ENRON NETWORK ANALYSIS AND COMMUNITY DETECTION

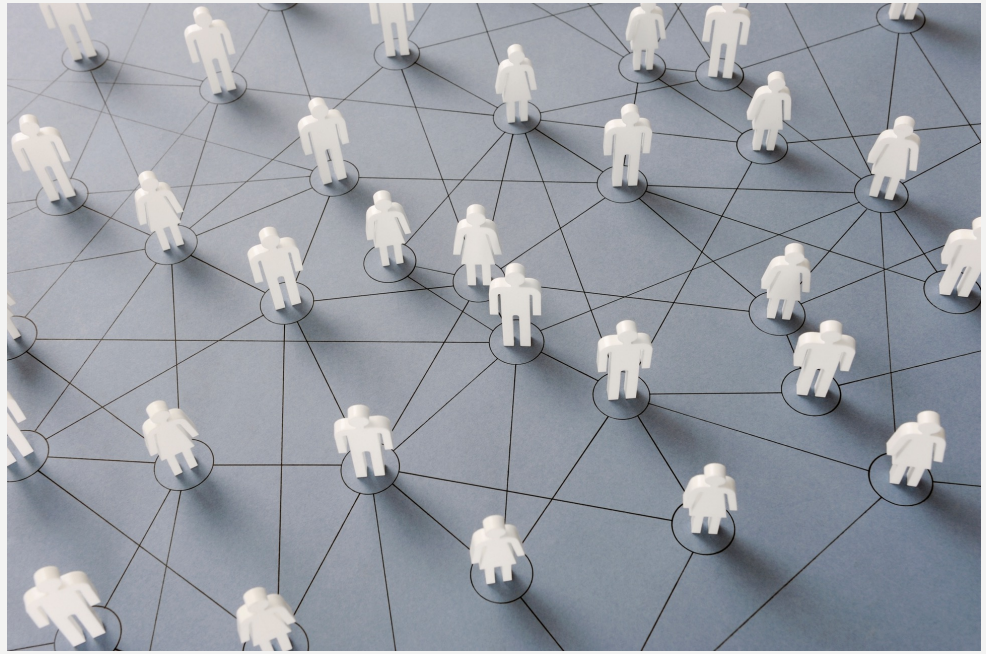
BY: KAUSTUBH LOHANI

AGENDA

1. Introduction
2. Problem Statement
3. Dataset Statistics
4. Data Cleaning and Preprocessing – The need
5. Data Cleaning and Preprocessing
6. Graph Exploratory Analysis
7. Community Detection
    i. Louvain Algorithm
    ii. Label Propogation Algorithm
    iii. Lieden Algorithm
    iv. Infomap Algorithm
    v. Walktrap Algorithm
    vi. Markov Clustering Algorithm
    vii. Girvan-Newman Algorithm (Level 7 to Level 20)

INTRODUCTION

- The Enron email communication network offers a case study for understanding complex organizational structures through the lens of social network analysis.
- This project focuses on the detection of communities within the Enron email dataset, which encapsulates interactions among over a thirty-six thousand unique emails
- Originally disclosed by the Federal Energy Regulatory Commission during its investigation.



PROBLEM STATEMENT

"Identify communities in the email-Enron dataset."

DATASET STATISTICS

Dataset statistics	
Nodes	36692
Edges	183831
Nodes in largest WCC	33696 (0.918)
Edges in largest WCC	180811 (0.984)
Nodes in largest SCC	33696 (0.918)
Edges in largest SCC	180811 (0.984)
Average clustering coefficient	0.4970
Number of triangles	727044
Fraction of closed triangles	0.03015
Diameter (longest shortest path)	11
90-percentile effective diameter	4.8

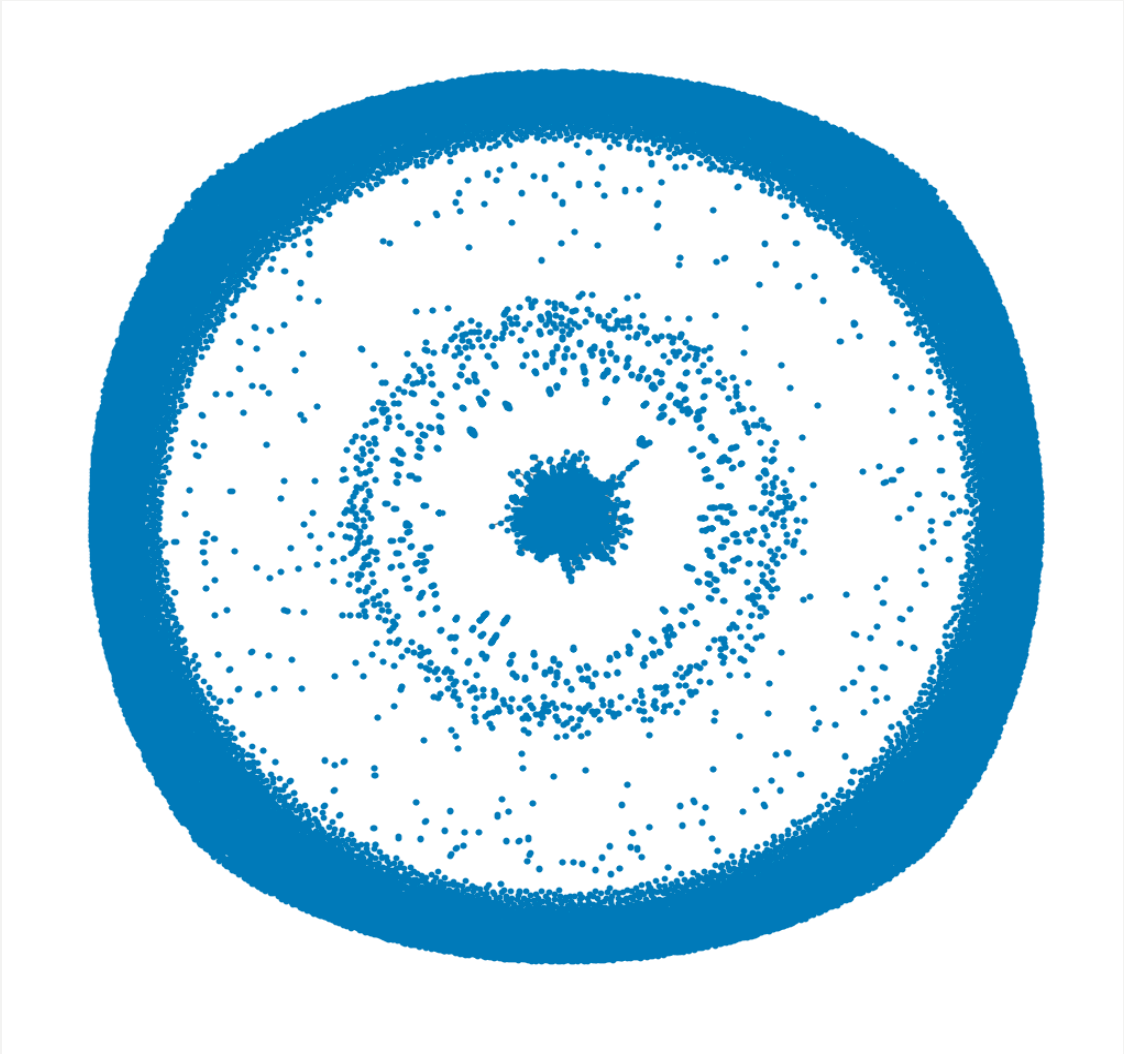
Average degree of the original network: 10.02

Additional Important Data Description

Note that non-Enron email addresses act as sinks and sources in the network as we only observe their communication with the Enron email addresses.

