

CS 176 Final Project

Group Name: Youtube Analysts

Group Members: Shreya Raj and Erica Xue

Introduction

- Using **NetworkX** to analyze and visualize large-scale communities in highly extensive datasets
 - Uncovering patterns and connections within a social network
 - Implementing sampling, community detection, and visualization
 - Goal: to understand the network's cohesiveness and structure
- Tools:
 - NetworkX, Matplotlib, Google Colab, Pandas
- Technologies:
 - Python, Network Analysis Algorithms, Graph Theory

Problem Description

- Large-scale network analysis overview
 - Degree distributions, clustering coefficient, average degree, node connectedness
 - Centrality algorithms (Degree, Betweenness, and Closeness)
 - Community detection algorithms (Louvain and Girvan-Newman)
 - Visualizations of communities and the overall network
- Challenges
 - Too large of a dataset
 - Random sampling to find optimal subset of the data
 - Detecting meaningful communities through modularity
 - Re-sampling
 - Clear, interpretable visualizations
 - Experimenting with different layouts and hyperparameters

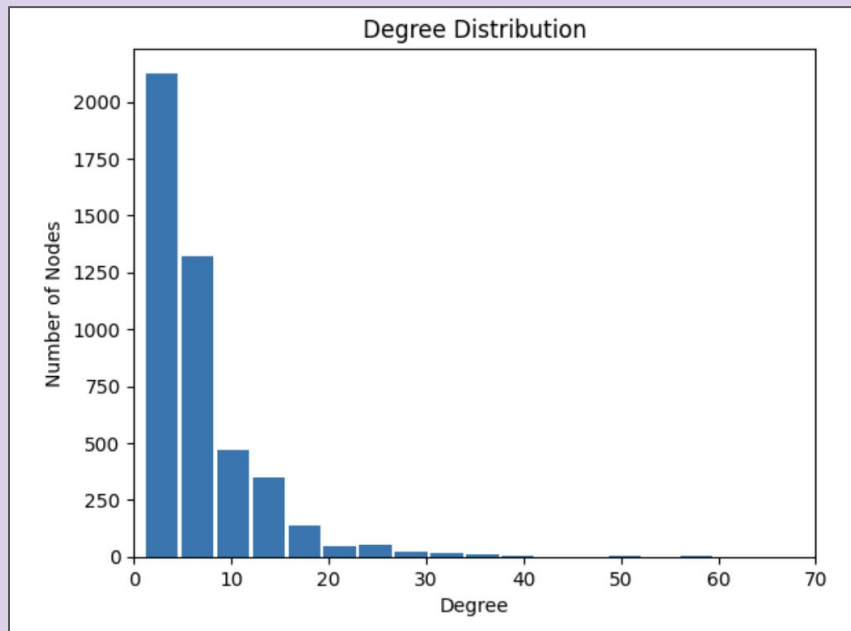
Dataset Description

- YouTube Social Network Dataset
 - Type: Undirected Graph
 - Nodes: 1,134,890 (users)
 - Edges: 2,987,624 (friendship connections)
 - Diameter: 20 (the shortest path connecting the two furthest apart users is a distance of 20 friendship connections)
 - Communities: 8,385 communities
- Sampling nodes with highest degrees
 - Creating subgraph with 15,000 edges for analysis

Analysis and Results

Number of Nodes: 4568
Number of Edges: 15000
Average Degree: 6.567425569176883
Clustering Coefficient: 0.015436066577593293
Is Graph Connected: False

- Sampled graph results:
- Number of bridges: 577
- Number of articulation points: 469
- Maximum diameter of connected component: 14
- Degree distribution histogram:
 - Power-law degree distribution



