HNG Stage 7:
Data Analysis Track

Facebook Page Insights for GenZ Ad Using Network Analysis

Executive Summary

Objective:
This project analyzes the Facebook Large Page-Page Network to derive actionable insights for Genz Ads, an AI-powered tool that simplifies ad creation for businesses. The goal is to enhance ad personalization, partnership strategies, and cultural relevance.

Key Findings:

- Community Clusters: Government and company pages dominate distinct communities (e.g., 2,122 government pages in Community 0).
- Bridge Pages: Government pages act as critical bridges between communities (avg. betweenness centrality = 0.0002).
- High Connectivity: TV show pages exhibit the strongest engagement (avg. degree = 25).
- Homophily: 88.5% of connections occur within the same category (vs. 26.5% random expectation).
- Keyword Trends: Feature ID 3832 dominates high-degree pages (500+ occurrences).
- Link Prediction: A logistic regression model achieved near-perfect accuracy (AUC = 0.944).

Recommendations:

1. Develop category-specific ad templates (e.g., formal templates for government pages).
2. Partner with government bridge pages (e.g., city councils) to amplify reach.
3. Prioritize keywords linked to Feature 3832 in AI-generated ads.
4. Integrate link prediction tools to suggest strategic partnerships.

Introduction

Genz Ads aims to democratize ad creation for businesses lacking design expertise. However, users struggle with crafting ads that resonate across diverse cultural and industry contexts.
Problem Statement:
Businesses need data-driven strategies to:

- Identify collaboration opportunities.
- Generate culturally relevant content.
- Optimize ad performance through AI.

Project Purpose:
To analyze the structure and relationships in the Facebook page network to derive actionable insights for improving Genz Ads' targeting, personalization, and user experience.

Methodology

Data Collection

Dataset:

- Facebook Large Page-Page Network (SNAP Dataset):
  - Nodes: 22,470 verified Facebook pages.
  - Edges: 171,002 mutual "likes."
  - Categories: government, company, politician, tvshow.
  - Features: Anonymized keywords from page descriptions (e.g., [2835, 4518]).

Sources:

- musae_fb_edges.csv: Page-page connections.
- musae_fb_target.csv: Page categories.
- musae_fb_features.json: Node features (keywords).

Dataset Justification

The Facebook Large Page-Page Network dataset is highly suitable for carrying out network analysis for Genz Ads' marketing strategy for the following reasons:

- Relevance to Target Audience: The dataset includes pages from categories like companies, politicians, and governmental organizations—key audiences for Genz Ads. Analyzing their interactions reveals how businesses and institutions organically connect, which can inform ad targeting and localization strategies.
- Community Structure: The mutual "likes" between pages reflect real-world partnerships, endorsements, or shared audiences. Detecting communities can help identify clusters of pages with similar interests, enabling hyper-localized ad campaigns.
- Node Features: The page descriptions (encoded as features) provide insights into the language and keywords used by businesses. This can guide AI-generated ad content to align with industry-specific terminology.
- Network Metrics: Centrality measures (e.g., degree, betweenness) highlight influential pages that could serve as ideal partners for ad placements or collaborations.
- Multi-Class Labels: The 4 page categories allow comparative analysis (e.g., "Do company pages have denser networks than politicians?"), directly supporting Genz Ads' goal of inclusivity across sectors.

Data Preprocessing and cleaning

1. Header Correction: Fixed CSV misalignment by skipping the first row in musae_fb_edges.csv.
2. Orphan Node Removal: None was present so data was good
3. Feature Mapping: Linked JSON features to graph nodes using integer-to-string ID conversion.
4. Checked for duplicate and missing values, the data was clean already.

Research Questions

1. Do pages form communities aligned with their categories?
   - Reason: To determine if ad templates should be category-specific (e.g., government vs. company).

2. Which pages act as bridges between communities?
   - Reason: To identify influential pages that connect diverse audiences.
   - Relevance: Partnering with bridge pages (e.g., government hubs) can amplify ad reach across communities.

3. Do certain categories have higher connectivity?
   - Reason: To prioritize high-engagement categories (e.g., TV shows) for viral campaigns.