DATA CLEANING AND PREPROCESSING – THE NEED

Visualization without cleaning

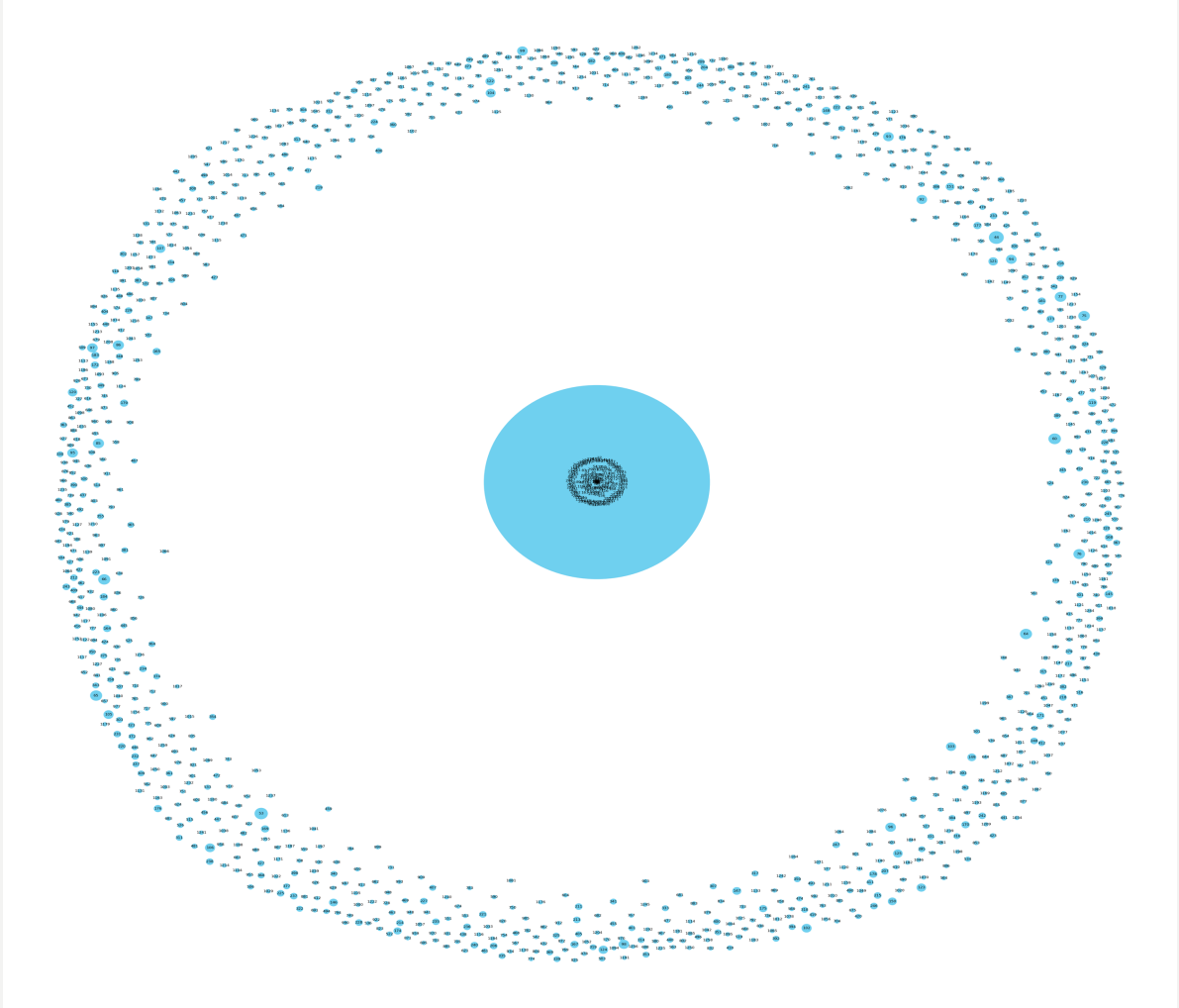


- This visualization is printed by random edge sparsification i.e. randomly removing 95% of the edges.
- Here we can see the sources and sinks (non-enron emails) as described in the data description.
- These emails as highlighted in data description are of no use for gaining insights about the communities of Enron.

DATA CLEANING AND PREPROCESSING - THE NEED - II

- Here, the blue circles are the nodes.
- The size of the blue circles represents the size of the community.
- Here we can see that the center community is the biggest and all the rest of them are represented with small circles.
- It can be inferred that the center community is the actual representation of the data and others are just outside Enron emails acting as source and sink.
- Therefore, they need to be removed for detecting communities in the meaningful central network.

Community Detection w/o cleaning
(Louvain Algorithm)
Modularity: 0.61



DATA CLEANING AND PREPROCESSING - II

Removing Non-Enron Nodes using Statistical analysis

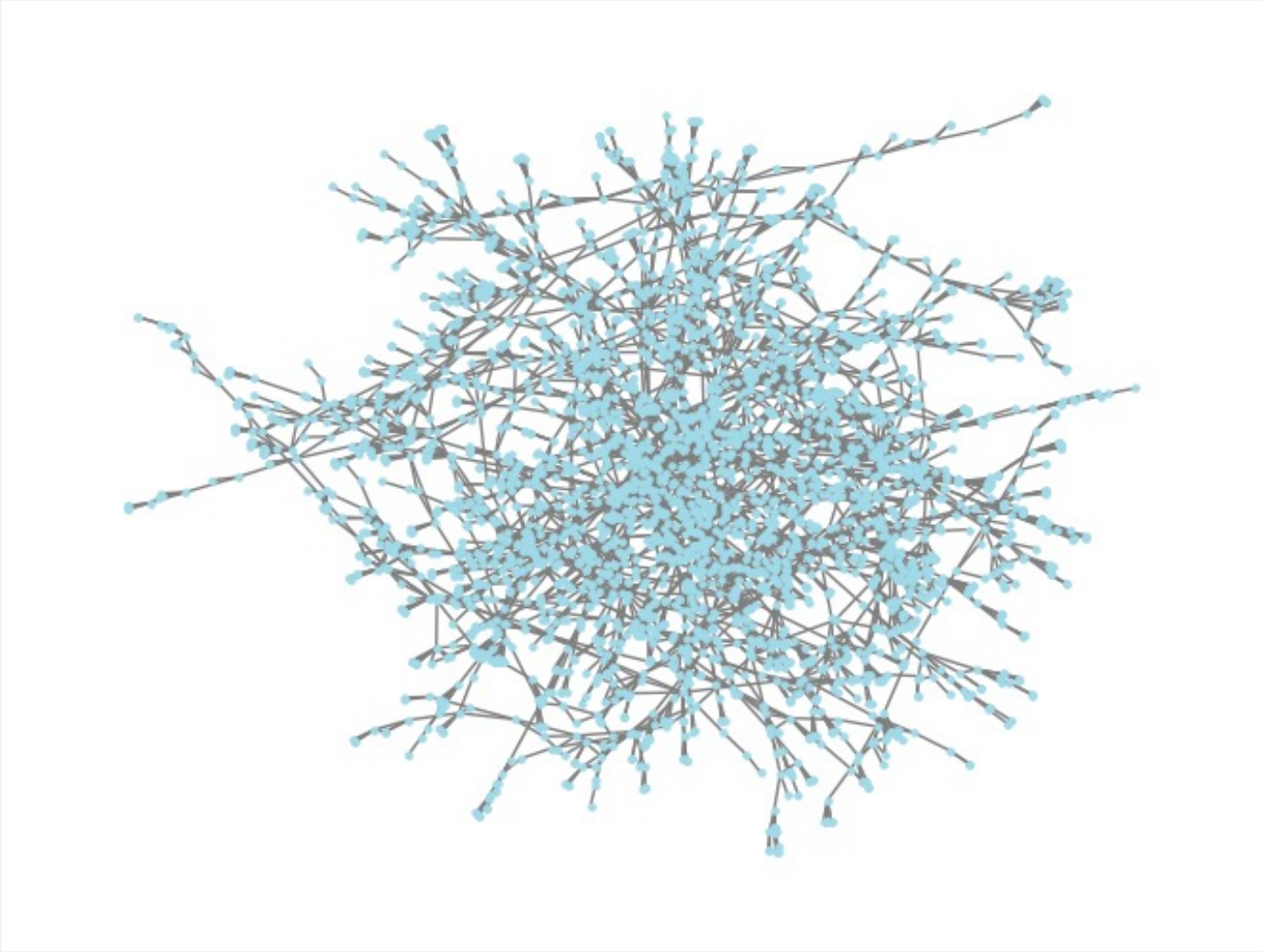
Component Analysis
Filtering small communities

Taking out the largest subsection for community detection

Removing Isolated Nodes

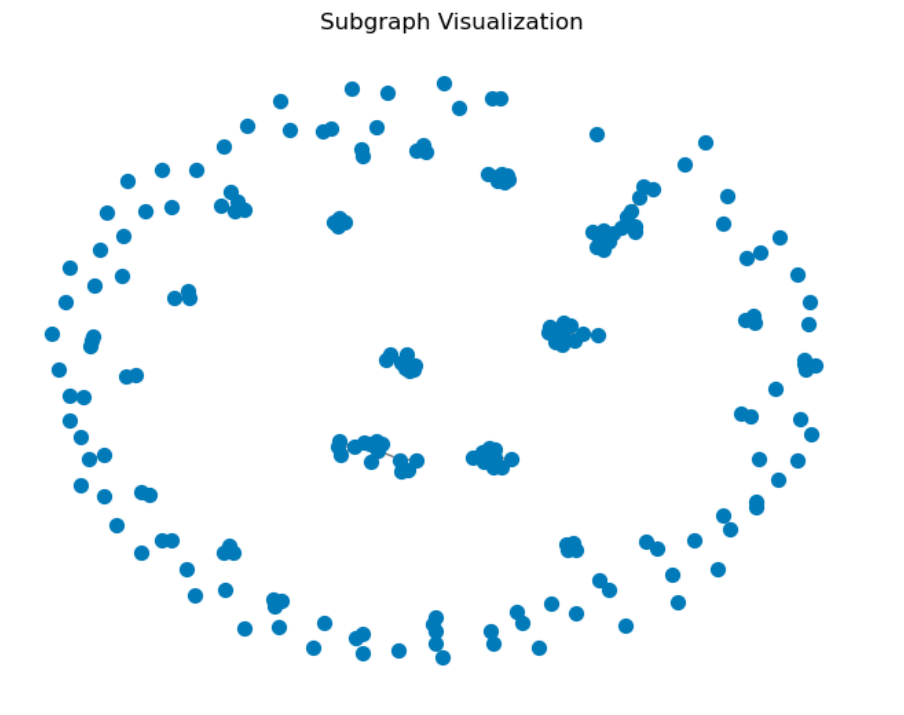
Final graph Statistics:
Number of New nodes: 3459
Number of New edges: 6221
Average degree: 3.6

