Community Detection and Node Classification in Social Networks

Hemanth Vennelakanti, Kevan Dedania

May 2024

1 Introduction

Our project aims to analyze and understand the dynamics of community formation within a Facebook network by examining both structural and content-based relationships among various page types such as Companies, Governments, TV shows, and Politicians. We have used network analysis techniques and community detection algorithms to explore how pages are interconnected and how similar content tends to cluster together in the detected communities. Furthermore, we evaluate the accuracy of our detected communities against a set of ground truth labels using Normalized Mutual Information (NMI). Based on this analysis we can better understand the making of group decisions and its importance for the targeted content and effective social media strategy. We also employ Machine Learning to identify key features influencing page type classification, enhancing our understanding of attributes that significantly impact community dynamics.

2 Data

For our study we utilized the Facebook Large Page-Page Network dataset available from the Stanford Large Network Dataset Collection. The dataset was obtained via the Facebook Graph API in November 2017.

2.1 Data Characteristics

- Nodes: Each node represents a Facebook page (22,470 nodes).
- Edges: Edges between nodes signify mutual likes between pages, indicating interactions and relationships (171,002).
- Page Categories: Each node (page) is classified into one of four categories: politician, governmental organization, television show, or company.

2.2 Data Cleaning

We removed duplicates, dropped missing values, standardized data types for columns, and eliminated duplicate edges and self-loops. The Stanford dataset is so robust that, after going through the cleaning pipeline, no records were lost.

2.3 Data Sampling

Since the dataset is very large, we cannot make meaningful insights about it directly. Therefore, we applied stratified data sampling to ensure that the sample is representative across different types of nodes. We first grouped the data based on page type and set the sampling fraction to 0.1 (i.e., 10% of random nodes under each page type). After sampling the data, we obtained 2247 nodes and 1604 edges. This size is easier to manage for computing network measures, detecting communities, and visualizing.

3 Project Process

Below are the overview of the things we have achieved:

- Network Measures
- Degree Distribution
- Community Detection
- Node Classification
- Evaluation of Detected Communities based on Ground Truth
- Identify Influential Nodes based on Network Measures
- Correlation of page types across different communities
- Comparison to Real World Graph Model (Random, Small, Preferential Attachment Graph Models)

3.1 Network Measures

After sampling the data, we removed isolated nodes and focused on the largest connected component, due to the presence of many disconnected nodes. We then calculated the average degree centrality, betweenness centrality, closeness centrality, average path length, and clustering coefficient on this component. Based on these measures we can have implications on resilience and vulnerability, information spread and community structure further based on these measures we can identify key connectors that have high betweenness centrality.

Network Measures	
Measure	Value
Average Degree Centrality	0.004463382157123832
Average Betweenness Centrality	0.01030936588331712
Average Closeness Centrality	0.11848284276098103
Average Path Length	8.721715046604528
Average Clustering Coefficient	0.2483903640204774

Table 1: Summary of Network Measures

Figure 1: Network graph of the largest connected component

3.2 Degree Distribution

The shape of the degree distribution helps in classifying the type of network. We tried to use this degree distribution to check if it's going to fit in any of the real world models. First we plotted based on the degree vs fraction of nodes having degree k.