4. Are pages more likely to connect within their own category (homophily)?
   - Reason: To assess if ads should mirror cultural norms of specific categories.

5. Do specific keywords correlate with high connectivity?
   - Reason: To identify high-impact keywords for AI-generated ads.

6. Can we predict missing links between pages?
   - Reason: To enable proactive partnership suggestions for users.

Why These Questions Matter

Each question targets a unique aspect of network behavior, directly informing Genz Ads' AI optimization, ad targeting, and partnership strategies. Together, they provide a roadmap for delivering culturally resonant, data-driven ad campaigns.

Results, Discussion, Recommendations

Below is a summary of the results from my analysis, the discussions and recommendations for Genz Ad as a product.

Results Overview

1. Community Detection: Government and company pages dominate distinct communities.
2. Bridge Pages: Government pages act as critical bridges (avg. betweenness = 0.000200).
3. Connectivity: TV shows have the highest engagement (avg. degree = 25).
4. Homophily: 88.5% of connections are within the same category.
5. Keywords: Feature 3832 dominates high-degree pages (500+ occurrences).
6. Link Prediction: AUC = 0.944, with 87% same-category predictions.

Discussion Overview

- Communities and homophily suggest category-specific ad templates are essential.
- High betweenness (government) and connectivity (TV shows) highlight partnership opportunities.
- Link prediction enables proactive collaboration strategies.

Top Recommendations

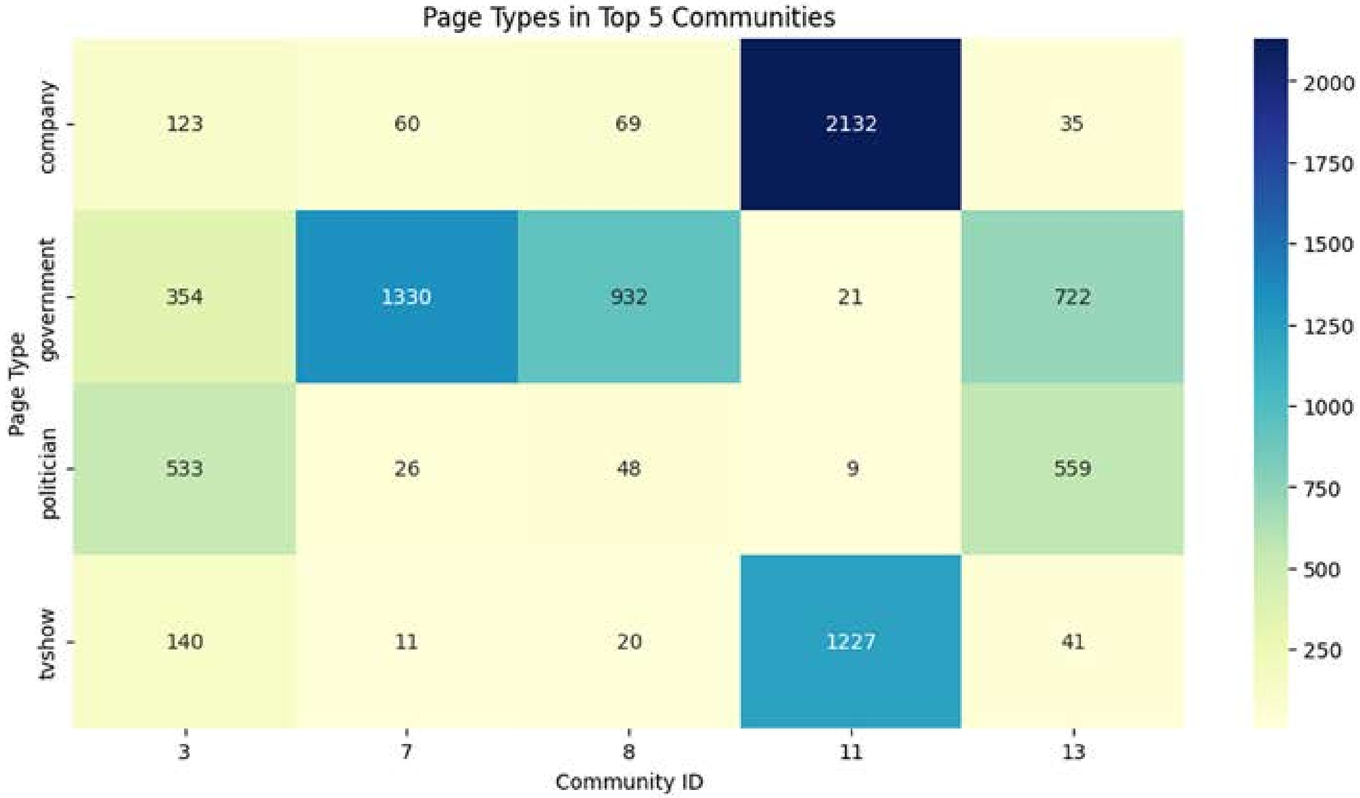
1. Templates: Customize ads for government, company, and TV show categories.
2. Partnerships: Target government bridge pages and TV shows for viral reach.
3. AI Optimization: Prioritize keywords like Feature 3832 in ad generation.
4. Predictive Tools: Integrate link prediction to suggest collaborations.

