NATIONAL RESEARCH UNIVERSITY

HIGHER SCHOOL OF ECONOMICS

Graduate School of Business

**Project**

**Analysis of Community Structures and User Interactions in the Facebook Social Graph**

**Group Members:**

Daria Zabelina

Artyom Gotovtsev

**Table of contents**

[**Introduction 3**](#_gu7zwr6wmsph)

[**1 Preprocessing and Data Loading 5**](#_8yye064fnla7)

[**2 Descriptive Statistics and Centralities 6**](#_7q04ldyqis6j)

[2.1 Identify number of nodes and edges 6](#_ndiy8hpscavy)

[2.2 Analyzing circles 7](#_vhw1dqf4jvce)

[2.3 Computing centrality measures 9](#_iyqu83gn6t00)

[2.4 Computing diameter 12](#_c9schg3n2zmx)

[2.5 Computing density 13](#_n29oj0w3ynwz)

[2.6 Visualizing graphs 15](#_j0r69p2gn7zp)

[**Appendix 16**](#_iyxe2x47bjsu)

# **Introduction**

Understanding how people interact and form communities on social media is an important problem that needs research and solutions. This project focuses on Facebook, aiming to uncover how users form groups and interact within these communities. The main problem we are addressing is how to analyze and understand the complex networks of user connections on Facebook. This is important both for scientific knowledge and practical applications, such as improving user engagement and community management on social media.

This topic is very relevant because there is a big gap between what is needed to understand social network dynamics and the current solutions available. By studying these interactions and community structures, we can provide useful recommendations for better community management and user engagement on Facebook.

*The purpose* of this project is to predict the results that will be obtained through the research. These results will include detailed insights into how users form communities and interact within these groups on Facebook.

To achieve this goal, the following *objectives* were set:

– to construct a social graph with nodes representing users and edges showing their interactions;

– to use advanced community detection algorithms to identify different user groups;

– to conduct centrality analysis to find key influencers in the network;

– to analyze interactions to understand the roles and influence of different users;

– to visualize the network using tools like Gephi or NetworkX.

*The object* of this research is the Facebook social graph, which represents the complex web of user interactions and community structures on the platform. *The subject* of this research is the most important aspects of the Facebook social graph, focusing on the properties and features of user interactions and community structures.

The research will use a *mix of methods* to solve these problems, including:

– theoretical methods like analysis, synthesis, comparison, and modeling;

– mathematical methods like quantitative data analysis and modeling.

By using these methods, this project aims to provide a better understanding of community structures and user interactions on Facebook, offering valuable insights and practical recommendations for improving social media strategies,

# **1 Preprocessing and Data Loading**

Firstly, we have downloaded the data from the website[[1]](#footnote-0). Then we have extracted the contents of the *facebook.tar.gz* and saw what files it contains. Then we could load the data with Python and start analysis. The extracted archive contains a directory named “facebook”. The contents on this directory is presented below:

– “.edges” – files containing edges for each node;s ego network;

– “.circles” – files containing circle informations for each node’s ego network;

– “.feat” – files containing feature information for nodes in the edge files;

– “.egofeat” – files containing feature information for the ego nodes;

– “.featnames” – files containing the names of each feature dimension.

After that we observed the available nodes from the directory (see Figure 1).



Figure 1. Available nodes

Node ID represents a specific ego network within the dataset. Each *node\_id* has corresponding files that describe its network structure and features. The corresponding files are the following:

– node\_id.edges;

– node\_id.circles;

– node\_id.feat;

– node\_id.egofeat;

– node\_id. featnames.

# **2 Descriptive Statistics and Centralities**

## **2.1 Identify number of nodes and edges**

Using Networkx we constructed a graph with the *.edges* for the specific *node\_id*. Networkx will automatically manage the nodes based on the edges added. Figure 2 presents the code to count the number of nodes and edges for each *node\_id*.



Figure 2. The number of nodes and edges for each *node\_id*

We have determined the following:

– The ego network for **node 0** contains 333 nodes and 2,519 edges;

– The ego network for **node 107** contains 1,034 nodes and 26,749 edges;

– The ego network for **node 1684** contains 786 nodes and 14,024 edges;

– The ego network for **node 1912** contains 747 nodes and 30,025 edges;

– The ego network for **node 3437** contains 534 nodes and 4,813 edges;

– The ego network for **node 348** contains 224 nodes and 3,192 edges;

– The ego network for **node 3980** contains 52 nodes and 146 edges;

– The ego network for **node 414** contains 150 nodes and 1,693 edges;

– The ego network for **node 686** contains 168 nodes and 1,656 edges;

– The ego network for **node 698** contains 61 nodes and 270 edges.

## **2.2 Analyzing circles**

Figure 3 presents the code to count the number of circles and the average circle size for each *node\_id*.

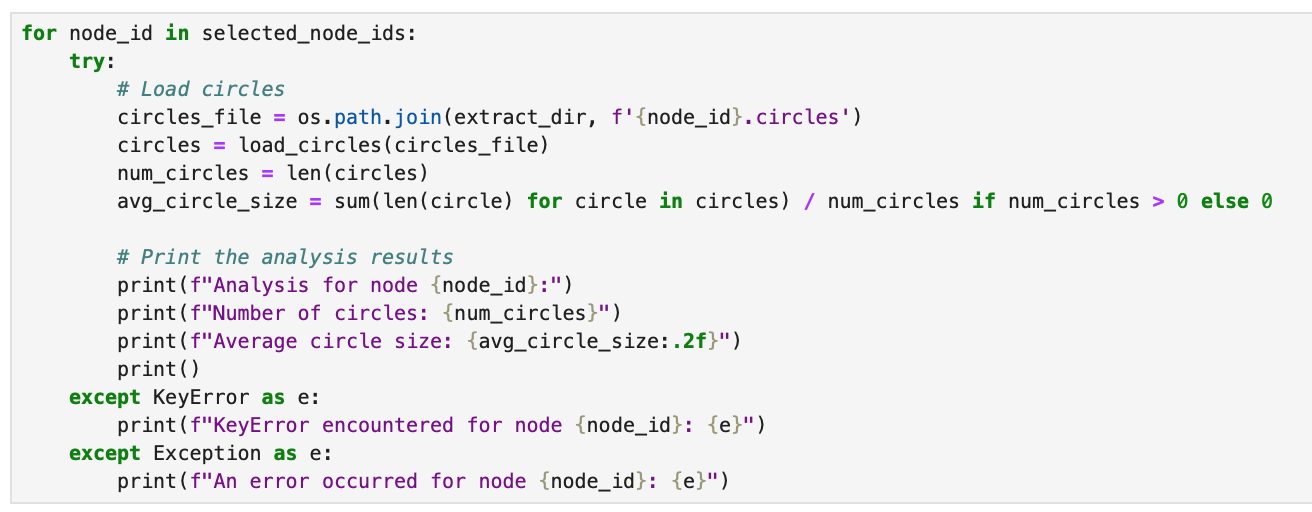


Figure 3. The number circles and average circle size for each *node\_id*

Below we will show the output and draw several conclusions about each *node\_id*.