Analysis and Results

Degree centrality

137767	0.000438
5754	0.001971
28233	0.000438
4964	0.001314
30816	0.002409

dtype: float64

Connectedness centrality

137767	0.178253
5754	0.228315
28233	0.222915
4964	0.216543
30816	0.218049

dtype: float64

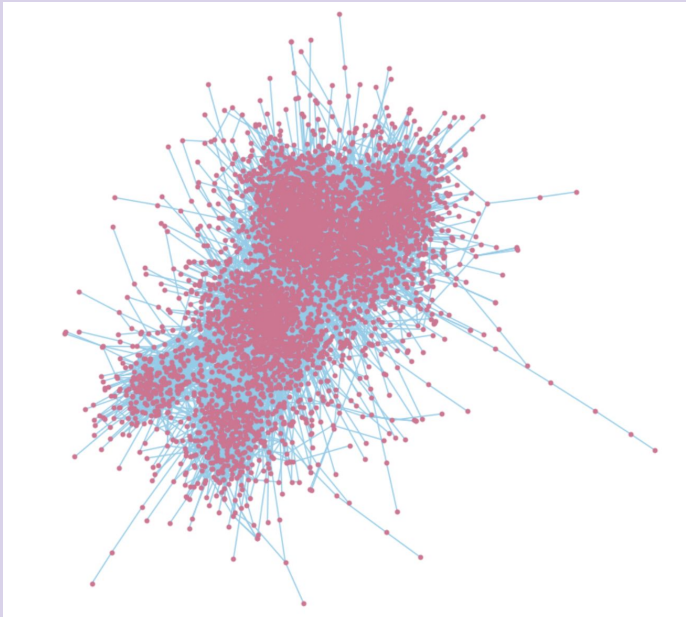
Betweenness centrality

137767	0.000218
5754	0.000225
28233	0.000008
4964	0.000491
30816	0.001756

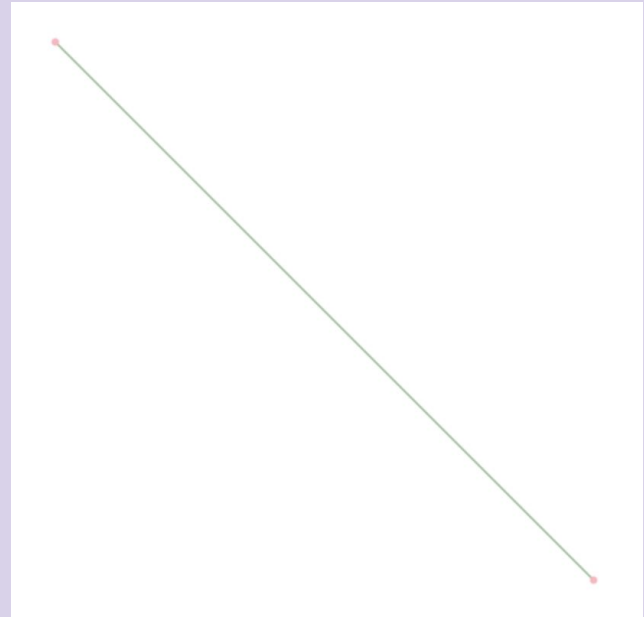
dtype: float64

Analysis and Results

- Largest Connected Component



- Second Largest Connected Component

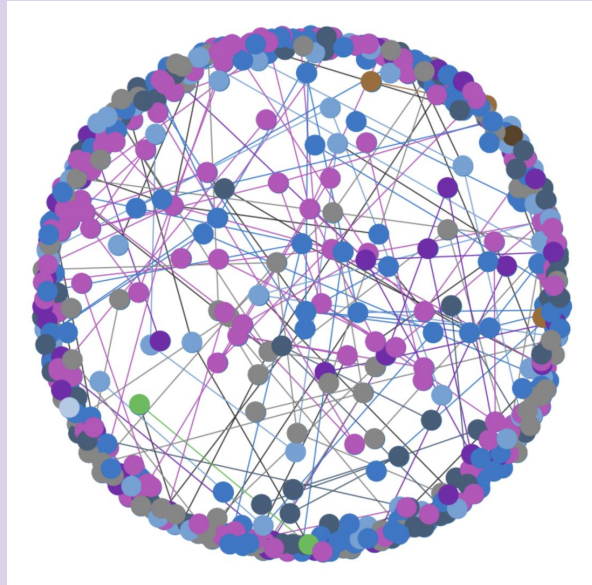


Analysis and Results- Louvain algorithm

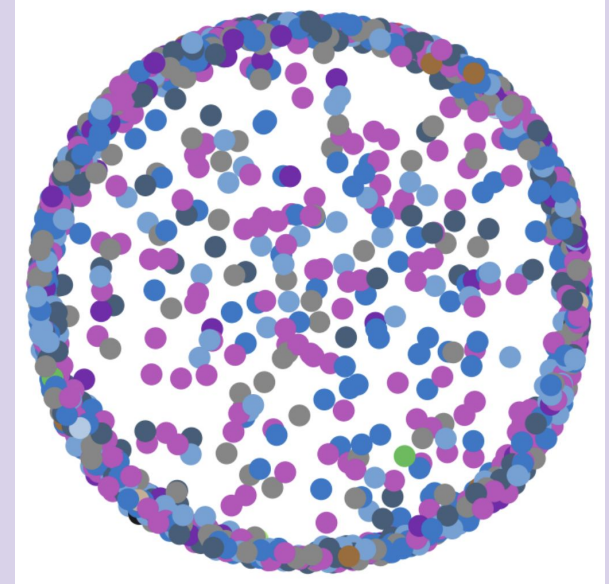
Number of communities: 24

Modularity score: 0.603

500 nodes with edges

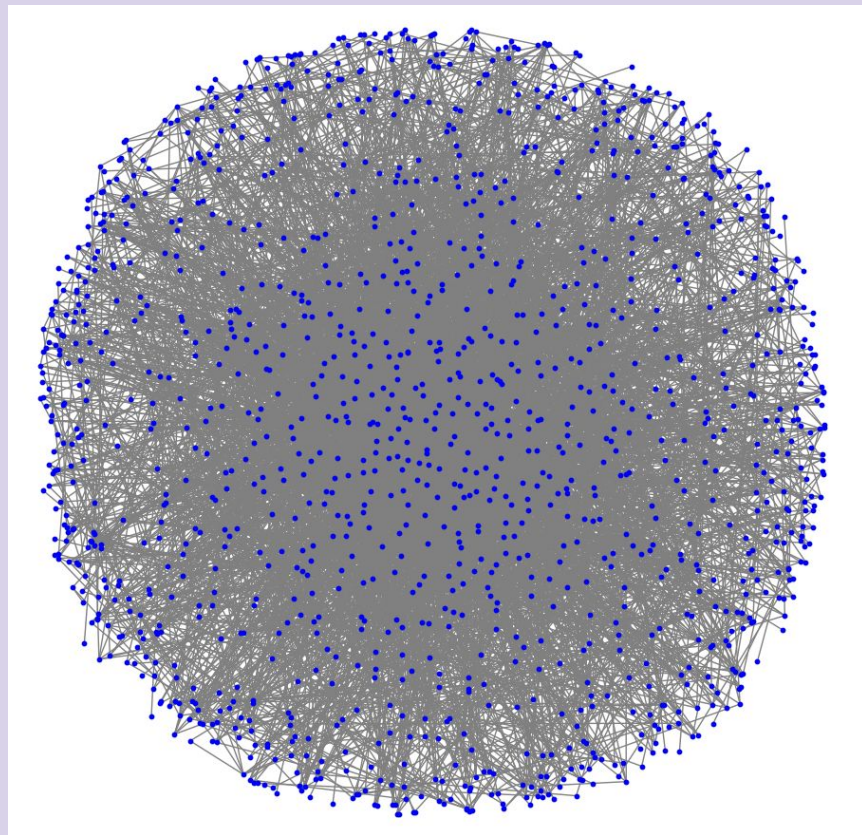


1250 nodes without edges



Analysis and Results- Louvain algorithm

- Largest community by node count
 - 1326 nodes



Analysis and Results- Girvan-Newman

Top 20 Communities Based
on Girvan-Newman

First Round