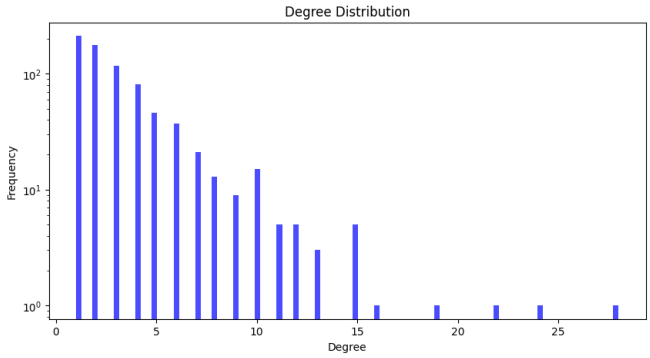

Figure 2: Degree distribution

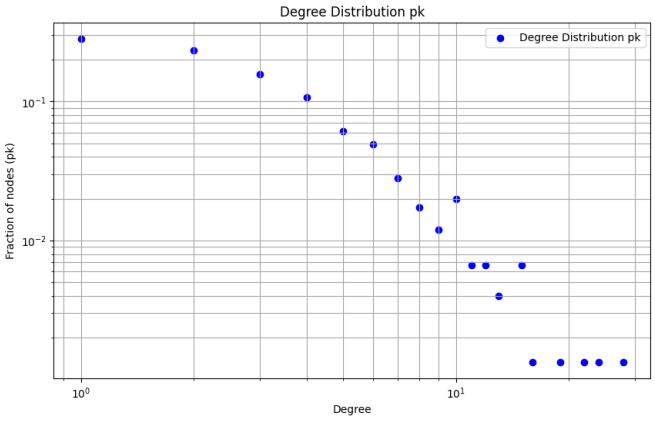

Figure 3: Degree distribution pk

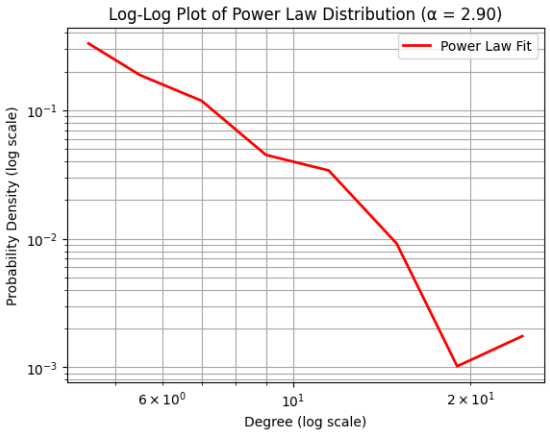

Figure 4: log-log plot of power law

From Figure 2, we observe that while a few nodes have a high degree, the majority of nodes have only a few connections. Based on this observation, in Figure 4 we plotted a log-log graph where the x-axis represents the logarithm of the degree, and the y-axis represents the logarithm of the fraction of individuals with that degree $\log(p_k)$. The plot shows an almost linear trend, typical of log-log plots for power-law distributions. This pattern indicates that the network exhibits a power-law degree distribution, classifying it as a scale-free network. Such networks fit the preferential attachment model, where the higher a node's degree, the greater the probability that new nodes will connect to it.

3.3 Influential Nodes

Based on the centrality measures discussed, we identified the top five influential nodes for each measure. Notably, page 12464, which has the highest centrality scores show how strong it is positoned within the network on multiple dimensions. It serves as a major hub or connector within the network, linking clusters or groups and controlling the flow of information between them. Additionally, it can rapidly gather and disseminate information throughout the network.

Node	Value
Node 15177	0.17337031900138697
Node 20135	0.17088174982911825
Node 12464	0.16552637386890312
Node 15236	0.16483516483516483
Node 11834	0.164257555847569

Table 2: Top 5 nodes based on degree centrality

Node	Value
Node 15177	0.2709968888076421
Node 3906	0.20728358414101053
Node 20135	0.1970479706394781
Node 11507	0.13960616591089942
Node 12464	0.1359372046474118

Table 3: Top 5 nodes based on betweenness centrality

Node	Value
Node 12464	0.007858284057544466
Node 13098	0.007826760485459114
Node 15236	0.00659822600033243
Node 3906	0.006512775024280039
Node 10503	0.005621713201788168

Table 4: Top 5 nodes based on pagerank centrality

3.4 Community Detection

In community detection we aim to identify groups of nodes (pages) within the network that are more densely connected to each other than to the rest of the network. This process is crucial for understanding the structure and dynamics of the network, as it reveals how nodes are organized into clusters or communities that can have distinct behaviors and properties. Out of all several community detection algorithms we started with Girvan Newman algorithm however it took so long to breakdown the communities and hence in our case it's affecting performance therefore we tried with Louvain Method which is efficient hierarchical clustering algorithm designed for large networks. The method works by optimizing the modularity score across the network, which quantifies the density of links inside communities compared to links between communities. We have used this approach particularly because it is favored for its speed and its ability to handle large-scale networks effectively. In Figure 5, we illustrate all the nodes belonging to the same community using a single color for each group. Algorithm detected in total of 25 communities and can be seen in Figure 7.

