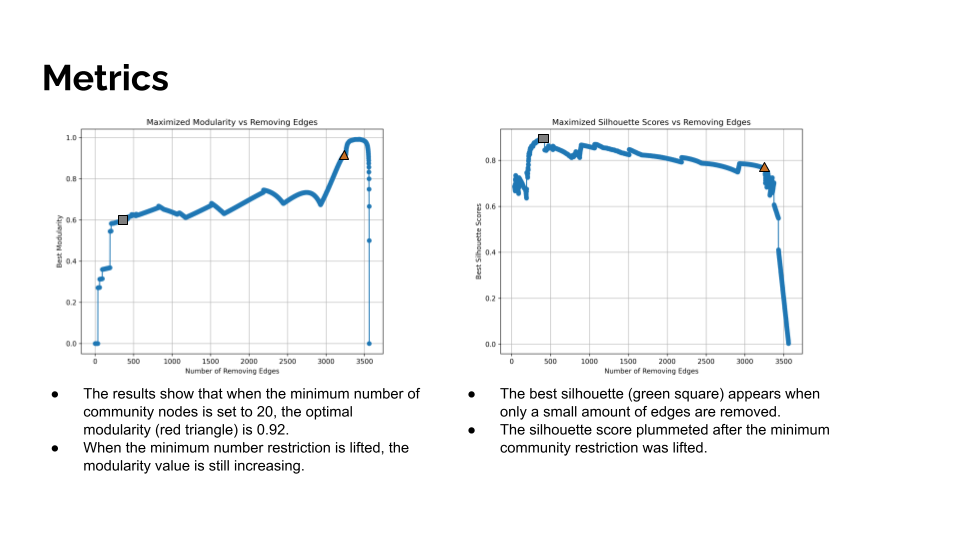
To effectively address the issue of the graph’s large scale, we adopted a degree-based sampling strategy during the data reduction phase, selecting the top 1.5% of nodes with the highest degree in each category. This method focuses on the key nodes that have the greatest impact on the network structure, significantly reducing data complexity while maintaining the original distribution of categories. We then set a minimum community size threshold of 20 nodes to filter out small, insignificant communities.

This process reduced the original dataset from 22,470 nodes and 171,002 edges to 319 nodes and 3,883 edges. It not only decreased the computational burden but also established a solid foundation for subsequent community detection. By preserving the category distribution in the original data and retaining the most critical node information, this sampling approach is particularly suitable for the structural analysis of large-scale networks.

**Modified Version of Girvan-Newman Algorithm**

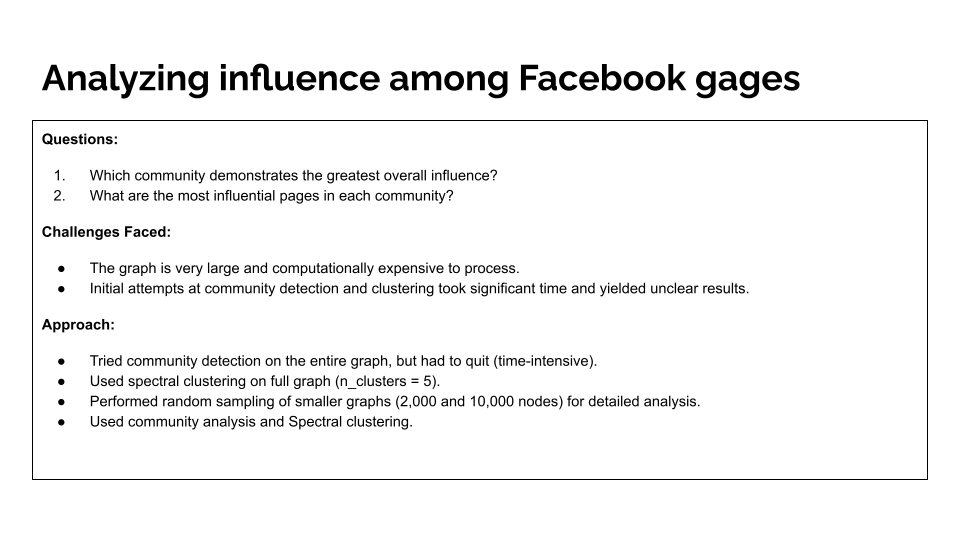
To enable the Girvan-Newman algorithm to detect communities with sufficient size and representativeness, we set a minimum threshold of 20 nodes per community. Under this restriction, modularity was used as a metric to identify the optimal community structure. Once no more edges could be removed, the restriction was lifted, but the best community structure was not updated further until all edges were removed.