DATA CLEANING AND PREPROCESSING – III

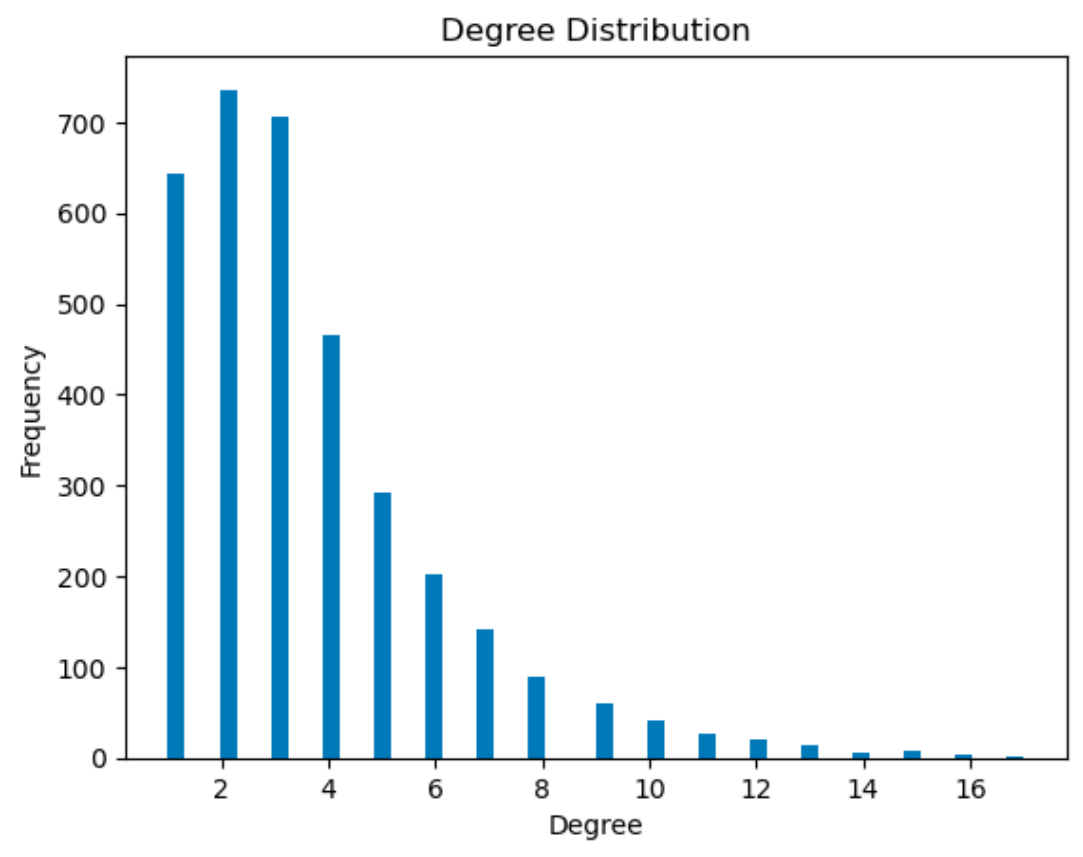


Cleaned Graph Visualization

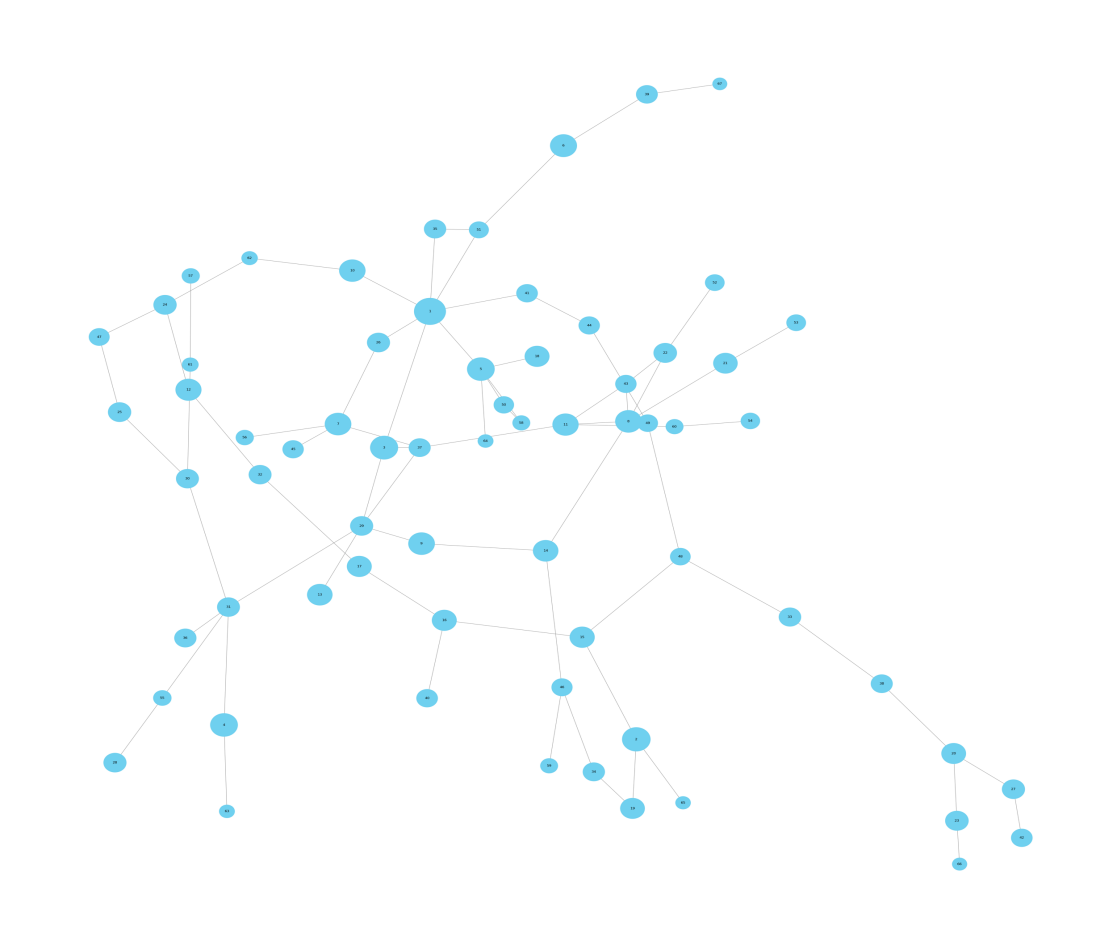
GRAPH EXPLORATORY ANALYSIS - I



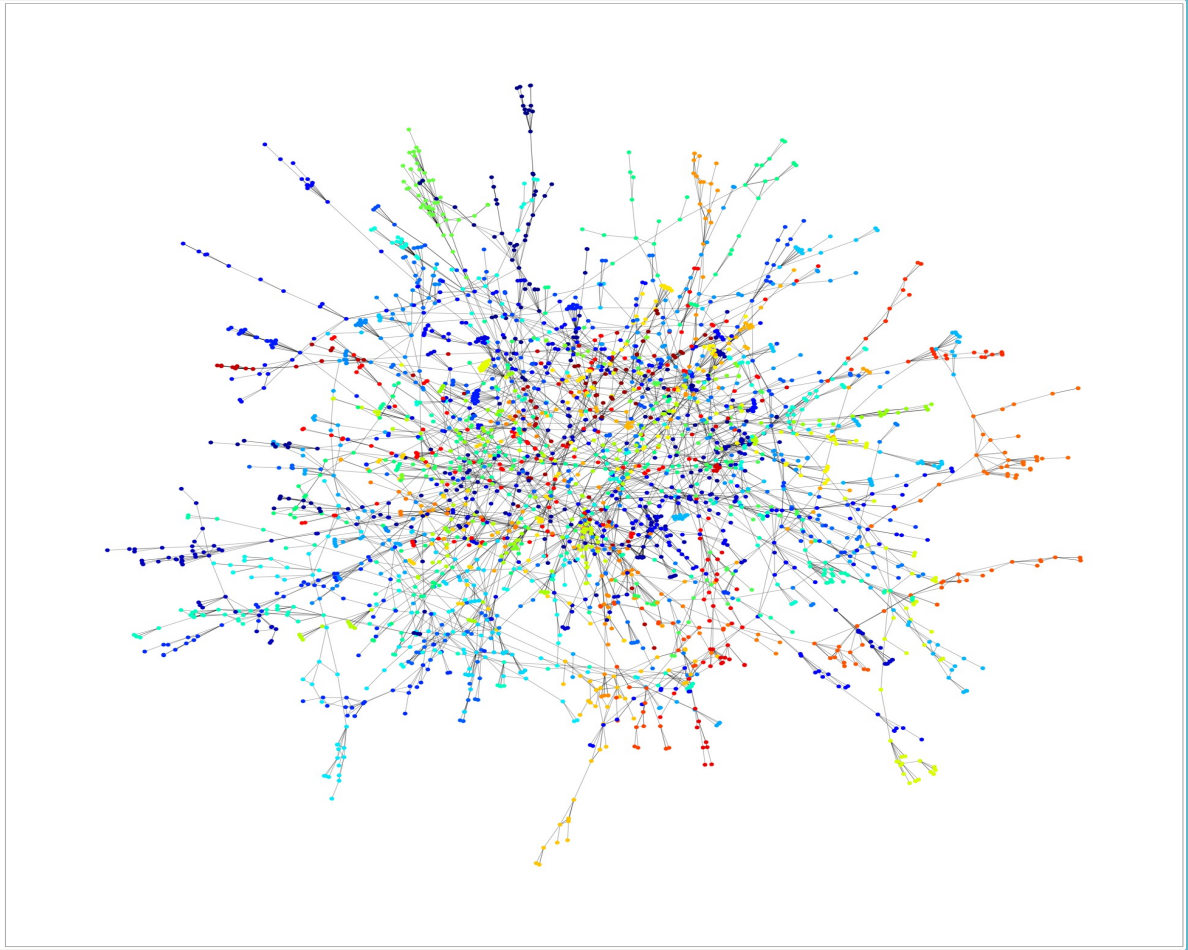
Average Degree	3.6
Density of the Graph	0.001
Average Clustering Coefficient	0.52