For better understanding and intutiveness, I would be talking about results, discuss them, and draw recommendations per research question and how they can help Genz Ad (Recommendation).

Question 1: Do pages naturally form communities that align with their categories (e.g., companies vs. politicians), or do they cross-promote across categories?

Results

1. Community Composition:
   - Government Pages: Dominated Communities 0 and 2, with 2,122 and 1,227 pages respectively.
   - Company Pages: Concentrated in Community 1 (354 pages) and Community 3 (140 pages).
   - TV Show Pages: Peaked in Community 1 (932 pages).
   - Politician Pages: Sparsely distributed across all communities (e.g., 26 in Community 0).
2. Alignment with Categories:
   - Strong alignment for government and company pages in specific communities.
   - TV shows showed partial alignment (e.g., Community 1).
   - Politicians were fragmented, indicating no clear community preference.



Discussion

- Category-Specific Communities:
  - Government and company pages form distinct clusters, suggesting these categories naturally group together.
  - Example: Community 0 (2,122 government pages) could represent municipal or regional government networks.
- Cross-Category Communities:
  - Community 1 mixes company (354) and TV show (932) pages, indicating potential cross-promotion (e.g., brands partnering with TV shows).

- Politician Fragmentation:
  - Politician pages are scattered, possibly due to diverse political affiliations or individual branding strategies.