*Analysis for node 0*

Number of circles: 24

Average circle size: 13.54

**Node 0** has a relatively high number of circles with a moderate average circle size, indicating a diverse network with many small to moderately sized groups.

*Analysis for node 107*

Number of circles: 9

Average circle size: 55.67

**Node 107** has fewer circles, but they are significantly larger. This suggests a network with fewer, but much larger, groups.

*Analysis for node 1684*

Number of circles: 17

Average circle size: 45.71

**Node 1684** has a moderate number of circles with large average sizes, indicating a network with several large groups.

*Analysis for node 1912*

Number of circles: 46

Average circle size: 23.15

**Node 1912** has the highest number of circles, each with a moderate size, suggesting a highly diversified network with many moderately sized groups.

*Analysis for node 3437*

Number of circles: 32

Average circle size: 6.00

**Node 3437** has a high number of circles, but they are relatively small, indicating a network with many small groups.

*Analysis for node 348*

Number of circles: 14

Average circle size: 40.50

**Node 348** has a moderate number of circles, each with a large size, indicating a network with several large groups.

*Analysis for node 3980*

Number of circles: 17

Average circle size: 3.41

**Node 3980** has a moderate number of circles, but they are very small, suggesting a network with many very small groups.

*Analysis for node 414*

Number of circles: 7

Average circle size: 25.43

**Node 414** has the fewest circles with moderately large sizes, indicating a network with a few large groups.

*Analysis for node 686*

Number of circles: 14

Average circle size: 34.64

**Node 686** has a moderate number of circles with large sizes, suggesting a network with several large groups.

*Analysis for node 698*

Number of circles: 13

Average circle size: 6.54

**Node 698** has a moderate number of circles, each with a small size, indicating a network with several small groups.

## **2.3 Computing centrality measures**

After this step we would like to compute the centrality measures: degree centrality, betweenness centrality, closeness centrality and eigenvector centrality. In Figure 4 presented the code to compute the above mentioned measures.

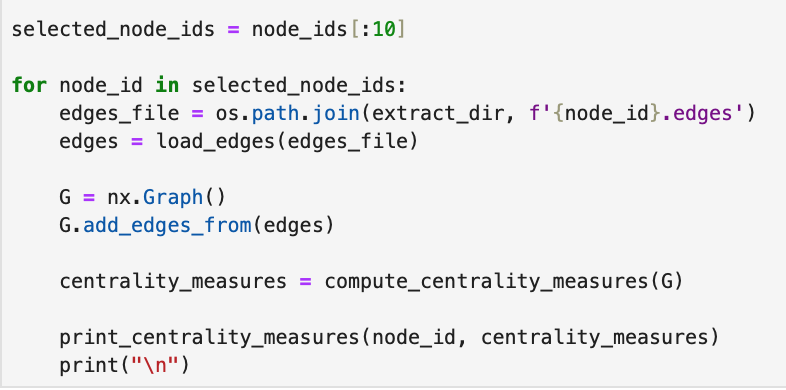


Figure 4. Computing centrality measures for each ego networks

The output is the following:

Centrality measures for the ego network of **node 0**:

The average degree centrality is 0.0456

The average betweenness centrality is 0.0079

The average closeness centrality is 0.2657

The average eigenvector centrality is 0.0302

Centrality measures for the ego network of **node 107**:

The average degree centrality is 0.0501

The average betweenness centrality is 0.0019

The average closeness centrality is 0.3454

The average eigenvector centrality is 0.0158

Centrality measures for the ego network of **node 1684**:

The average degree centrality is 0.0455

The average betweenness centrality is 0.0025

The average closeness centrality is 0.3268

The average eigenvector centrality is 0.0161

Centrality measures for the ego network of **node 1912**:

The average degree centrality is 0.1078

The average betweenness centrality is 0.0021

The average closeness centrality is 0.3981

The average eigenvector centrality is 0.0201

Centrality measures for the ego network of **node 3437**:

The average degree centrality is 0.0338

The average betweenness centrality is 0.0046

The average closeness centrality is 0.2948

The average eigenvector centrality is 0.0205

Centrality measures for the ego network of **node 348**:

The average degree centrality is 0.1278

The average betweenness centrality is 0.0069

The average closeness centrality is 0.4163

The average eigenvector centrality is 0.0476

Centrality measures for the ego network of **node 3980**:

The average degree centrality is 0.1101

The average betweenness centrality is 0.0224

The average closeness centrality is 0.2948

The average eigenvector centrality is 0.0985

Centrality measures for the ego network of **node 414**:

The average degree centrality is 0.1515

The average betweenness centrality is 0.0111

The average closeness centrality is 0.3740

The average eigenvector centrality is 0.0512

Centrality measures for the ego network of **node 686**:

The average degree centrality is 0.1180

The average betweenness centrality is 0.0086

The average closeness centrality is 0.4264

The average eigenvector centrality is 0.0556

Centrality measures for the ego network of **node 698**:

The average degree centrality is 0.1475

The average betweenness centrality is 0.0076

The average closeness centrality is 0.2585

The average eigenvector centrality is 0.0844

Nodes with higher average degree centrality indicate that on average nodes in these networks have more connections. **Node 414** has the highest score (0.1515), **Node 3437** has the lowest one (0.0338)

Nodes with higher average betweenness centrality are essential for information flow as those nodes act like bridges within the network. **Node 3980** (0.0224) plays a significant role in connecting different parts of its network. **Node 107** (0.0019) has less influence on the network’s connectivity.

Nodes with higher average closeness centrality are more central (so, they can reach other nodes more quickly). **Node 348** has the highest score 0.4163 indicating the most centralized structure among other networks. It ensures faster information spread. **Node 698** has the lowest score 0.2585 indicating that it is less centralized.

Nodes with higher eigenvector centrality are more influential, having connection to other well-connected nodes. **Node 3980** has the most influential nodes (score 0.0985) which tells us about highly interconnected and influential structure. **Node 107** has the lowest score (0.0158).