COMMUNITY DETECTION - LOUVAIN



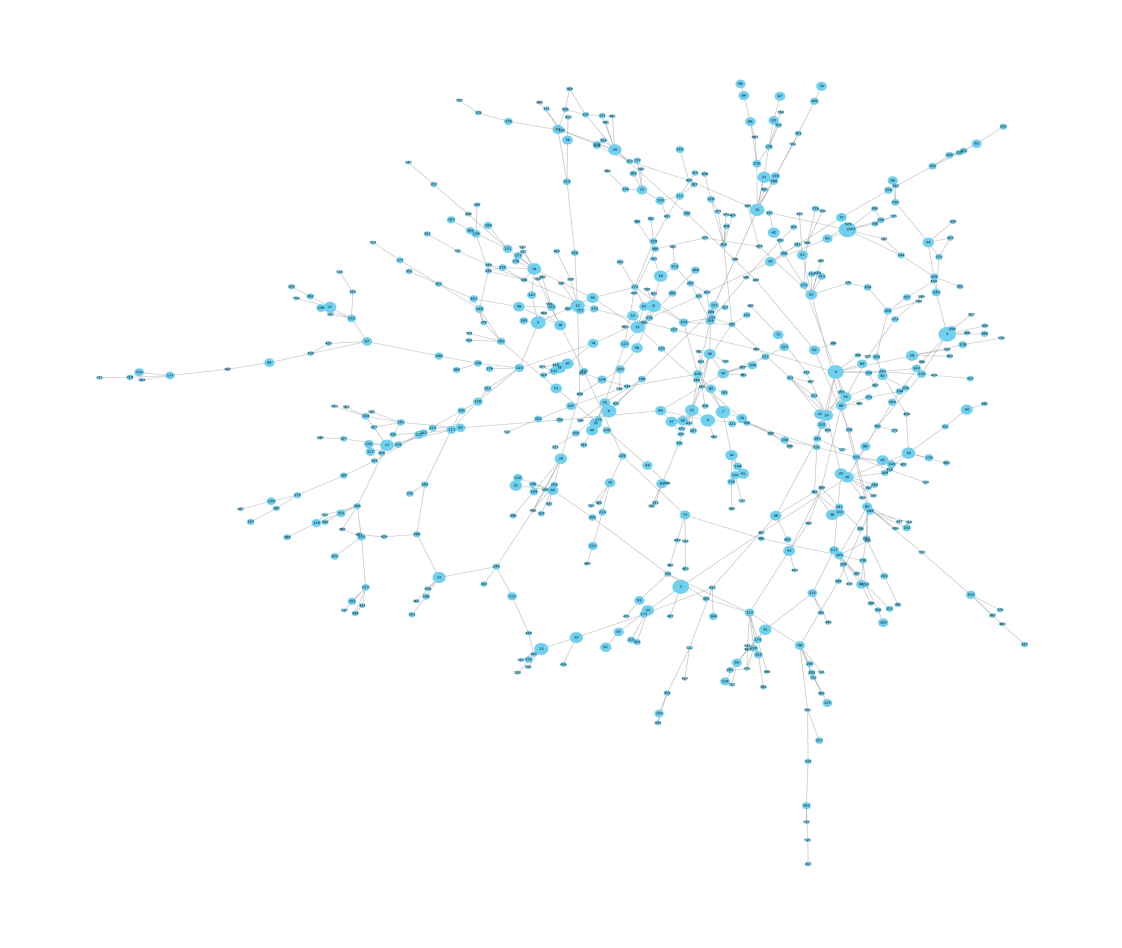
Condensed Graph



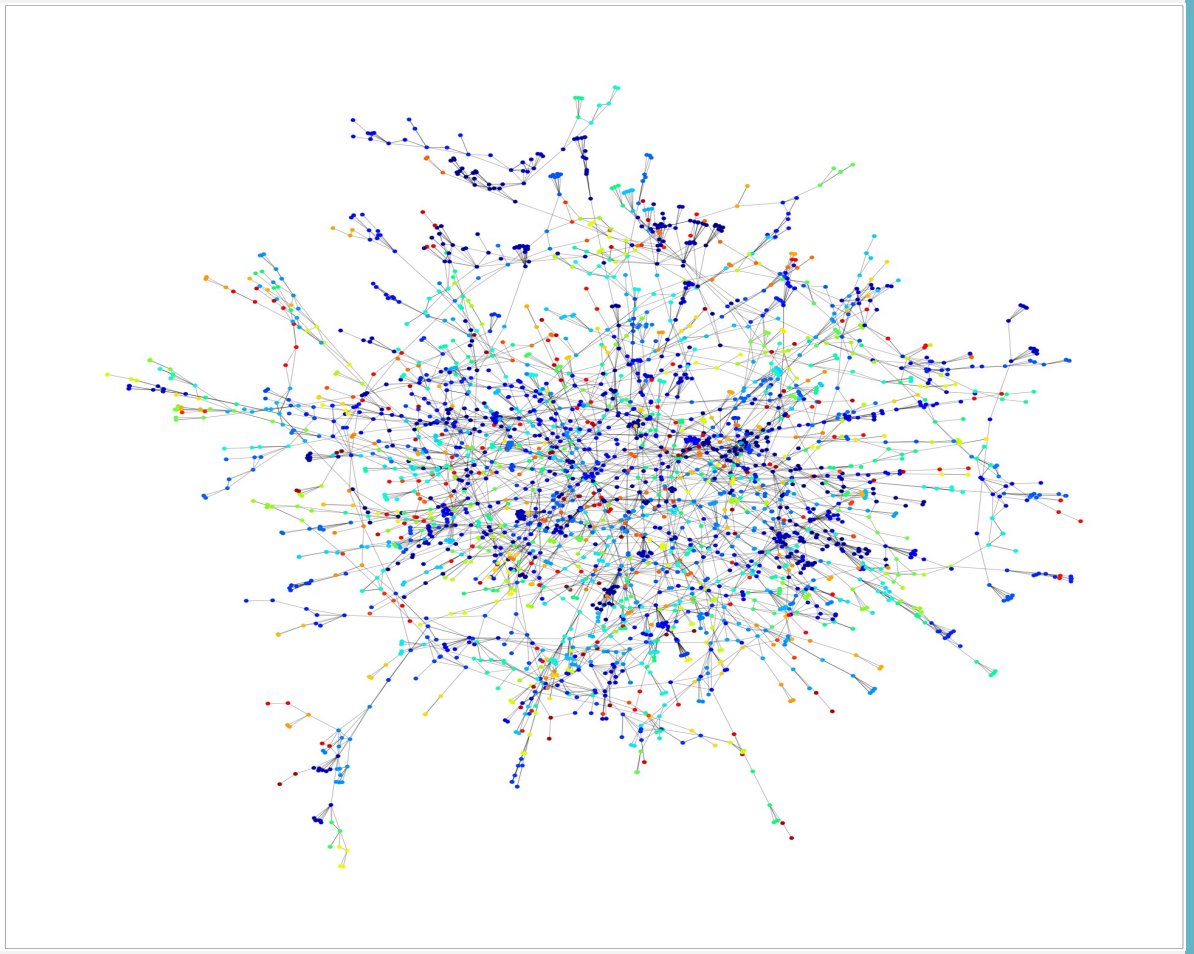
Full Data Graph

Modularity: 0.97
Conductance: 0.015
Number of Communities: 67

COMMUNITY DETECTION - LPA



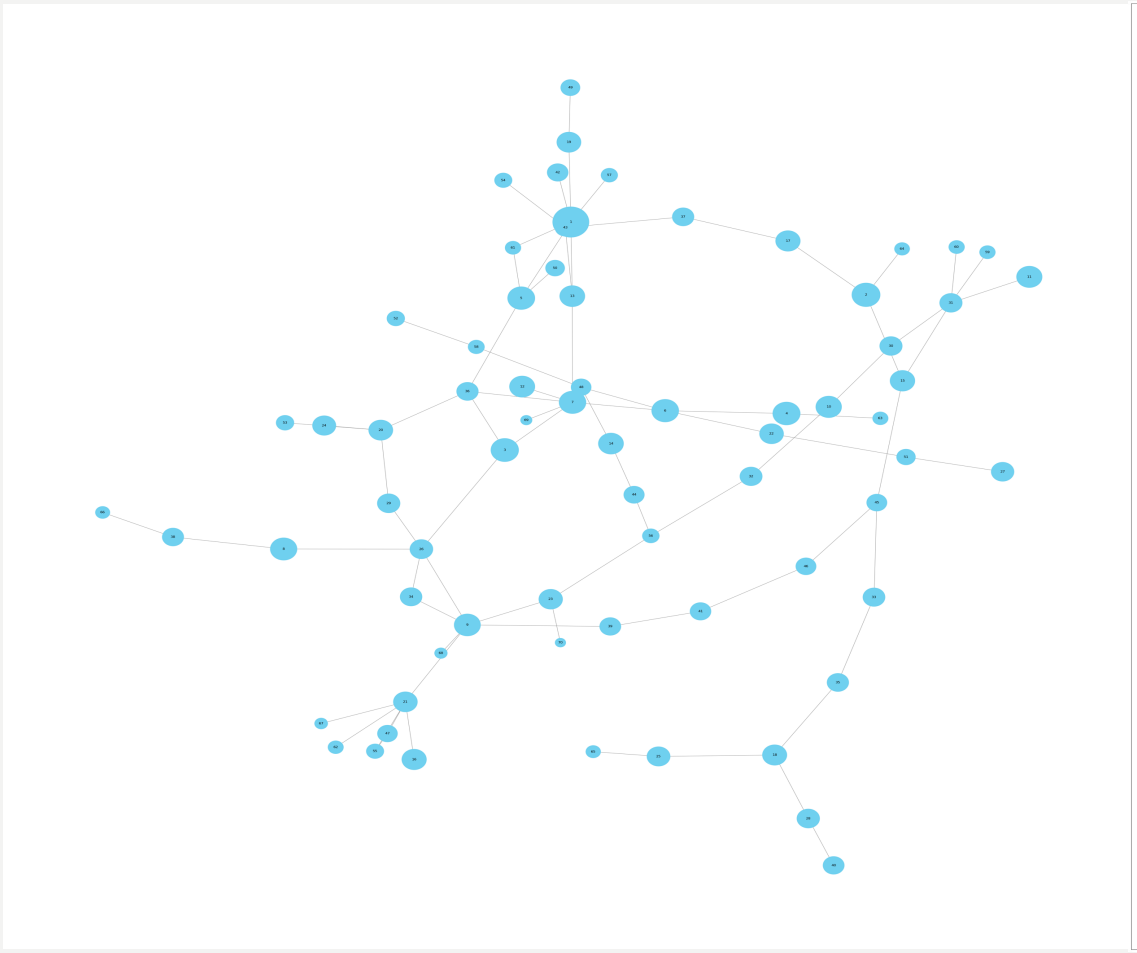
Condensed Graph



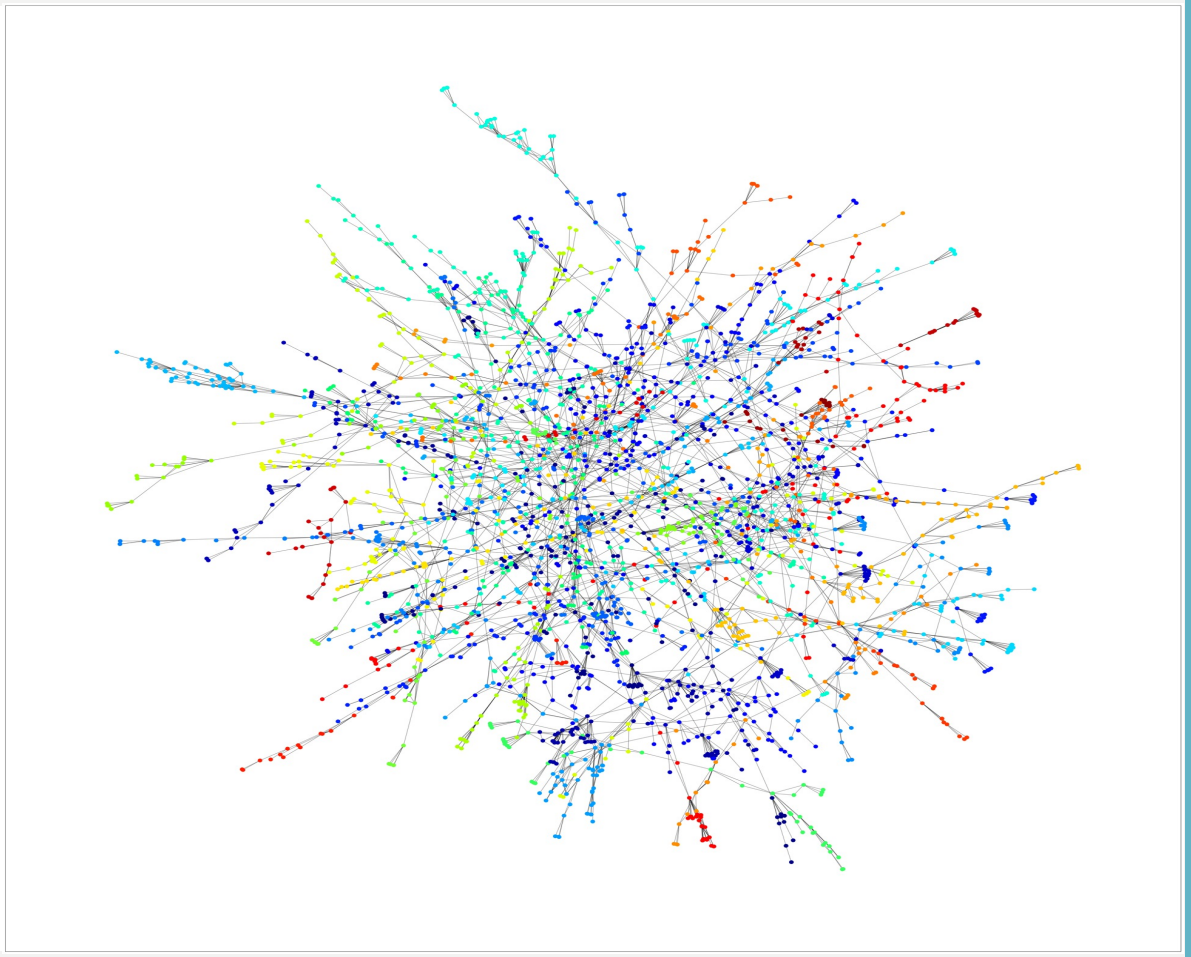
Full Data Graph

Modularity: 0.82
Conductance: 0.25