Using this improved Girvan-Newman algorithm, we identified 13 significant communities within the network. Among these, five communities were dominated by government pages, highlighting their prominence; three communities were dominated by company pages, reflecting the importance of business nodes. Other types, such as politicians and TV shows, dominated two communities each. Additionally, the dominant category in each community accounted for more than 50% of its pages. This analysis not only reveals the concentrated distribution of categories within the network and the strong association between communities and categories but also provides critical insights into the interactions and distributions among communities. These findings offer a reliable foundation for further exploration of the relationship patterns and potential value within social networks.

**Metrics for Evaluation**

To validate the effectiveness of community detection, we introduced two key metrics: modularity and silhouette scores. Modularity measures the compactness and rationality of community divisions, achieving its optimal value of 0.92, represented by the red triangle in the figure. This result indicates a well-structured community division. Silhouette scores reflect the separation between communities and the cohesion within them, achieving the best score when only a small number of edges were removed. Notably, the silhouette score remained relatively high throughout the process.

After removing the node count restriction, modularity initially increased slightly but then dropped rapidly, while the silhouette score plummeted to 0, indicating blurred community boundaries and significantly reduced quality of division. These results demonstrate that setting a reasonable node count threshold not only improves computational efficiency but also ensures the scientific validity and interpretability of community detection. The evaluation results provide strong evidence for the rationality of the algorithm design.

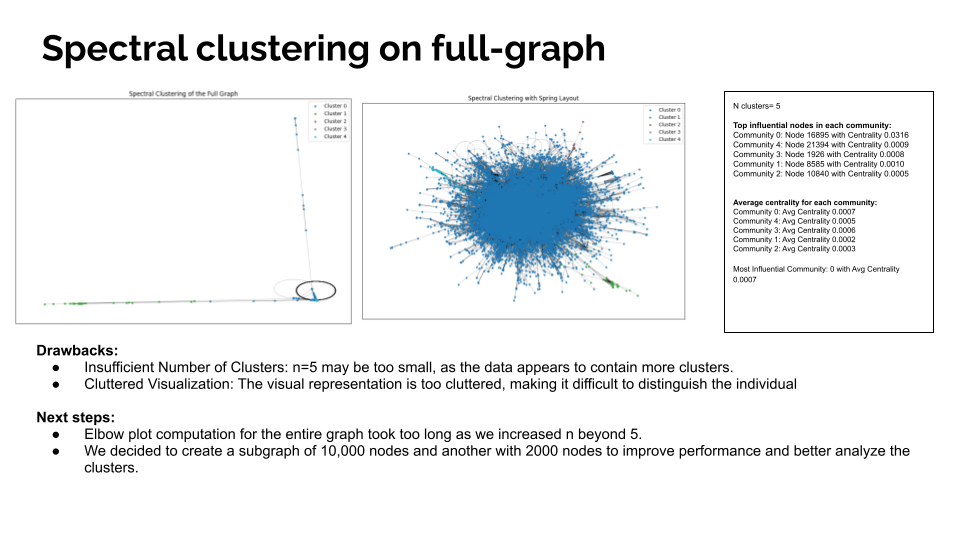
****

**Influence Analysis**

In this segment of the project, we focused on exploring the concept of influence within the Facebook graph network. The analysis was structured around three guiding questions:

1. Which community demonstrates the greatest overall influence?
2. What are the most influential pages within each community?