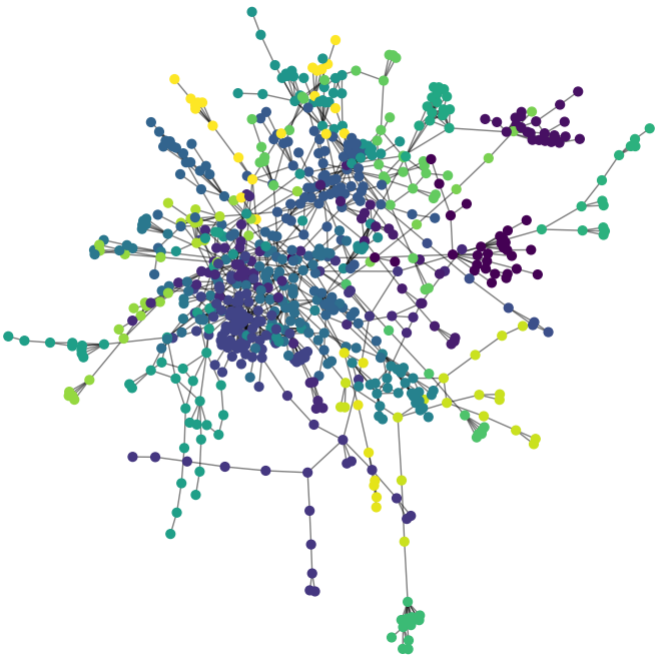

Figure 5: Detected communities

Community	Majority Category	Homogeneity	Total Nodes	Category Breakdown
0	Company	0.68	31	Company: 21, Government: 7, Tvshow: 3
1	Government	0.54	24	Government: 13, Politician: 11
2	Government	0.63	19	Government: 12, Company: 6, Politician: 1
3	Government	0.96	45	Government: 43, Politician: 2
4	Company	0.89	27	Company: 24, Tvshow: 3
5	Government	0.92	86	Government: 79, Company: 5, Tvshow: 2
6	Politician	1.0	6	Politician: 6
7	Government	0.49	70	Government: 34, Politician: 32, Company: 3, Tvshow: 1
8	Politician	0.54	50	Politician: 27, Government: 15, Company: 5, Tvshow: 3
9	Government	0.91	68	Government: 62, Company: 4, Politician: 2
10	Tvshow	0.74	19	Tvshow: 14, Company: 5
11	Politician	0.83	29	Politician: 24, Tvshow: 3, Government: 2
12	Government	0.56	16	Government: 9, Politician: 7
13	Politician	0.61	44	Politician: 27, Government: 15, Tvshow: 2
14	Government	0.47	40	Government: 19, Politician: 17, Tvshow: 2, Company: 2
15	Politician	1.0	17	Politician: 17
16	Company	0.64	14	Company: 9, Tvshow: 5
17	Politician	1.0	16	Politician: 16
18	Government	1.0	9	Government: 9
19	Politician	0.69	35	Politician: 24, Government: 10, Company: 1
20	Company	0.67	6	Company: 4, Politician: 2
21	Government	0.74	23	Government: 17, Politician: 6
22	Government	0.89	9	Government: 8, Company: 1
23	Government	1.0	19	Government: 19
24	Politician	1.0	8	Politician: 8
25	Politician	0.57	21	Politician: 12, Government: 9

Figure 6: Communities with majority category and homogenity score

3.4.1 Evaluation of Detected Communities

After finding the detected communities we aggregated the true and predicted labels across all nodes in the graph, rather than evaluating them separately for each community. In the code we collected all true and predicted binary classifications across all nodes in all communities. The true-binary array for each community is created by comparing each node's category to the majority category of the community it belongs to.

From the Table 5, We can see that precision 74% of the nodes assigned to their respective communities are correctly categorized based on the majority category of each community. This indicates that when a community claims a node belongs to its predominant category, it is correct about 74% of the time. A recall of 1.00 indicates that every node that should have been identified as belonging to the majority category within their communities was correctly identified. NMI of 0.34 indicates that there is some, but not a strong, correlation between the communities detected by the algorithm and the actual page types. High Precision and low NMI means that some true clusters might not be effectively captured by the model, leading to a lower overall mutual information score.

Metric	Value
Precision	0.74
Recall	1.0
Purity	0.74
NMI	0.34

Table 5: Assessing quality of community detection

3.5 Correlation of Pages Across Communities

Correlation of page types across communities measures content-based relationships within a social network. Here we compare how same page types relate to each other across different communities within the entire network. Here we examine if certain types of pages consistently appear together in various communities throughout the network.
Diagonal (1.00): Represents the correlation of each page type with itself, which is always perfect (1.00).
Off-diagonal values: Show how often pages of one type tend to co-occur in the same community with pages of another type.
Positive values (e.g., 0.35 between TV Show and Company): Indicate that these page types are more likely to appear together in the same community.
Negative values (e.g., -0.40 between Company and Government): Suggest that these page types tend not to co-occur frequently in the same community, indicating a segregation of community types based on page category.