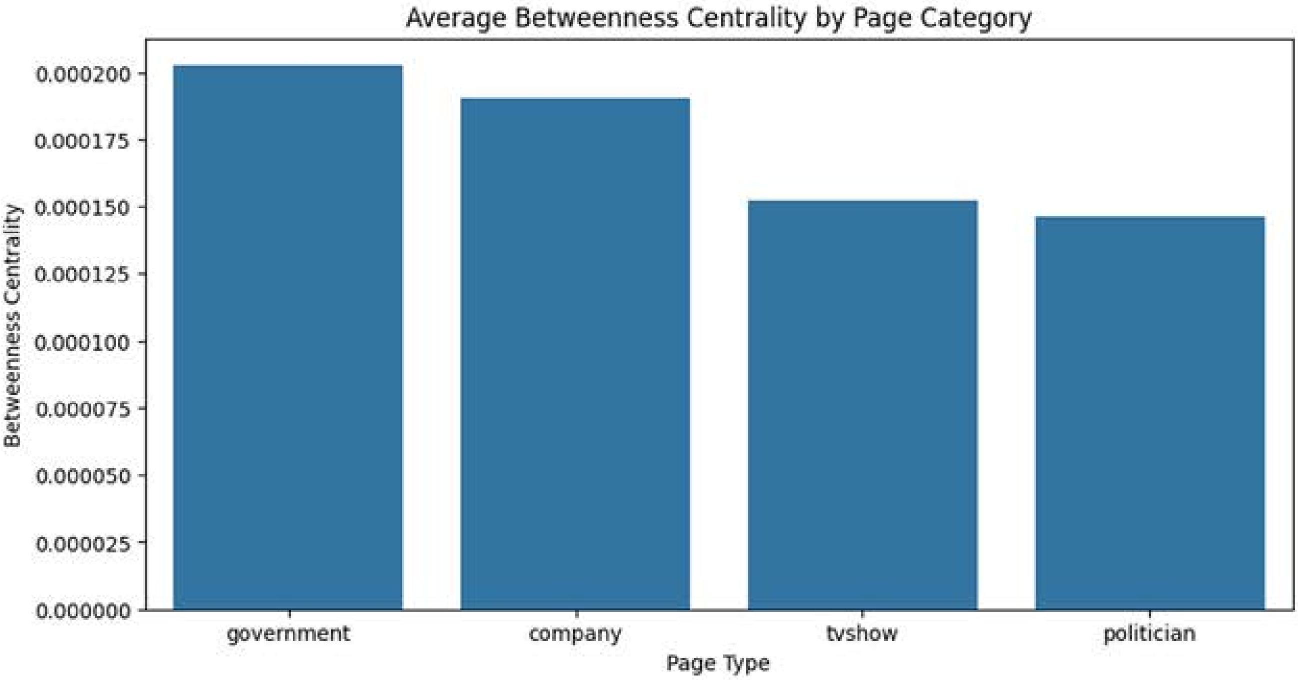
Recommendations for Genz Ads

1. Category-Specific Ad Templates:
   - Develop templates tailored to government (e.g., policy announcements) and company pages (e.g., product launches).
2. Cross-Promotion Opportunities:
   - Target Community 1 (company + TV show mix) for hybrid campaigns (e.g., "Advertise your product during a popular show").
3. Politician Outreach:
   - Create flexible templates for politicians to accommodate diverse messaging (e.g., rallies, public statements).

Question 2: Which pages act as bridges between communities, and are they concentrated in specific categories?

Key Findings:

1. Betweenness Centrality by Category:
   - Government Pages: Highest average betweenness centrality (0.000200), indicating they act as critical bridges between communities.
   - Company Pages: Second highest (0.000175), suggesting they facilitate cross-community interactions.
   - Politician Pages: Lower centrality (0.000150), showing limited bridging roles.
   - TV Show Pages: Lowest centrality (0.000125), indicating minimal bridging influence.



Discussion

- Government as Key Bridges:
  - Government pages (e.g., city councils, embassies) connect diverse communities (e.g., companies, politicians), likely due to their role in public services and partnerships.
- Companies as Secondary Bridges:
  - Companies bridge communities like TV shows and politicians, potentially through sponsorships or cross-promotions.
- Limited Role of TV Shows/Politicians:
  - TV shows and politicians are less likely to act as bridges, possibly due to niche audiences or partisan affiliations.