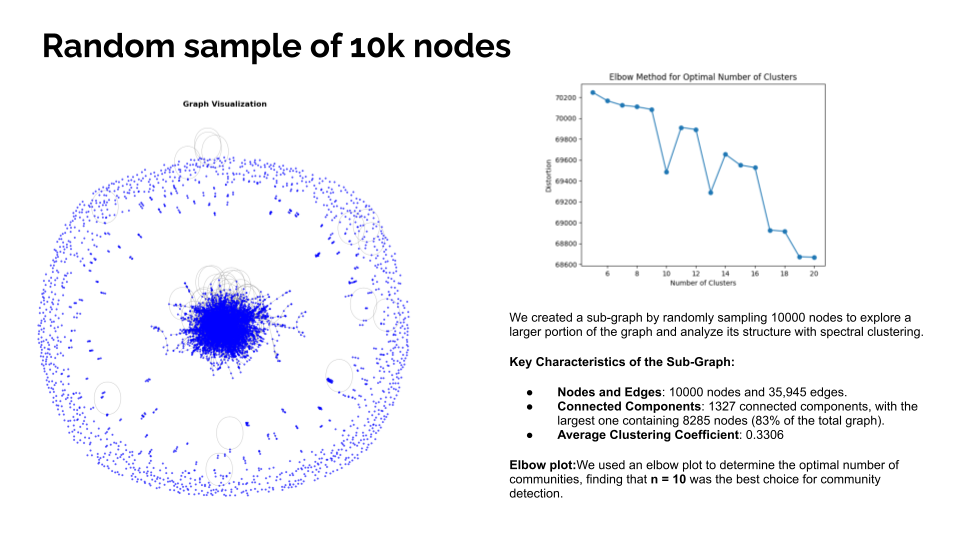
To tackle these questions, we attempted to assess influence across the graph’s structure. However, the analysis was not without challenges. The graph's vast size posed significant computational hurdles. Processing the graph to detect communities and clusters required extensive time and resources, often leading to unclear or suboptimal results in our initial attempts. This necessitated iterative refinement of our methods to manage the computational complexity effectively and derive meaningful insights.

****

**Spectral Analysis**

Our initial approach involved performing spectral clustering on the full graph, starting with five clusters. This method provided some preliminary insights into the community structure. However, we quickly recognized that using only five clusters was insufficient to capture the graph's complexity. The resulting visualization was overly cluttered, making it difficult to discern clear patterns or interpret the data meaningfully.

To refine our approach, we attempted to generate an elbow plot for a higher number of clusters. However, this process proved computationally expensive and time-consuming due to the graph's size. Consequently, we shifted our strategy towards creating subgraphs, allowing us to improve computational performance and analyze the clusters more effectively.

****

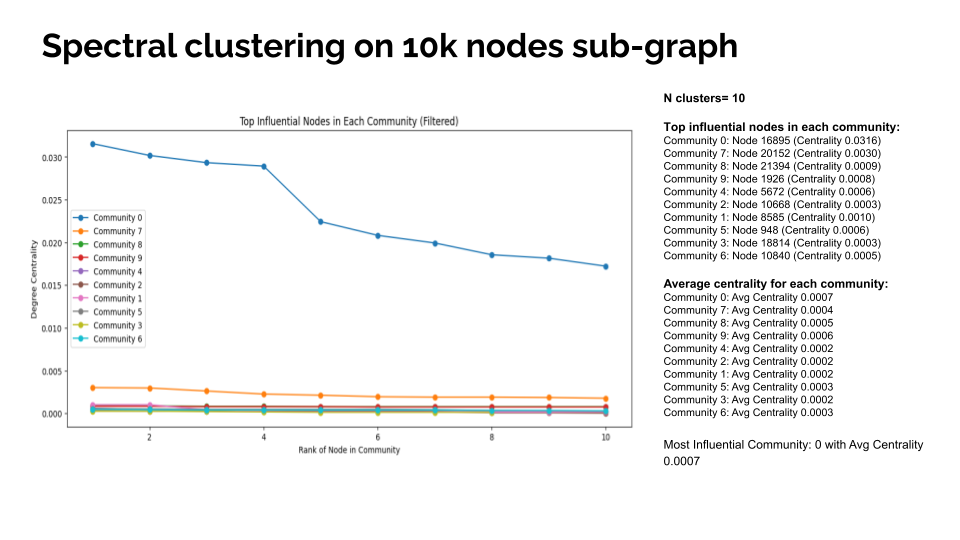
**Subgraph Creation and Cluster Analysis**