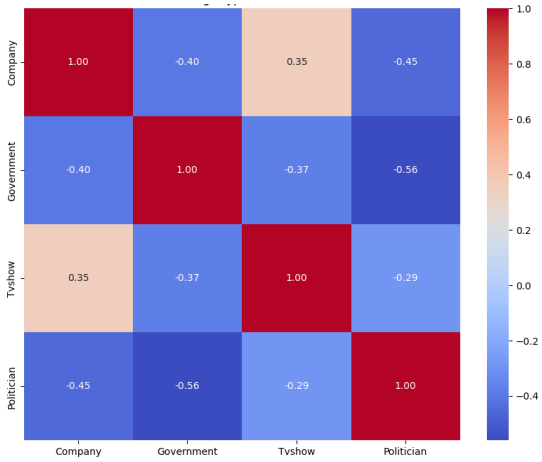

Figure 7: Correlation Matrix for different page types tending to cluster together

3.6 Node Classification

In this project, we apply machine learning to attempt understanding and predicting the types of various Facebook pages, such as companies, governments, TV shows, and politicians, based on their network characteristics. We capture the influence and connectivity of each page using various measures of centrality, including degree, closeness, betweenness, eigenvector, and PageRank.

For example, Let's consider from the Table:6 we see that government pages have a high degree and betweenness centrality and act as a hub connecting

various people and organizations. If we can train a RandomForestClassifier on these network-derived features, we will thus be able to predict whether a new, unlabeled node in this network is likely to be a government page, TV show, company, or politician, based on its position and connections within the network.

Overall Accuracy: 71%			
Page Type	Precision	Recall	F1-Score
Company	0.39	0.32	0.35
Government	0.81	0.76	0.78
Politician	0.65	0.80	0.72
Tvshow	0.50	0.30	0.37

Table 6: Performance metrics for a classifier

From the Table 6, we observe that the overall accuracy is 71%, indicating that the model correctly predicted the category of nodes (such as Company, Government, Politician, TV Show) 71% of the time across all predictions made. The model is particularly effective at identifying 'Government' pages (exhibiting high precision and recall), but it struggles with 'TV Show' pages, which exhibit lower precision and recall.

4 Project Results

After sampling the data, a few nodes became isolated, and others formed edges but remained part of disconnected components far from the main clusters. We removed such nodes and focused on the largely connected components. We identified network measures and influential pages, finding that page 1264 is highly influential. Using the Louvain Method, we detected a total of 25 communities; notably, one community predominantly consisted of 78 government pages. The purity (0.74) indicates that the communities are relatively clean but still contain a notable portion of nodes from other categories, which could be a point of improvement. Although the detection algorithm performed well overall, as indicated in Table 5, the Normalized Mutual Information (NMI) was not high. This lower NMI suggests that some true clusters might not have been effectively captured by the model. Additionally, as shown in Figure 7, the correlation of page types across communities indicates that TV Show and Company page types are more likely to cluster together. Finally, with an accuracy of 71%, we can predict the category of new unlabeled nodes in the network

5 Discussion - Takeaways

After spending time on the project, we were able to achieve most of the initial goals, as evident from the results. However, one aspect that did not go as expected was the low NMI for the detected communities. In future efforts, we plan to experiment with multiple community detection algorithms to improve NMI. Additionally, although the purity was balanced, we observed a minor presence of nodes from other categories, which will be looked after further. In addition to that we learned how crucial centrality measures are for identifying influential nodes and their potential use in social media marketing.

6 Future Work

In the future, we would love to compare multiple community detection algorithms on this Facebook page dataset and observe which maintains a good balance of NMI and other metrics. Additionally, we are very keen to apply Community Robustness Testing, as we learned in class that the entire network can be disrupted if influential nodes are removed. We want to see how community structure remains consistent after the addition or removal of nodes.

References

[1] Linhares, C.D.G., Ponciano, J.R., Pereira, F.S.F. et al. Visual analysis for evaluation of community detection algorithms. Multimed Tools Appl 79, 17645–17667 (2020). https://doi.org/10.1007/s11042-020-08700-4

[2] Prof. Cynthia Hood Teaching Material

[3] Social Media Mining An Introduction. https://doi.org/10.1017/CBO9781139088510

[4] Maximilian Jerdee, Alec Kirkley and M. E. J. Newman. Normalized mutual information is a biased measure for classification and community detection

[5] Louvain Communities, NetworkX Documentation