ANALYSIS OF A SOCIAL NETWORK

Sotirios Panagiotis Koulouridis

Theoretical Foundation for Network Analysis

Introduction to Network Analysis

net·work: (n.) an interconnected or interrelated chain, group, or system.

"*Behind each complex system there is a network that encodes the interactions between the system's components*"

Networks are divided in 4 categories :

1. Technological networks
2. Information networks
3. Social networks
4. Biological networks

Network analysis is a powerful framework used to study the relationships and interactions between entities within a system. These entities, called nodes, are connected by edges, which represent the relationships or interactions between them. Networks can model various real-world systems, such as social relationships, communication patterns, transportation systems, or biological systems.

The topic we are more interested in is Social networks . In Social networks, nodes represent individuals, and edges represent any kind of relationship between them. With the rise of social media platforms, the study of social networks has become a subject of research, as it allows us to draw many specific conclusions , as we can uncover patterns, identify influential entities, detect communities, and understand the structural properties of the system.

Generally, the best way to analyze a complex social network is to visualize it first. Once we understand the network's structure, with its nodes and edges, we try to calculate global metrics.Then, we find local metrics for each node and visualize the network again based on our findings.

Key Concepts in Network Analysis

To understand and analyze networks effectively, several fundamental concepts and metrics are essential:

Nodes and Edges:

- Nodes: Represent the entities in the network (e.g., people).
- Edges: Represent the connections or relationships between nodes. Edges can be directed (e.g., follower relationships) or undirected (e.g., mutual friendships) and may have weights to indicate the strength or intensity of a relationship (e.g. likes).

Strongly Connected Network :

- A directed graph (or network) is strongly connected if every node is reachable from every other node, following the direction of the edges.

Weakly Connected Network :

- A directed graph is weakly connected if replacing all its directed edges with undirected edges makes the graph connected.

Degree:

- The degree of a node is the number of connections it has.
- In-degree: The number of edges directed toward a node.
- Out-degree: The number of edges directed away from a node.
- Degree distribution provides insight into the overall connectivity of the network and can help identify highly connected nodes (hubs).

Centrality Measures:

Captures the idea of how central a node is in the network:

- Degree Centrality: Based on the number of direct connections a node has.
- Betweenness Centrality: Measures how often a node lies on the shortest path between other nodes, highlighting nodes that act as bridges.
- Closeness Centrality: Reflects how close a node is to all other nodes in terms of shortest paths.
- Eigenvector Centrality: Assigns importance to nodes based on their connections to other important nodes.

Clustering Coefficient:

- Measures the tendency of nodes to form tightly-knit groups or "triangles." A high clustering coefficient indicates that the neighbors of a node are likely to be connected.

Bridges:

- An edge is a bridge if removing it increases the number of connected components in the graph. Bridges often represent critical connections in the network . They have a crtitical role in information flow and vulnerability.

Triadic Closure :

- If two nodes *A* and *B* are both connected to a common node C, there is a tendency for A and B to also form a connection.

PageRank:

- A variant of eigenvector centrality that measures the importance of nodes based on the quality and quantity of incoming connections.

Modularity and Community Structure:

- Modularity measures the strength of division of a network into communities or clusters. Communities are groups of nodes that are more densely connected internally than with the rest of the network.

Graph Density:

- Density quantifies how connected the network is. It is the ratio of the number of edges to the maximum possible number of edges.

Homophily:

- Homophily describes the tendency of nodes with similar attributes (e.g., gender, location) to connect more frequently. Understanding homophily provides insights into the social dynamics of a network.

The Role of Gephi in Network Analysis

Gephi is an open-source software tool designed specifically for network visualization and analysis. It provides an intuitive interface to import, analyze, and visualize networks, making it accessible for users with various levels of technical expertise. Some key features of Gephi include:

Graph Visualization:

- Gephi offers various layout algorithms (e.g., ForceAtlas2, Fruchterman-Reingold) that arrange nodes in visually meaningful ways, helping to uncover patterns and structures within the network.

Statistical Analysis:

- Gephi can calculate important metrics like degree distribution, clustering coefficient, centrality measures, and modularity directly within the software.

Partitioning and Ranking:

- Nodes and edges can be partitioned or ranked based on attributes (e.g., gender, degree, or centrality), allowing for targeted visual analysis.

Community Detection:

- Gephi's modularity algorithm identifies communities and assigns nodes to clusters, which can then be visualized using distinct colors.

Export Options:

- Gephi allows users to export their visualizations as images or their metrics as datasets, making it suitable for both presentation and further analysis.

Conclusion

Network analysis provides a structured way to study relationships and interactions within complex systems. By using tools like Gephi, we can visualize and quantify the properties of networks, uncover hidden structures, and derive meaningful insights. This theoretical foundation underpins the practical application of network analysis in this assignment, where we aim to analyze a social network, calculate key metrics, and visualize it.