## **2.4 Computing diameter**

In Figure 5 presented the code to compute the diameter for each *node\_id*.



Figure 5. Computing diameters for each ego networks

The results are the following:

**Node 0**: inf;

**Node 107**: 9;

**Node 1684**: inf;

**Node 1912**: inf;

**Node 3437**: inf;

**Node 348**: 9;

**Node 3980**: inf;

**Node 414**: inf;

**Node 686**: 6;

**Node 698**: inf.

The presence of isolated nodes or subgraphs in many networks lead to the idea that there may be significant fragmentation within these ego networks. Fully connected networks with moderate diameters indicate more cohesive structures where information spreads more efficiently.

For nodes with infinite diameter further analysis might be needed to identify and understand the isolated subgraphs or disconnected nodes.

## **2.5 Computing density**

Figure 6 presented the code to compute the density for each *node\_id*.

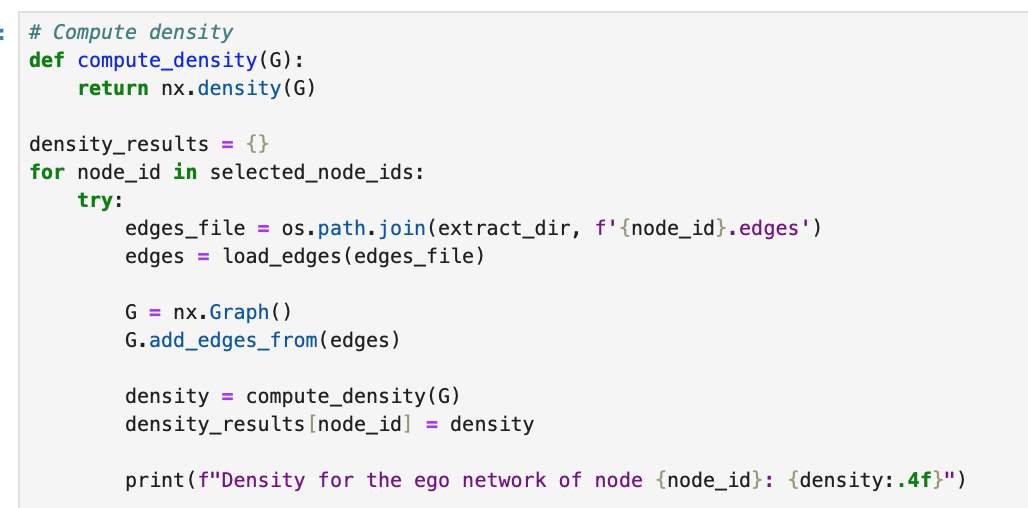


Figure 6. Computing density for each ego networks

The output is the following:

Density for the ego network of**node 0**: 0.0456

The ego network for **node 0** is sparsely connected with only 4.56% of the possible connections.

Density for the ego network of **node 107**: 0.0501

The ego network for **node 107** is a little bit more connected with 5.01% of the possible connections being present.

Density for the ego network of **node 1684**: 0.0455

The ego network for **node 1684** has a similar low density and showed 4.55% of the possible connections.

Density for the ego network of **node 1912**: 0.1078

The ego network for **node 1912** is more connected with 10.78% of the possible connections indicating moderately interconnected network.

Density for the ego network of **node 3437**: 0.0338

The ego network for **node 3437** obtained the lowest score and showed only 3.38% of the possible connections.

Density for the ego network of **node 348**: 0.1278

The ego network for **node 348** obtained a high score of density and showed 12.78% of the possible connections indicating a highly interconnected network.

Density for the ego network of **node 3980**: 0.1101

The ego network for **node 3980** is also highly connected showing 11.01% of the possible connections.

Density for the ego network of **node 414**: 0.1515

The ego network for **node 414** has the highest score for density of 15.15% of the possible connections indicating a well-connected network

Density for the ego network of **node 686**: 0.1180

The ego network for **node 686** is well connected with only 11.80% of the possible connections.

Density for the ego network of **node 698**: 0.1475

The ego network for **node 698** is also highly connected with 14.75% of the possible connections.

## **2.6 Visualizing graphs**

Figure 7 presented the code to visualize each *node\_id*.