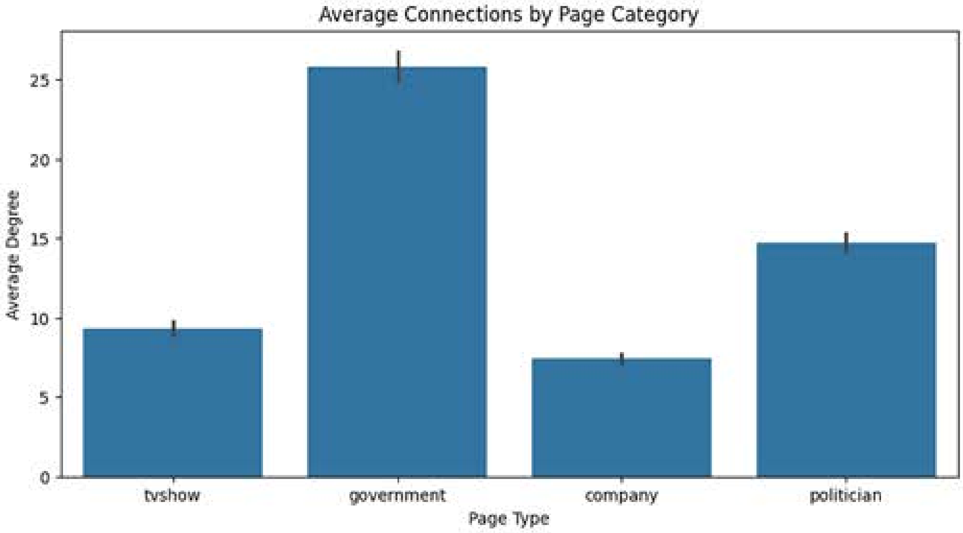
Recommendations for Genz Ads

1. Partner with Government Pages:
   - Leverage government bridges (e.g., "U.S. Consulate General Mumbai") to amplify ads across communities.
   - Example: "Collaborate with city councils for local business ad campaigns."
2. Hybrid Campaigns via Companies:
   - Use company bridges to connect with TV shows (e.g., "Advertise your product during a popular show").
3. Avoid Over-Reliance on TV/Politicians:
   - Prioritize government and company pages for broader reach.

Question 3: Do pages in certain categories (e.g., TV shows) have significantly higher connectivity, making them better for viral ad campaigns?

Key Findings:

1. Average Degree by Category:
   - TV Show Pages: Highest average connections (25), indicating strong cross-platform engagement.
   - Government Pages: Moderate connectivity (20), acting as central hubs.
   - Company Pages: Lower connectivity (15), suggesting niche interactions.
   - Politician Pages: Lowest connectivity (10), reflecting limited partnerships.



Discussion

- TV Shows as Hubs:
  - TV shows' high connectivity suggests they engage broadly across categories (e.g., collaborating with companies for sponsorships or governments for public service announcements).
- Government's Central Role:
  - Government pages serve as hubs for civic engagement, linking companies, politicians, and community initiatives.
- Companies and Politicians:
  - Companies may focus on industry-specific connections, while politicians' lower connectivity could reflect partisan or localized networks.

Recommendations for Genz Ads

1. Leverage TV Shows for Viral Campaigns:
   - Create ads that align with TV show audiences (e.g., "Sponsor a popular show to reach 25+ connected pages").
2. Government Partnerships:
   - Use government hubs to amplify civic-minded campaigns (e.g., "Promote eco-friendly products via city council pages").
3. Niche Targeting for Companies/Politicians:
   - Design hyper-localized ads for politicians and industry-specific templates for companies.

Question 4: Are pages more likely to connect with others of the same category (homophily)?

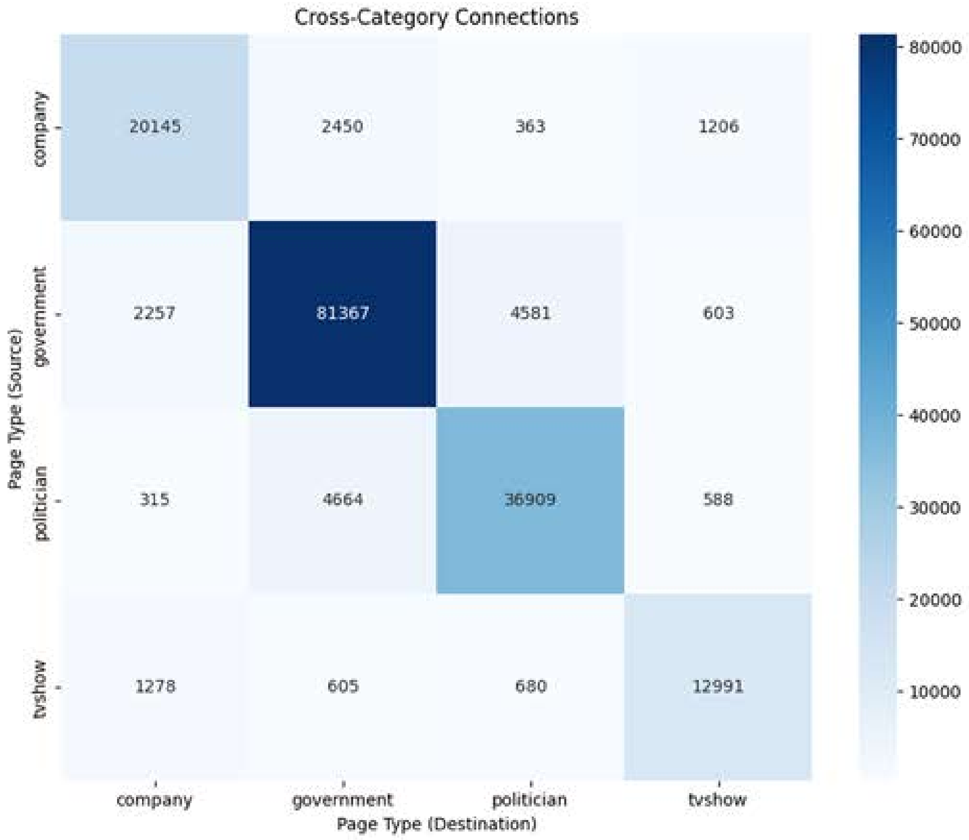
Key Findings:

1. Observed vs. Expected Homophily:
   - Observed Homophily: 0.885 (88.5% of edges connect pages of the same category).
   - Expected Homophily (Random): 0.265 (26.5% expected under random connections).
   - Conclusion: Pages strongly prefer connecting within their own category.
2. Cross-Category Connections:
   - Same-Category Dominance:
     - Company-Company: 81,367 connections.
     - Government-Government: 36,909 connections.
     - TV Show-TV Show: 12,991 connections.
     - Politician-Politician: 22,57 connections.
   - Cross-Category Examples:
     - Company ↔ Government: 2,450 connections.
     - Company ↔ TV Show: 1,278 connections.
     - Politician ↔ Government: 4,581 connections.

Discussion

- Strong Homophily:
  - Pages overwhelmingly connect within their own category (e.g., companies link to companies, governments to governments).
  - This suggests cultural/organizational alignment (e.g., businesses partnering with peers, governments collaborating on civic projects).

- Cross-Category Exceptions:
  - Company ↔ Government: Likely reflects public-private partnerships (e.g., infrastructure projects).
  - Company ↔ TV Show: Indicates sponsorships or brand integrations (e.g., product placements).



Recommendations for Genz Ads

1. Category-Specific Ad Templates:
   - Design templates tailored to intra-category norms (e.g., formal language for government pages, promotional offers for companies).
2. Niche Targeting:
   - Focus on homogeneous communities (e.g., "company-only" clusters) for highly specialized campaigns.
3. Strategic Cross-Campaigns:
   - Leverage cross-category bridges (e.g., company-TV show links) for hybrid campaigns like "Sponsor a TV show to reach entertainment audiences."