To address computational challenges, we created a subgraph by sampling 10,000 nodes from the original graph. This subgraph contained 35,945 edges and 1,327 connected components. While random sampling introduced some disconnected nodes, the subgraph retained enough data to provide meaningful insights while significantly reducing computation time.

Key observations included:

* The largest connected component within the subgraph comprised 8,285 nodes.
* The average clustering coefficient for the subgraph increased to 0.33, reflecting stronger local connectivity compared to the full graph.

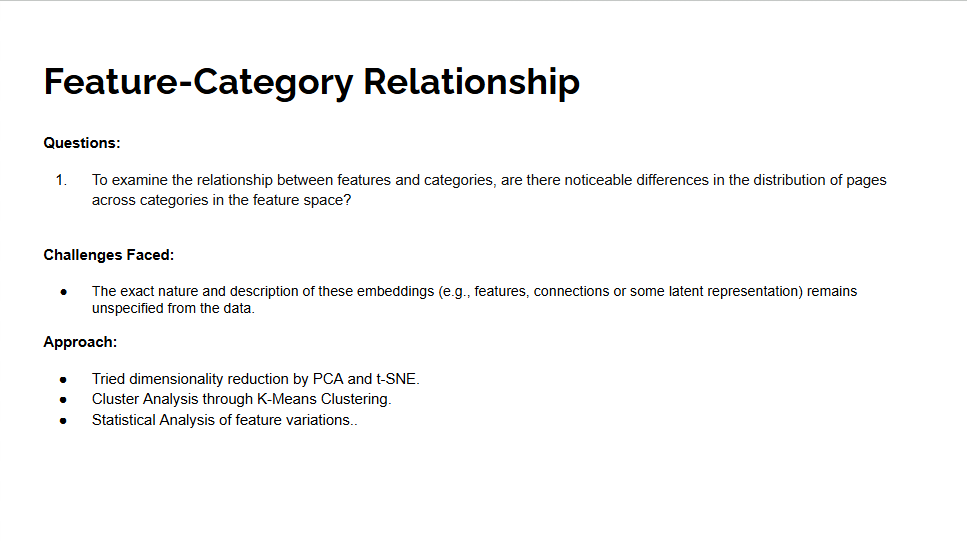
We then generated an elbow graph to determine the optimal number of clusters, testing values from to . The plot revealed a significant drop in distortion values between and . Based on this result, we selected the optimal number of clusters for further analysis.

****

**Spectral Clustering Results**

Using spectral clustering on the largest connected component of the subgraph, we successfully identified 10 clusters. Among these, Community 0 emerged as the most influential, with an average centrality score of 0.0007. The most influential node within Community 0 exhibited a centrality of 0.0316.

The analysis revealed significant variations in centrality scores across the identified communities. Some clusters exhibited very low centrality, indicating minimal influence, while others, like Community 0, demonstrated high centrality values, signifying greater influence. This variation underscores the heterogeneous nature of influence within the Facebook graph network and highlights the importance of tailored community-specific strategies in network analysis.