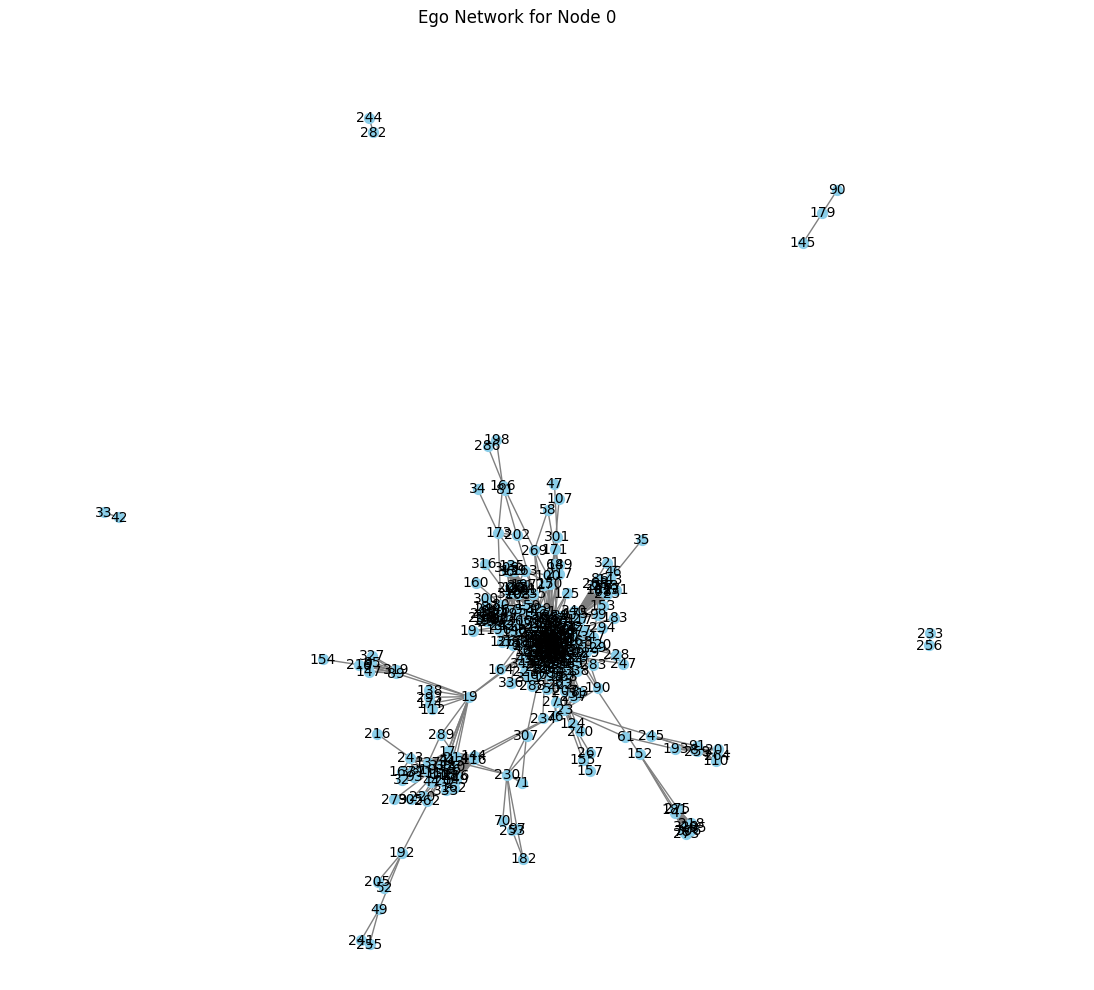


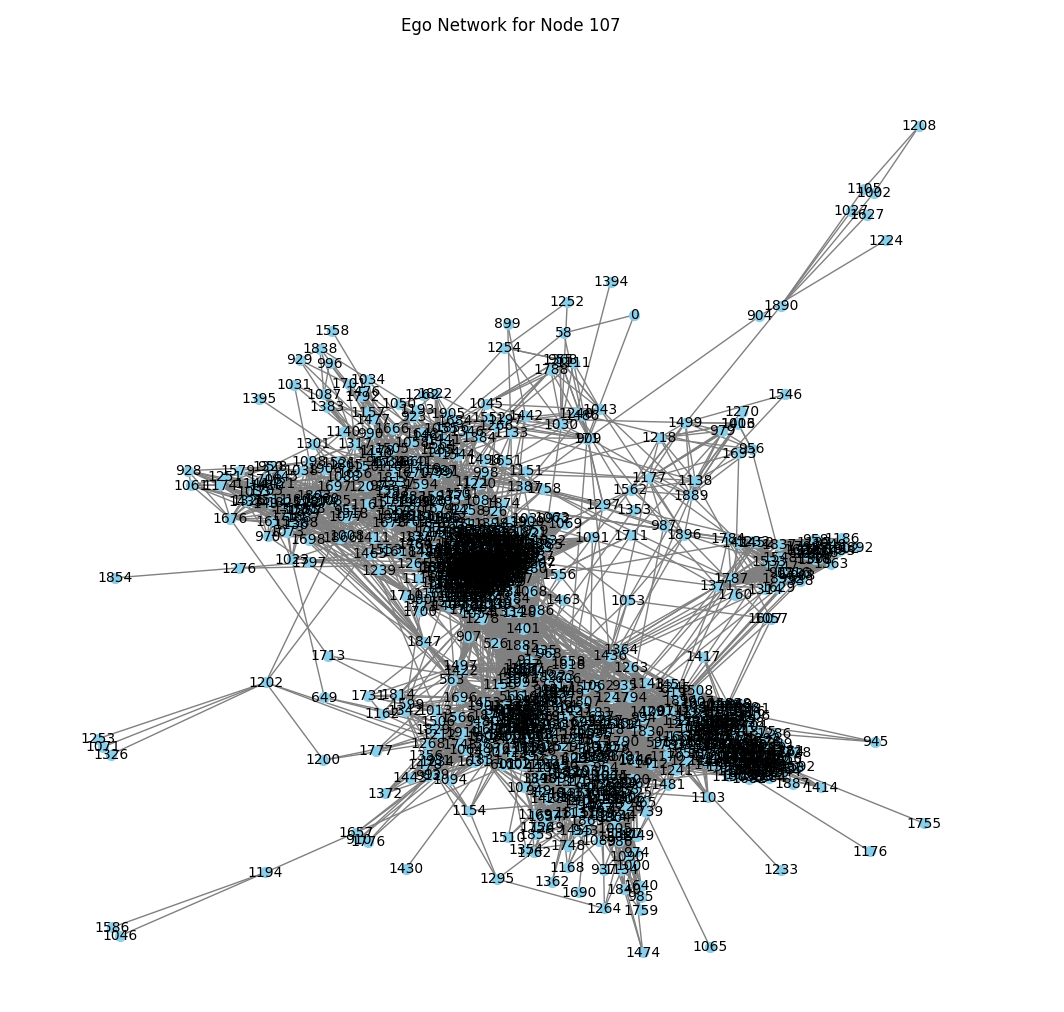
Figure 7. Code for visualization each ego networks