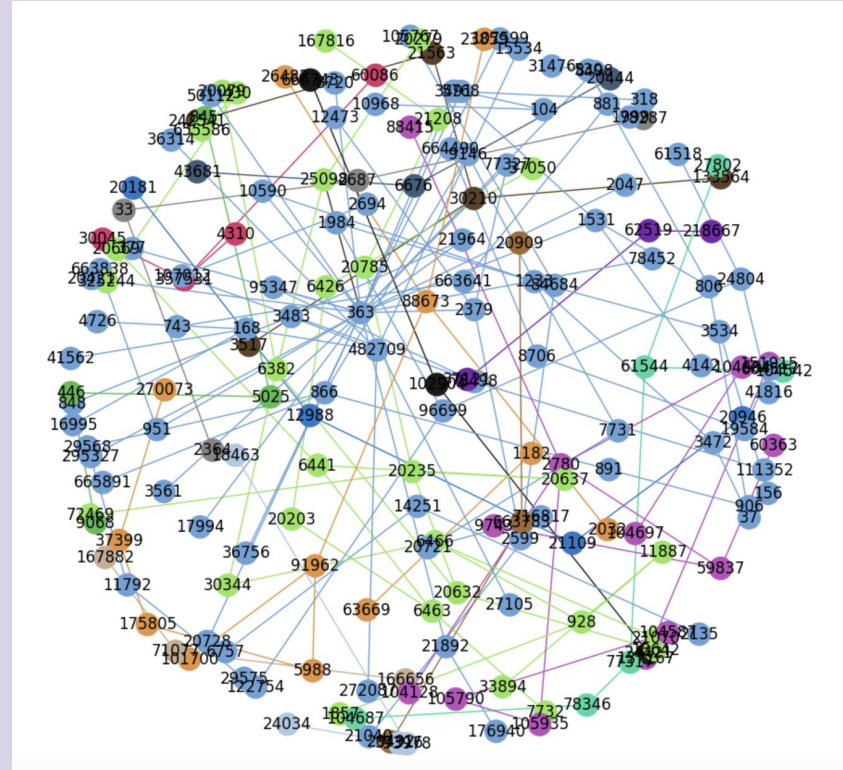
Number of communities: 318

Modularity score: 0.874

Second Round

Number of communities: 319

Modularity score: 0.896



Conclusion

- Sampling of the entire dataset allowed for more effective analysis and visualization
- Low clustering coefficient
 - Nodes are more randomly connected, nodes of a node do not have edges between them
- The degree distribution histogram supports the power-law degree distribution
- Louvain Algorithm identified a strong community structure with a modularity score of 0.603
 - Girvan Newman Algorithm produced an average modularity score of 0.885
 - Bridges and articulation points are very important components that ensure the network is connected
- Overall, Youtube Social Network has a decently strong community structure

Future Work

- Integrate Gephi for advanced graph visualization
- Implement a Machine learning algorithm to handle a larger quantity of nodes
- Explore maximum influence
- Utilize the results for a recommendation system

Thank You!