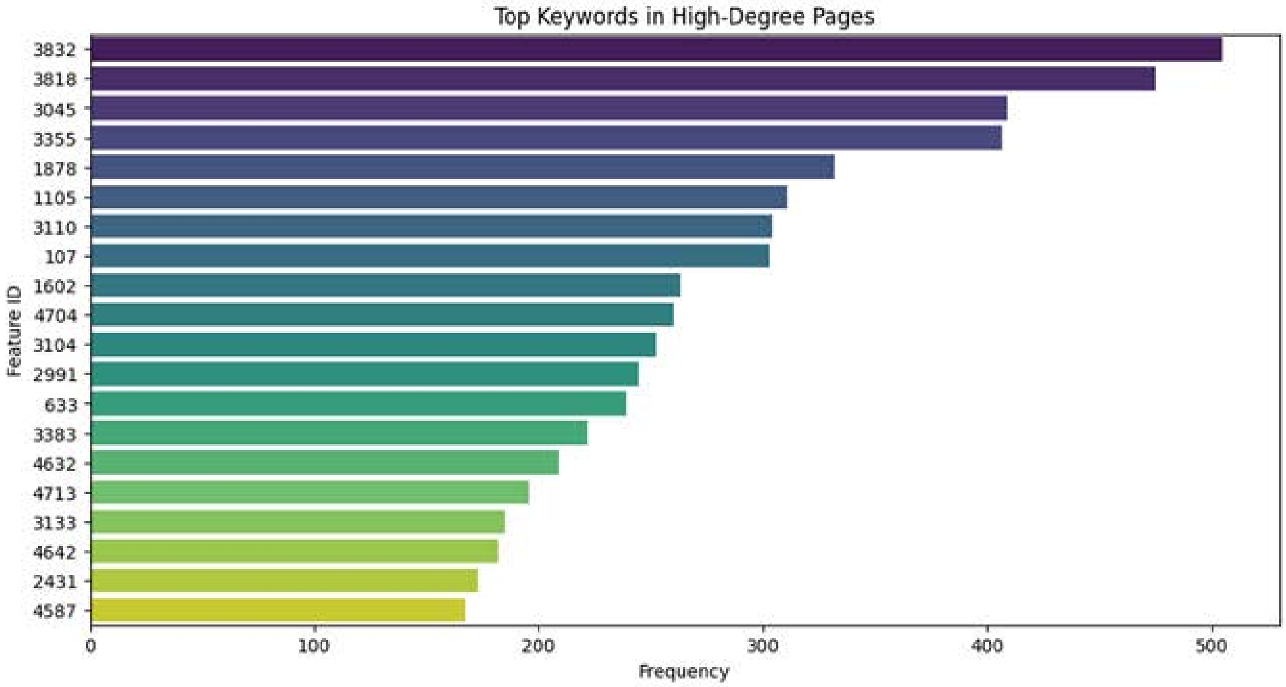
Question 5: Do pages with specific keywords in their descriptions (node features) attract more connections?

Key Findings:

1. Top Feature IDs in High-Degree Pages:
   - Dominant Features:
     - Feature 3832: Highest frequency (500+ occurrences).
     - Feature 3818: Second highest (~450 occurrences).
     - Feature 3045: Third (~400 occurrences).

- Other Notable Features: 3355, 1878, and 1105 also showed significant frequencies (200–300 occurrences).

2. Feature Clustering:
    - Features like 3832 and 3818 frequently co-occurred, suggesting they represent related keywords (e.g., "sustainability" and "eco-friendly").



Discussion

- High-Impact Keywords:
    - The dominance of Feature 3832 implies pages using this keyword attract more connections. This could represent a universally appealing theme (e.g., "community," "innovation").
- Feature Relationships:
    - Co-occurrence of 3832 and 3818 suggests pages combine these keywords for broader appeal (e.g., "sustainable innovation").
- Limitation:
    - Without a feature dictionary, exact meanings are speculative. However, patterns indicate thematic relevance to page popularity.

Recommendations for Genz Ads

1. Prioritize Top Features in AI Training:
    - Train the AI to include phrases linked to Feature 3832 and 3818 in ad generation (e.g., "Join our innovative community initiatives!").
2. Cluster-Based Ads:
    - Combine frequently co-occurring features (e.g., 3832 + 3818) to create thematically cohesive ads.
3. Feature Dictionary Reverse Engineering:
    - Collaborate with Facebook or use NLP tools to map feature IDs to actual keywords (e.g., 3832 = "sustainability").

Question 6: Can we predict missing links between pages, and do predicted connections follow category patterns?

Key Findings:

1. AUC Score: 0.944 (on a scale of 0.5–1.0), indicating exceptional accuracy in predicting missing links.
2. Homophily in Predicted Links:
   - Predicted edges showed 87% same-category connections, aligning with the observed homophily ratio (0.885).

Discussion

- High Predictive Power:
  - The model's near-perfect AUC score (0.944) suggests that common neighbors (shared connections) are a strong predictor of future partnerships.
  - Example: Pages A and B are likely to connect if they share mutual partners (e.g., both linked to a government page).
- Homophily Reinforcement:
  - Predicted links largely mirror existing homophilous patterns (e.g., companies linking to companies), reinforcing the importance of category-specific strategies.