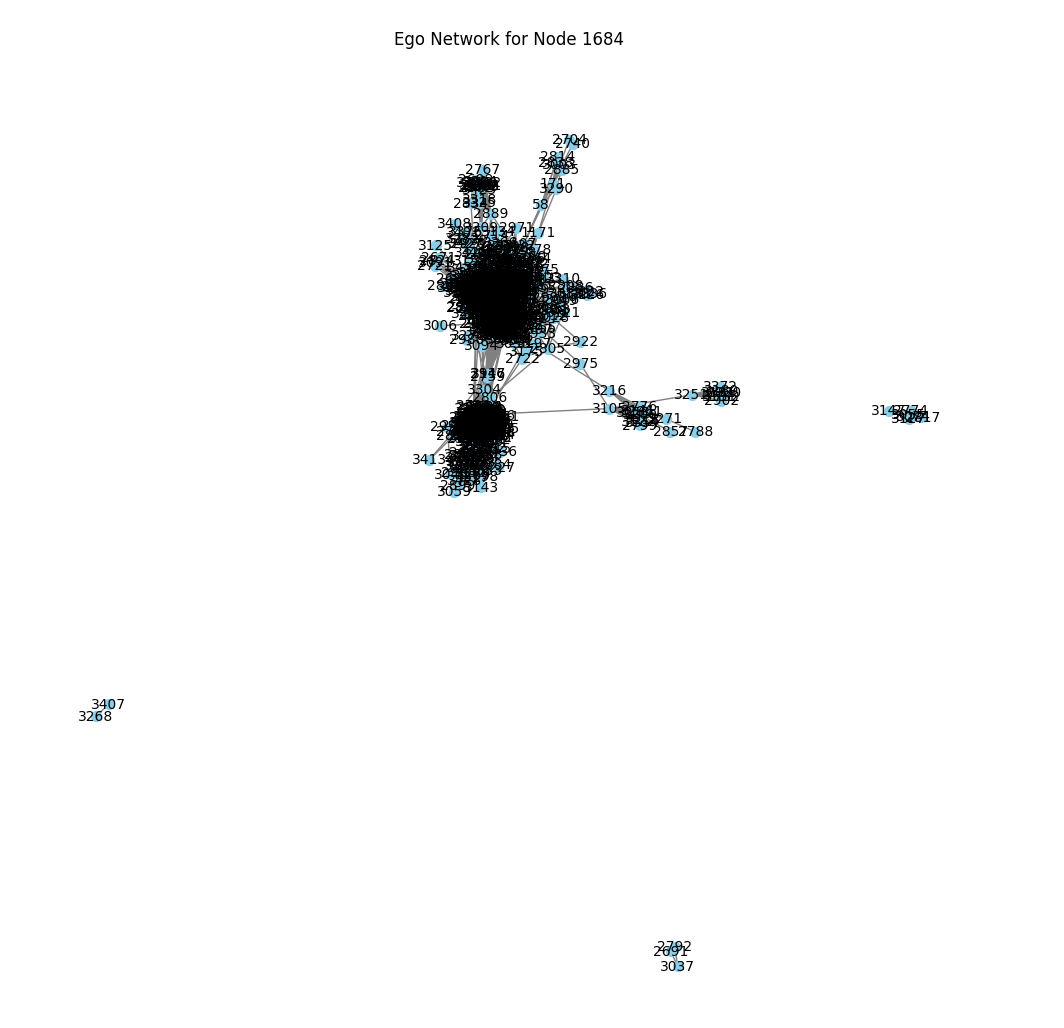
All visualizations are located in Appendix 1.