Empirical Validation for Network Analysis

At this point, we will analyze a social network, specifically the one formed by the followers and followings of the well-known actor Mark Hamill. The main goal of the analysis is to understand who the key users influencing the network are, whether there are clusters within the network, and if Hamill acts as the central hub .

Fetching of data

For our analysis, we used Gephi as our tool. To fetch the data of the social network, our initial intention was to use the Bluesky Gephi plugin; however, it proved to be malfunctioning. This challenge led us to use Python for fetching the data. As a result, in addition to retrieving followers, we were able to identify whether there are interactions via likes between two connected users, and the gender of the users. The API used for this process was Atproto.

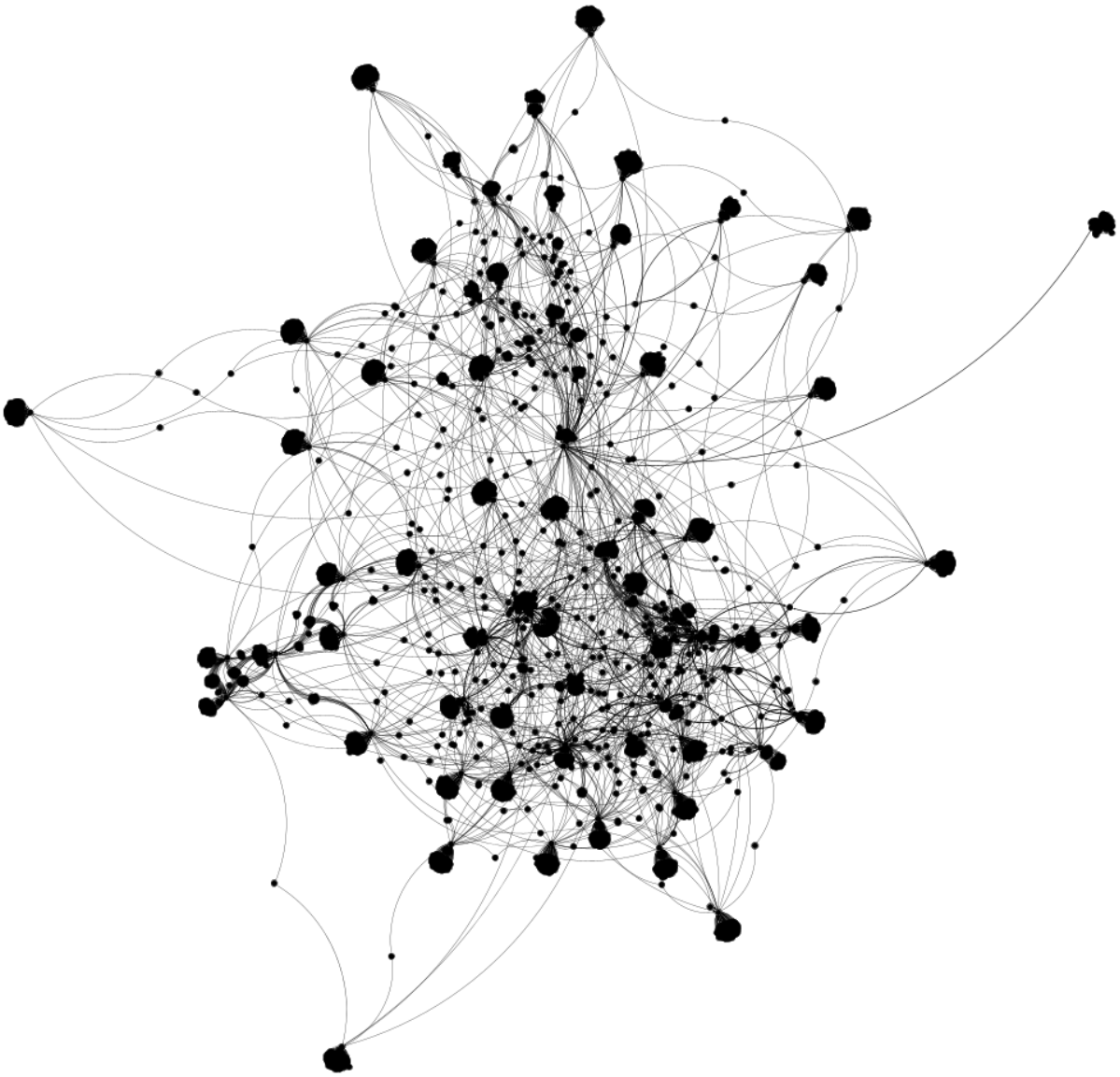
We used 3 python files:

- Fetch.py in order to get the wanted results and write them in a csv . The csv includes searches for up to 50 connections for each user and the depth of the analysis is 2 .
- Preproccess.py in order to clean the data and split into nodes and edges , a format that is suitable for Gephi , getting them ready to be imported as spreadsheets .
- Gender-Finder.py in order to obtain a prediction of the user's gender based on their display name.