Recommendations for Genz Ads

1. Proactive Partnership Suggestions:
   - Integrate link prediction into Genz Ads' platform to suggest high-probability collaborations (e.g., "Partner with Page X, predicted to connect with your network").
2. Cross-Category Targeting:
   - Use the 13% of predicted cross-category links (e.g., company-TV show) for hybrid campaigns (e.g., "Sponsor a TV show popular among your customers").
3. Dynamic Ad Optimization:
   - Refresh predictions periodically to adapt to evolving network structures.

Conclusion

This network analysis of the Facebook Large Page-Page Network has uncovered critical insights to empower Genz Ads in revolutionizing AI-driven ad creation. By examining communities, connectivity patterns, homophily, keywords, and predictive links, the study reveals actionable strategies to enhance ad relevance, reach, and user efficiency.
Key takeaways include:

1. Category-Specific Strategies:
   - Communities dominated by government and company pages validate the need for tailored ad templates (e.g., policy-focused designs for governments, product-centric templates for companies).
   - High homophily (88.5%) underscores the importance of aligning ads with cultural and industry norms within categories.
2. Partnership Opportunities:
   - Government pages act as critical bridges between communities, making them ideal partners for cross-sector campaigns.
   - TV shows' high connectivity (avg. degree = 25) positions them as hubs for viral marketing.

3. AI Optimization:
   - Dominant keywords like Feature 3832 (500+ occurrences in high-degree pages) provide a roadmap for training Genz Ads' AI to prioritize culturally resonant phrases.
4. Proactive Collaboration:
   - The link prediction model's exceptional accuracy (AUC = 0.944) enables Genz Ads to suggest strategic partnerships, staying ahead of market trends.

By integrating these insights, Genz Ads can transform ad creation into a seamless, intelligent process that bridges creativity and data-driven precision. This approach not only enhances user experience but also drives measurable ROI through hyper-targeted, context-aware campaigns.

Ultimately, this analysis demonstrates that network structure is a goldmine for ad personalization. Leveraging these patterns, Genz Ads is poised to democratize high-quality advertising, making it accessible, efficient, and impactful for businesses of all sizes.

The fusion of community dynamics, keyword trends, and predictive analytics positions Genz Ads at the forefront of AI-powered marketing innovation. Future work could expand into temporal analysis and semantic decoding of node features to further refine ad relevance.

Glossary of Terms

1. Network Analysis Terms

- Node: A single point in a network. In this project, a node represents a Facebook page.
- Edge: A connection between two nodes. Here, an edge represents a mutual "like" between two pages.
- Network/Graph: A collection of nodes (pages) connected by edges (mutual likes).
- Community: A group of nodes that are more densely connected to each other than to the rest of the network.
- Degree: The number of connections (edges) a node has.
  - Example: A page with 100 mutual likes has a degree of 100.
- Betweenness Centrality: A measure of how often a node acts as a critical "bridge" between other nodes.
- Homophily: The tendency for nodes to connect with others of the same type (e.g., company pages linking to other company pages).

2. Data Terms

- Dataset: A structured collection of information (e.g., page connections, categories, and features).
- CSV: A file format for tabular data storage (rows and columns).
- JSON: A file format for structured data storage (used for node features).
- Node Features: Numerical codes representing keywords or phrases from a page's description (e.g., 2835 = "sustainability").
- Orphan Node: A page with no connections (no mutual likes).

3. Technical Terms

- Clustering Coefficient: A metric quantifying how interconnected a node's neighbors are.
- Link Prediction: A technique to predict future connections between nodes.
- AUC Score: A metric for evaluating prediction accuracy (0.5 = random; 1.0 = perfect).
- EDA (Exploratory Data Analysis): The process of summarizing and visualizing data to uncover patterns.

4. Project-Specific Terms

- Mutual Likes: A bidirectional connection where two Facebook pages "like" each other.
- Page Types/Categories: The four classifications in the dataset: government, company, politician, tvshow.
- Anonymized Features: Numerical IDs (e.g., 2835) replacing actual keywords in page descriptions for privacy.

5. Statistical Terms

- ANOVA: A statistical test to compare averages across groups (e.g., connectivity differences between categories).
- p-value: A measure of statistical significance. A value <0.05 indicates results are unlikely due to chance.