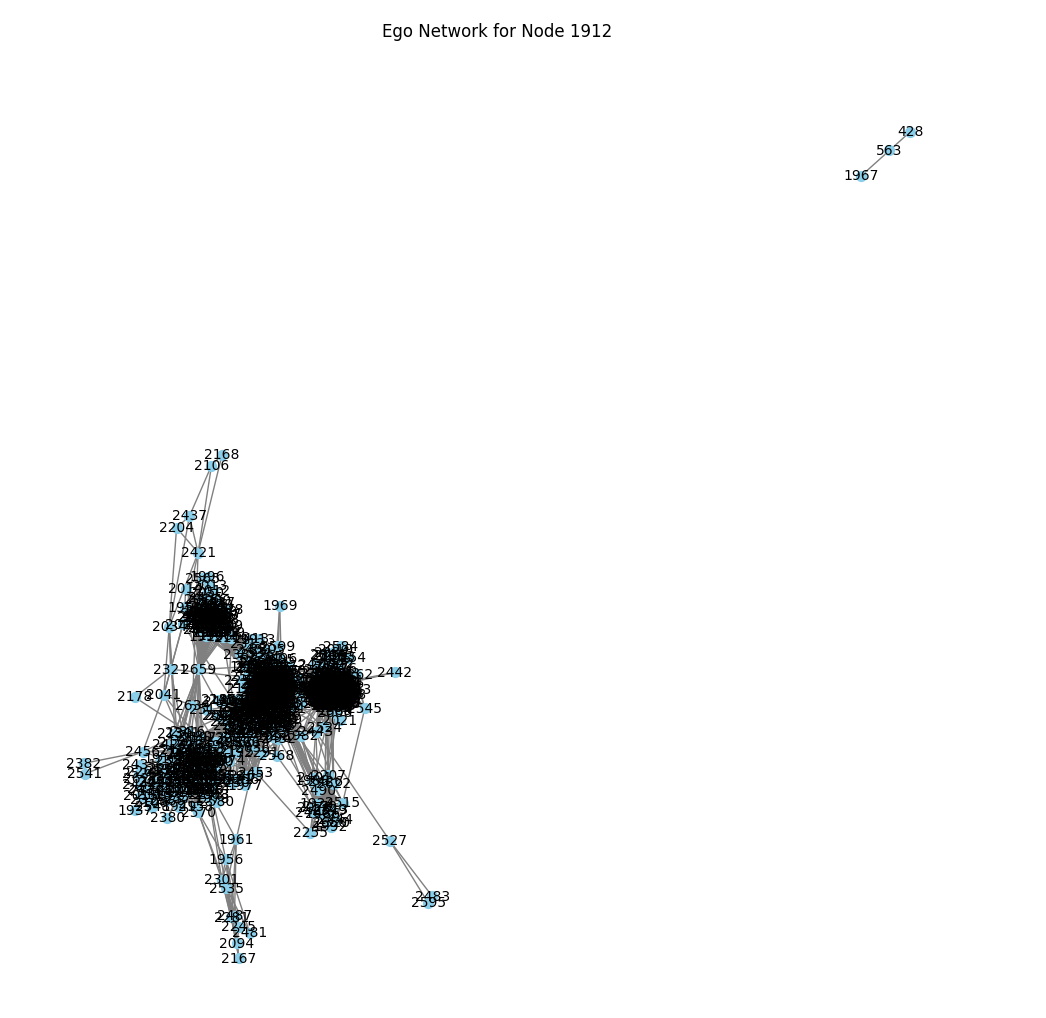
# **Appendix**

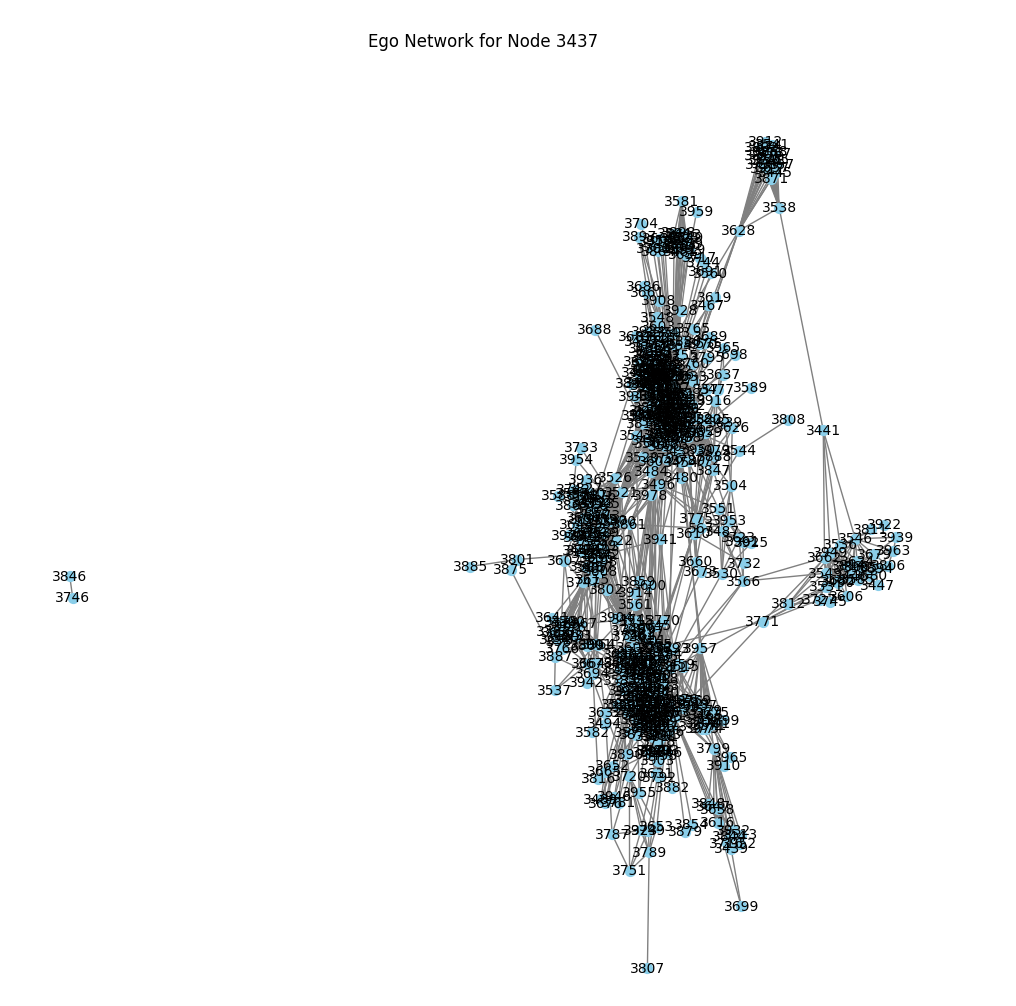
Appendix 1. Ego networks visualization