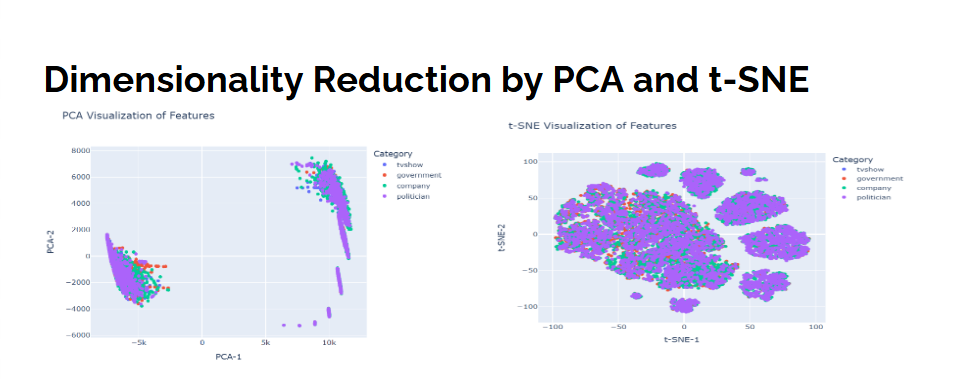
****

**Feature-Category Relationship**

We examined feature-category analysis by addressing one key question:

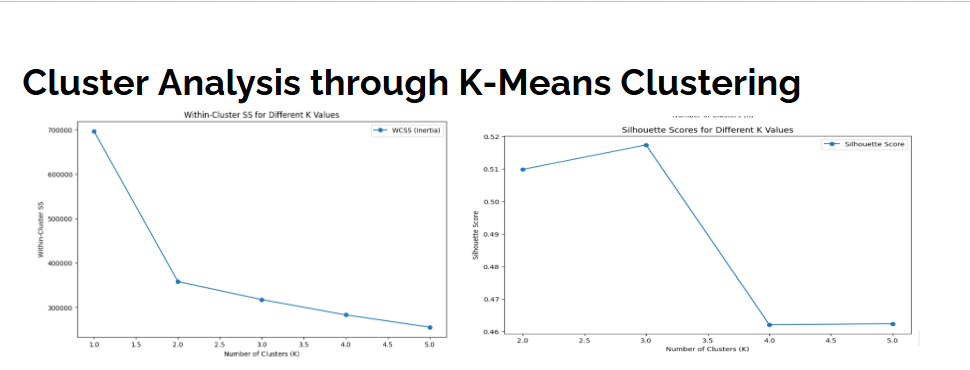
1. To examine the relationship between features and categories, are there noticeable differences in the distribution of pages across categories in the feature space?

The main challenge in this analysis lies in the lack of clarity regarding the exact nature and interpretation of these embeddings, which could be specific features, relationships, or latent representations they capture from the data.

****

**Dimensionality Reduction by PCA and t-SNE**

We focussed on implementing PCA and t-SNE plots with a number of components equal to 2. From the PCA visualization, it appears that the clusters of data points overlap significantly across categories. This is because PCA may fail to capture non-linear relationships in the data, leading to overlapping clusters when the features have complex dependencies. The t-SNE plot shows tighter and more distinct clusters compared to PCA. From the t-SNE plot, the purple category(Politician) seems to form tightest clusters. This suggests that its features are well-represented and distinct, with less overlap with other categories. Other categories, such as red category (government) and green category (company), seem more dispersed and intermixed and overlap significantly in the feature space.

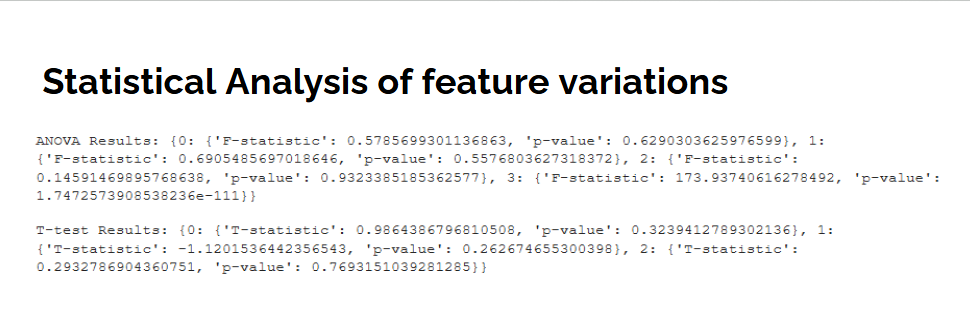
****

**Cluster Analysis through K-Means Clustering**

We also implement K-Means clustering with the number of clusters equal to 5 and max number of iterations equal to 100. We compute within cluster sum of squares and silhouette scores for each iteration and also implement their plots.