Number of Communities: 563

COMMUNITY DETECTION - LIEDEN



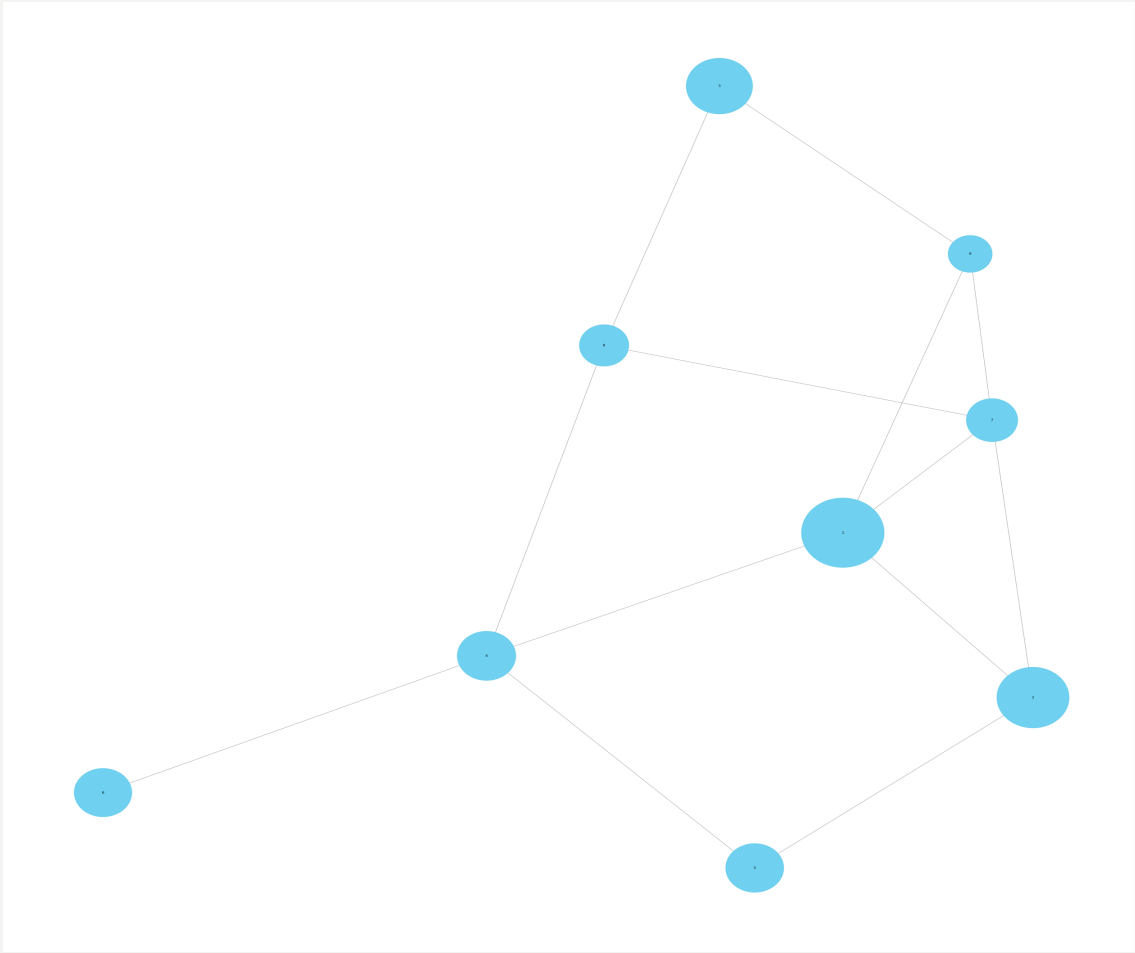
Condensed Graph



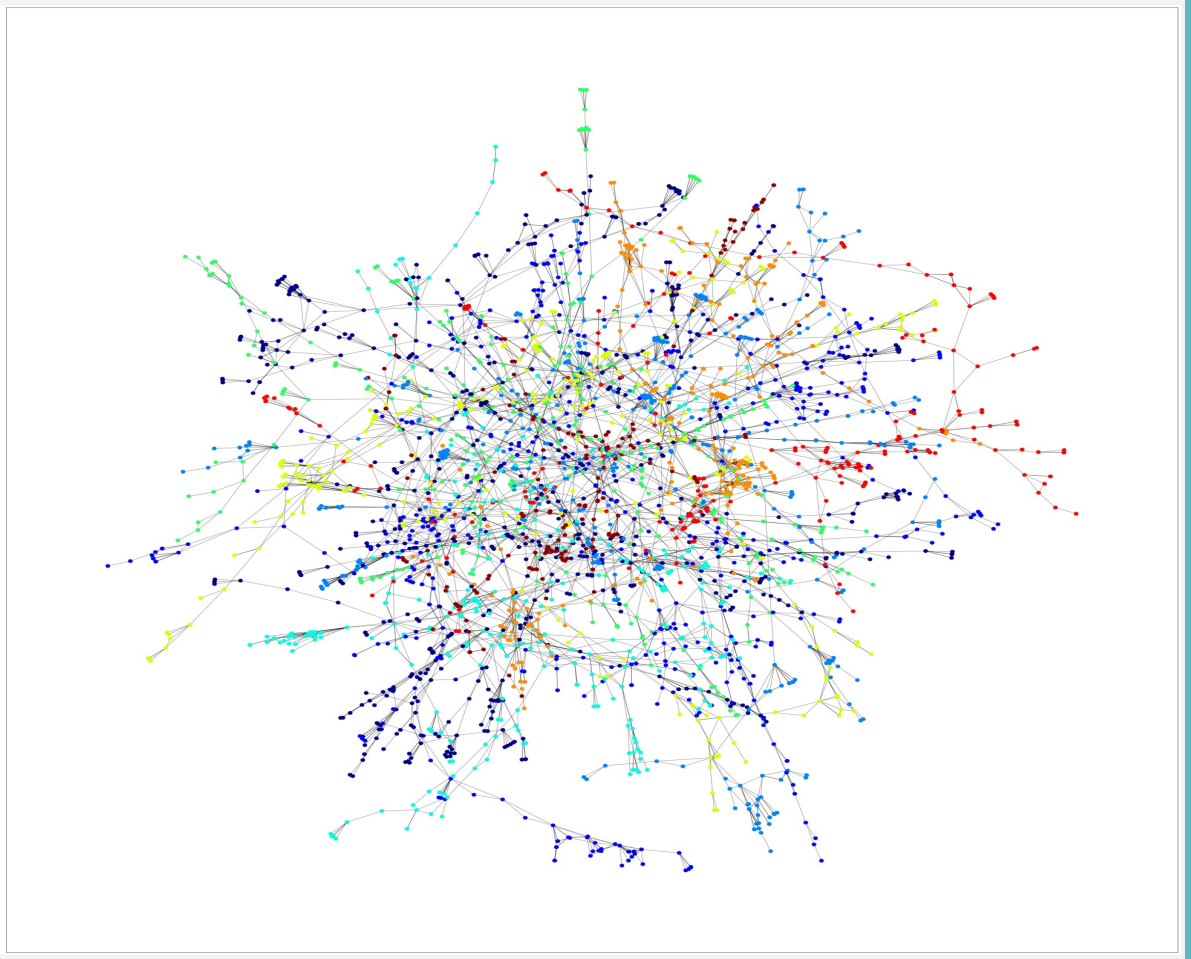
Full Data Graph

Modularity: 0.96
Conductance: 0.151
Number of Communities: 70

COMMUNITY DETECTION - INFOMAP



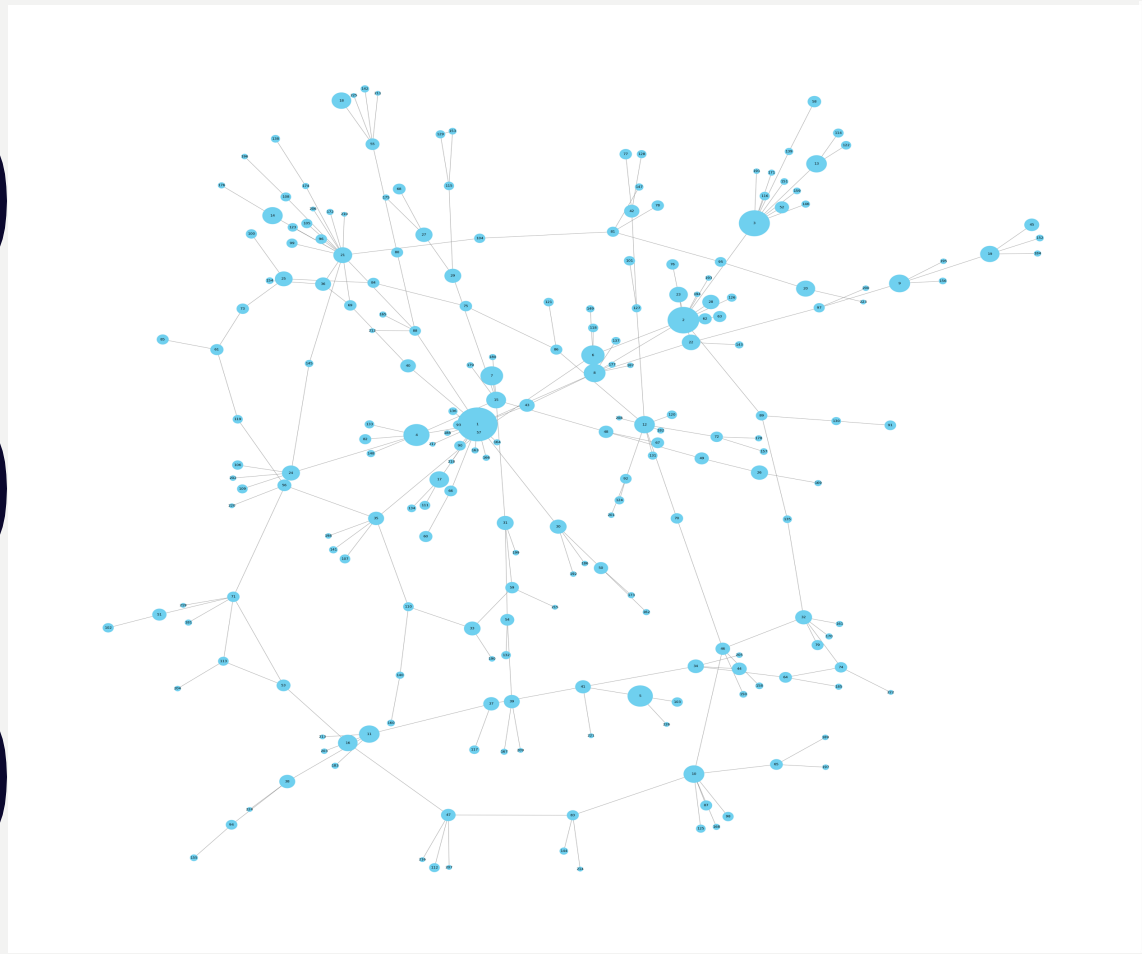
Condensed Graph



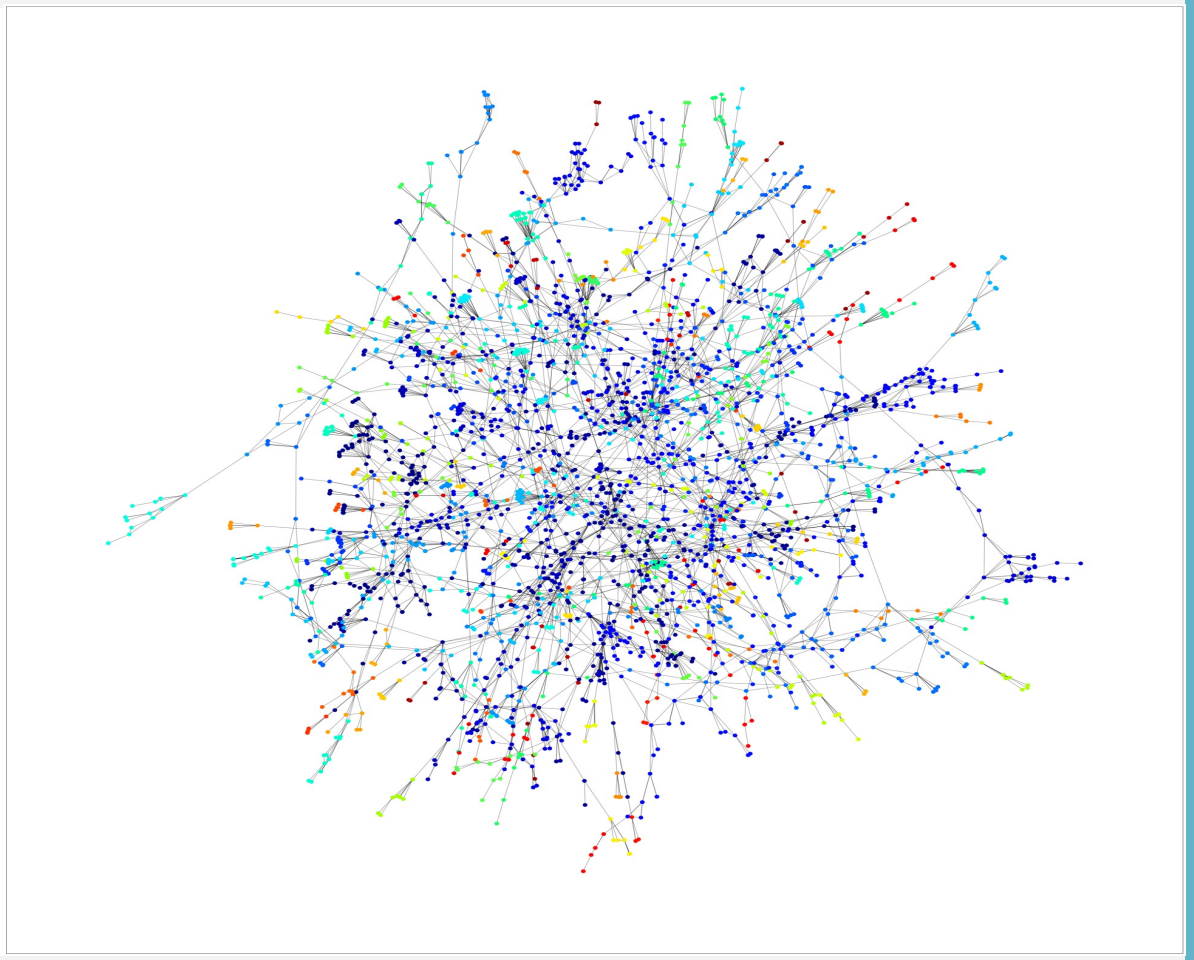
Full Data Graph

Modularity: 0.87
Conductance: 0.002