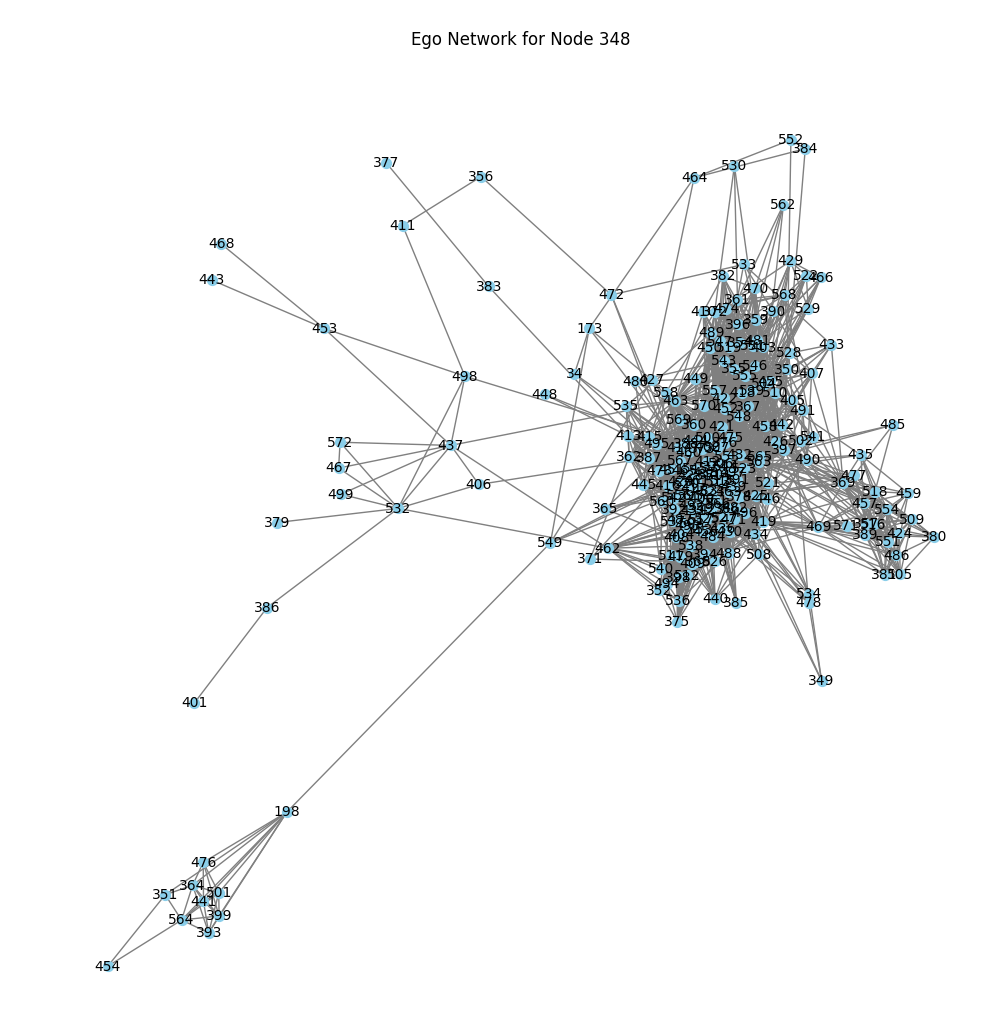


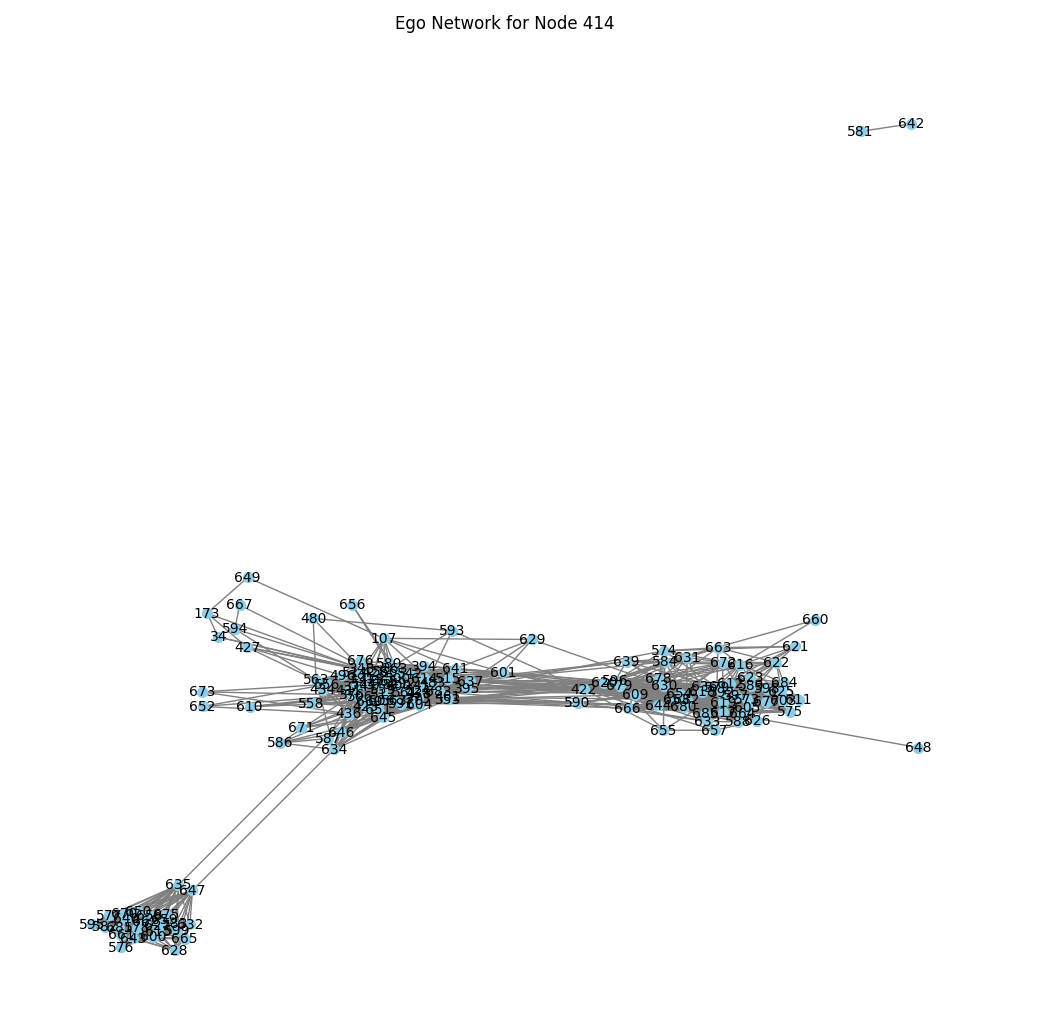
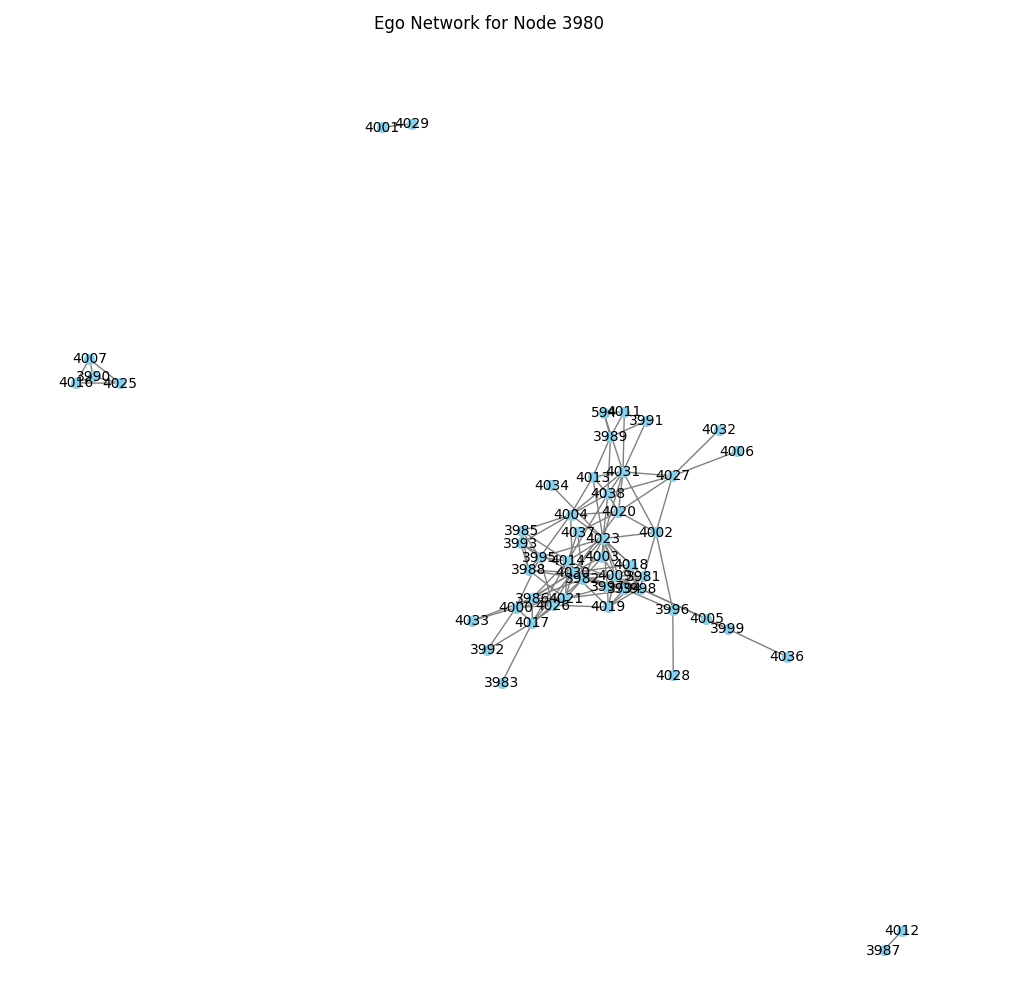