The "Within-Cluster SS (WCSS)" plot is used to identify the optimal number of clusters (K). Through the elbow method, we see a sharp decline in WCSS between K=1 and K=2, followed by a more gradual reduction, indicating that three clusters may effectively capture the structure of the data. This suggests meaningful groupings in the feature space and also potentially representing distinct category patterns.

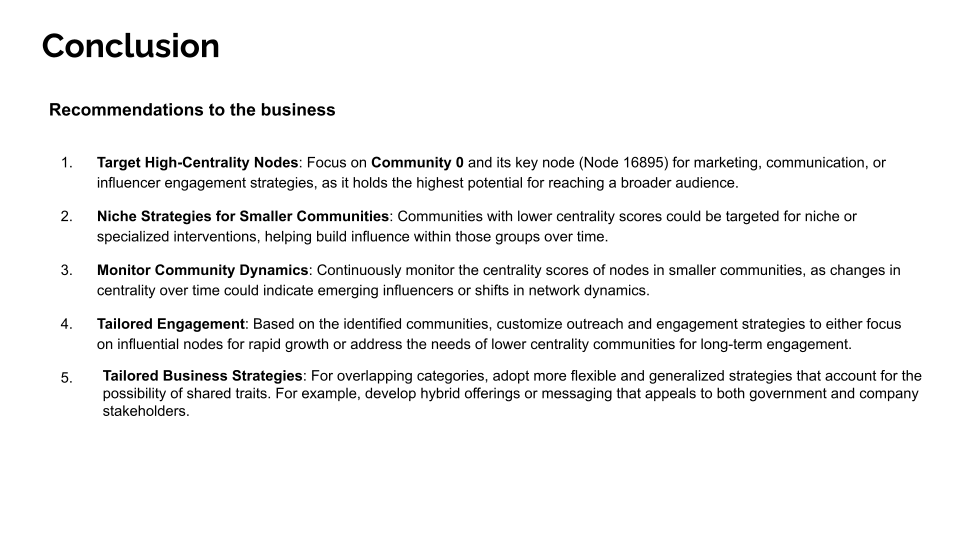
Similarly, the “Silhouette” plot evaluates the quality of clustering. From the Silhouette plot, we find that K=3 achieved the highest score, indicating well-defined clusters. Beyond K=3, the drop in the score suggests that adding more clusters results in overlapping or less distinct groupings.

****

**Statistical Analysis of feature variations**

We perform ANOVA and T-tests between two categories: TV-Show and Government.

From the ANOVA test, features with high F-statistics and low p-value indicates significant differences in the distribution of the feature across categories such as Feature 3, which has a F-statistic of 173.93 and p-value of 1.74\*10^(-111). From the T-tests, features with high T-statistics and low p-value indicate a large difference between two groups for a given feature. From the above figure, we have no features which have high T-statistic and low p-value.



**Conclusion**

Based on our analysis, we recommend the following business strategies:

**Target High-Centrality Nodes**: Focus on Community 0 and its key node (Node 16895) for marketing and influencer engagement, as it has the highest potential to reach a broad audience.

**Niche Strategies for Smaller Communities**: Communities with lower centrality scores could be targeted for specialized interventions to gradually build influence.

**Monitor Community Dynamics**: Track changes in centrality scores to identify emerging influencers and shifts in community influence.

**Tailored Engagement**: Customize outreach strategies for each community based on their centrality and influence levels to optimize engagement.

This study demonstrates the power of graph analysis in uncovering community structures, influence distributions, and feature relationships. By addressing computational challenges and leveraging advanced techniques, we provide actionable insights that can inform practical applications in social network analysis.