Number of Communities: 9

COMMUNITY DETECTION - WALKTRAP



Condensed Graph

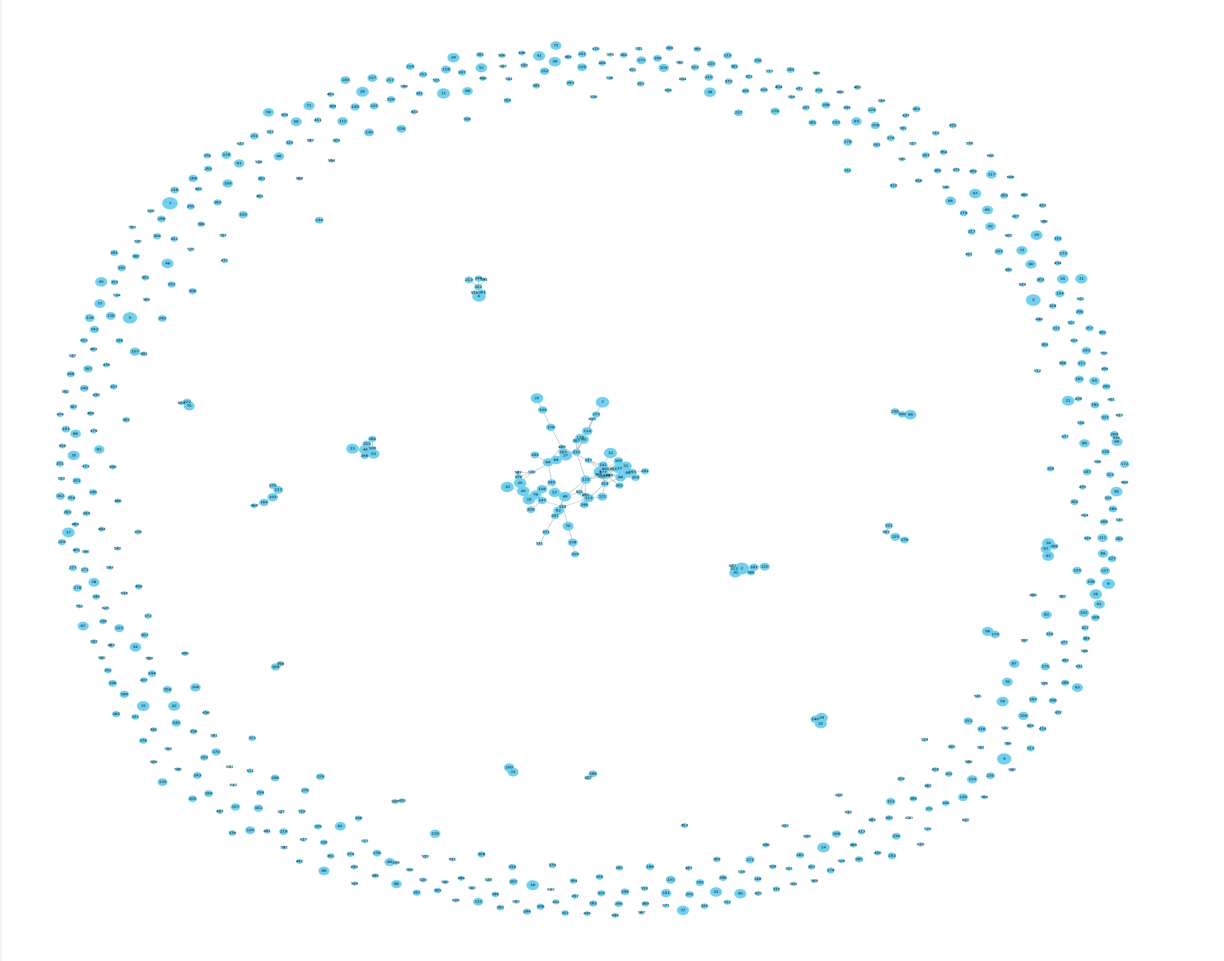


Full Data Graph

Modularity: 0.94
Conductance: 0.08
Number of Communities: 226

COMMUNITY DETECTION – MARKOV CLUSTERING

Full Data Graph

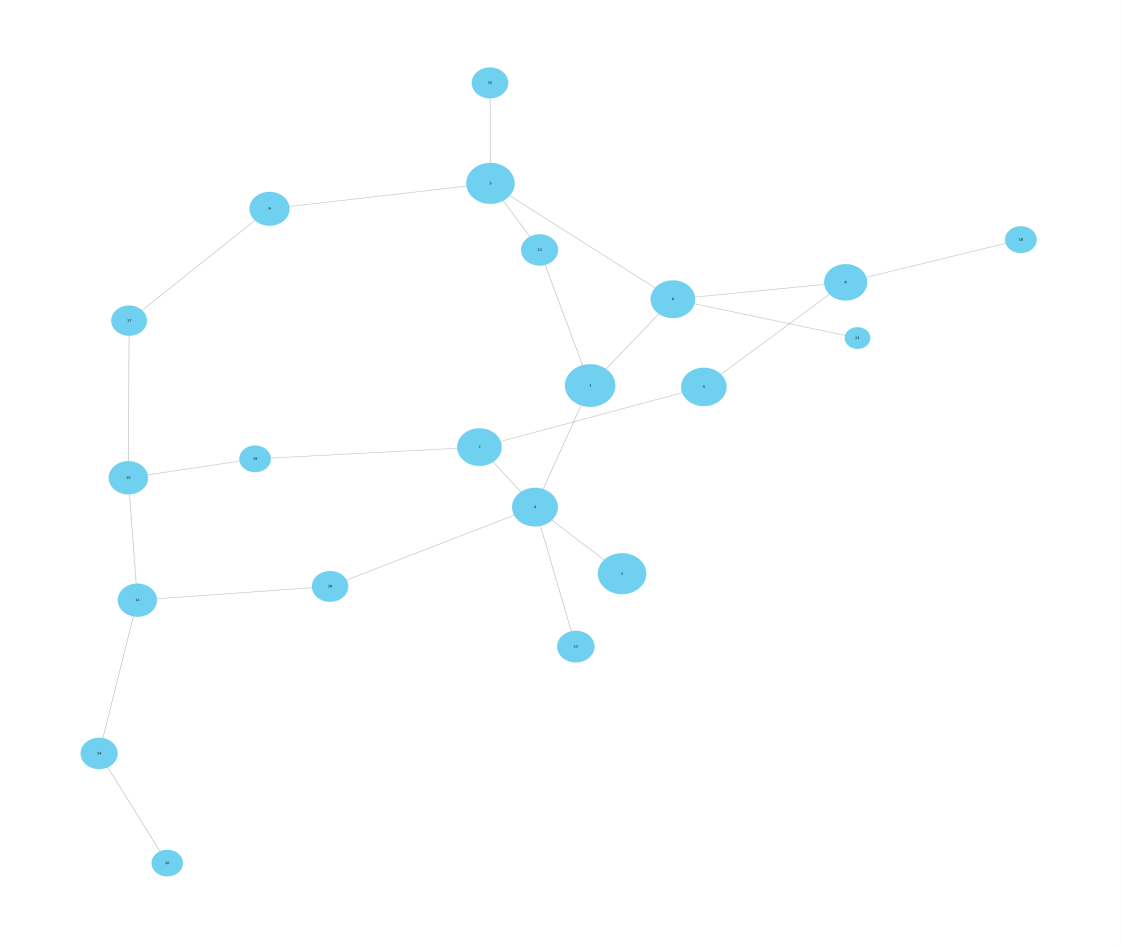


Modularity: 0.001
Conductance: 0.26
Number of Communities: 633

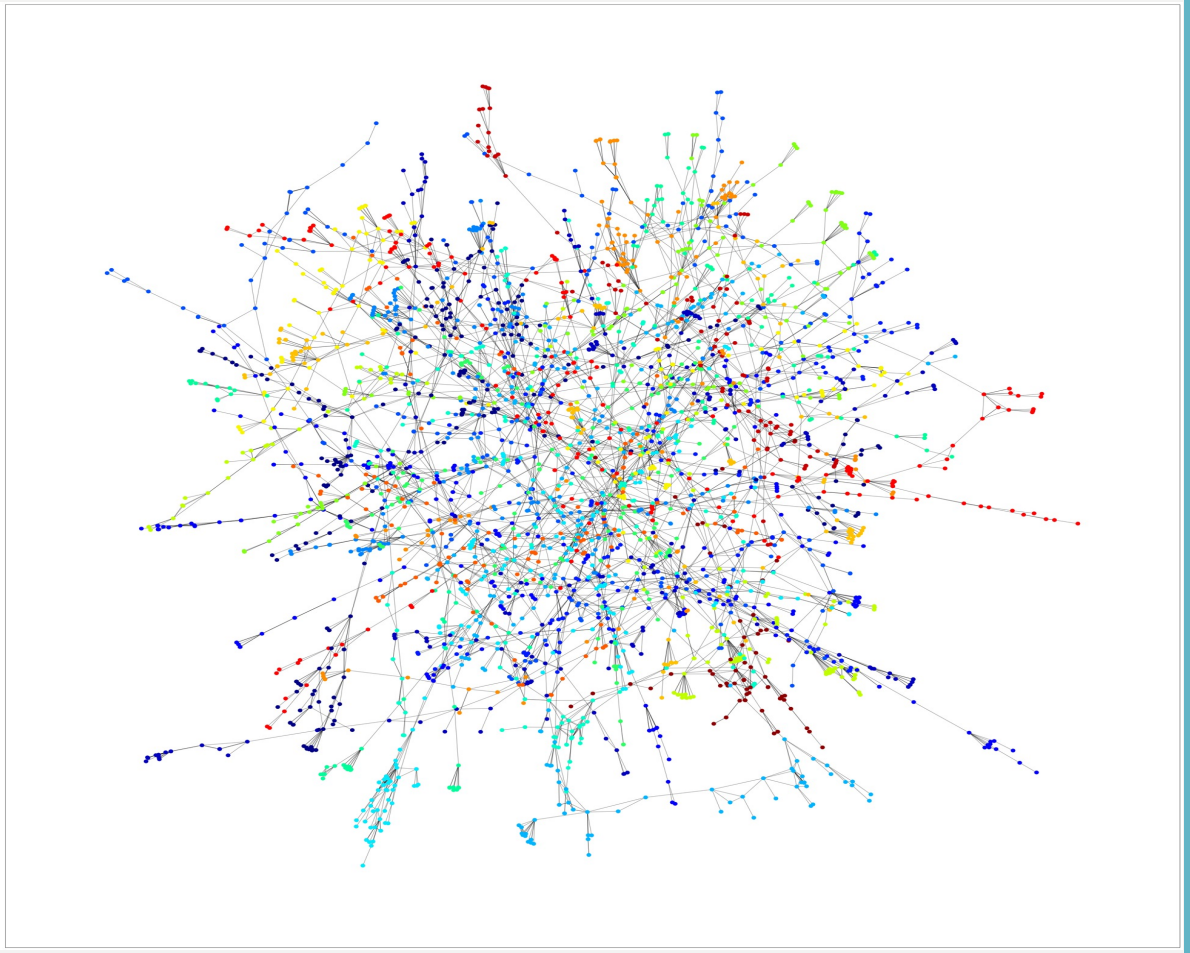
COMMUNITY DETECTION – GIRVAN NEWMAN

Level	Modularity	Conductance
Level 7	0.86	0.002
Level 10	0.89	0.0035
Level 12	0.91	0.0038
Level 14	0.92	0.0043
Level 16	0.93	0.0047
Level 18	0.93	0.0048
Level 20	0.94	0.005

COMMUNITY DETECTION – GIRVAN NEWMAN (L20)



Condensed Graph



Full Data Graph

Modularity: 0.94
Conductance: 0.005
Number of Communities: 21

RESULTS

- **Best performing community detection model**: Girvan-Newman Algorithm (Level 20 - 21 communities) with the modularity of 0.94 and conductance of 0.005.
- Girvan-Newman was chosen the best instead of Louvain (modularity- 0.97) due to the conductance of Louvain (0.015) being higher
- **Worst performing community detection model**: Markov Clustering Algorithm (Communities - 633) with the modularity of 0.001 and conductance of 0.26.

THANK YOU