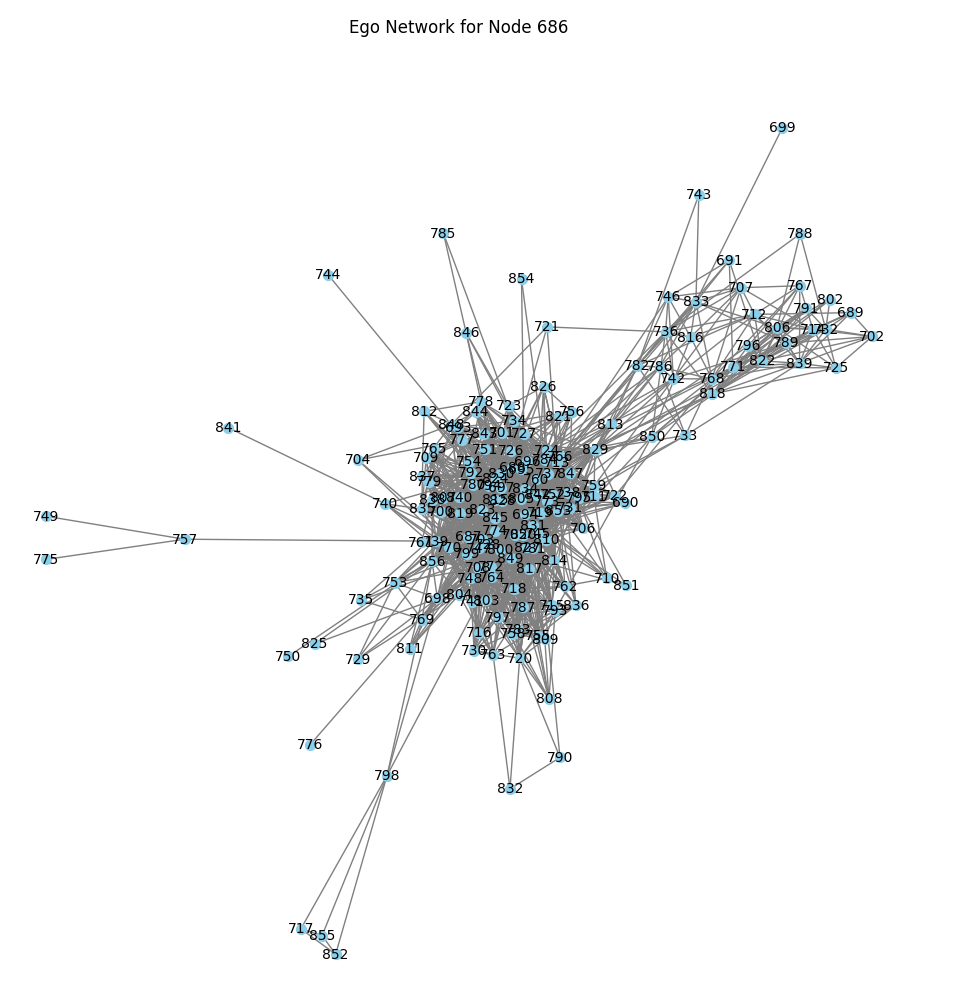


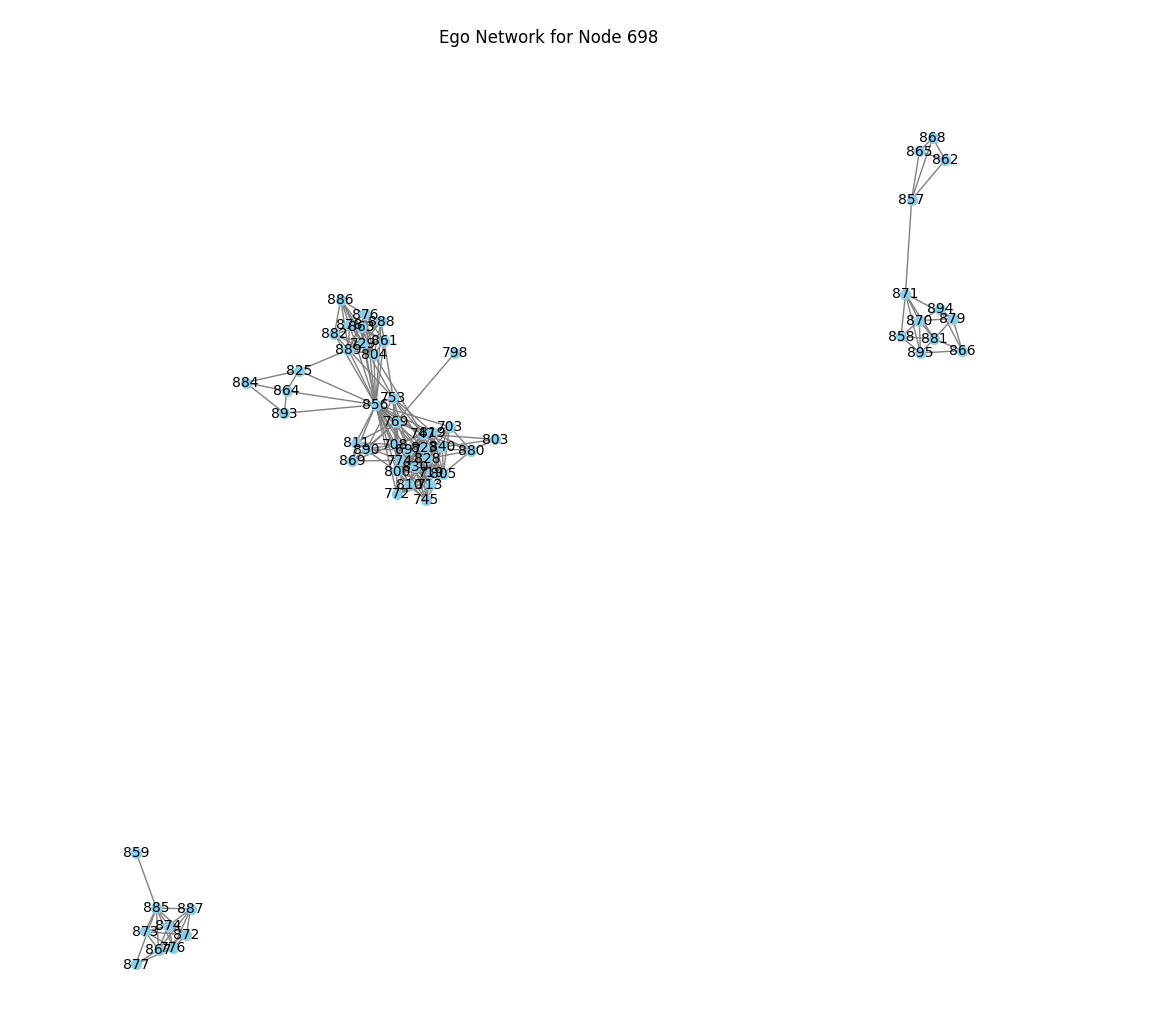


****

****

****

****

****

1. Social circles: Facebook. [Electronic resource] URL: <https://snap.stanford.edu/data/ego-Facebook.html> (access date 14/05/2024) [↑](#footnote-ref-0)