Next step is loading nodes.csv and edges.csv into Gephi .

Graphical Representation of the network

Our nodes and edges are loaded . We will use ForceAtlas2 as a layout . At first , we will not use any filter or color for our network .



Then, we visualize the graph with edge colors representing interaction strength: edges with weight 2 (mutual likes) appear distinct from those with weight 1 (non-mutual interactions), enhancing the clarity of relationship intensity within the network.

Basic topological properties, such as numbers of nodes and edges, network diameter, and average path length.

- Number of Nodes: 4770
  Number of Edges: 6116

- Diameter: 13
  Radius: 0
  Average Path length: 5.445674786378045

A diameter of 13 suggests that the farthest two nodes in the network are separated by 13 steps. The network is relatively spread out , but this may be typical for BlueSky .

The radius of 0 implies that some nodes are unreachable .

The average path length is the average of the shortest paths between all pairs of nodes. A number of 5.4 means we have some very sparse regions in our network .

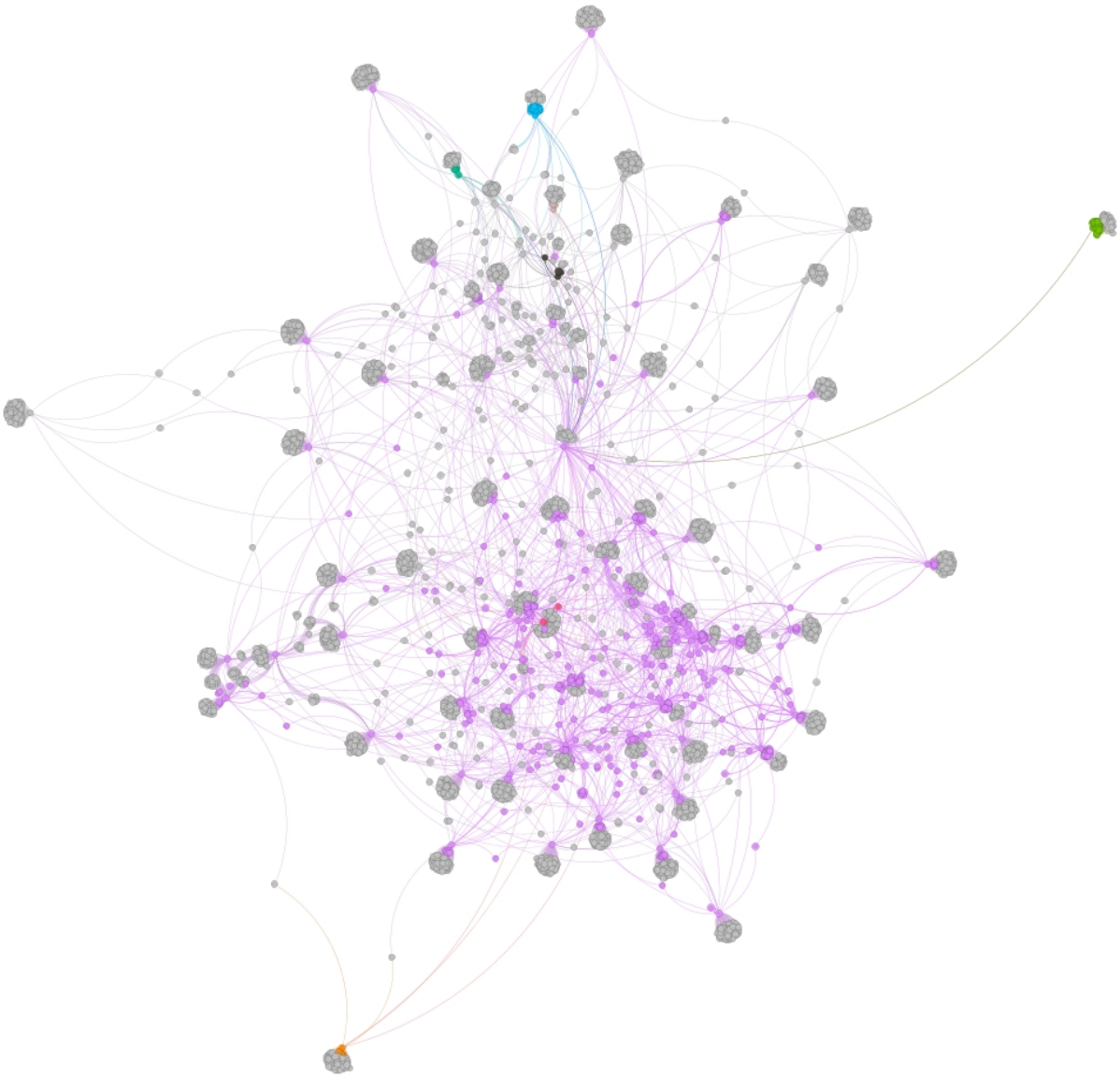
Component measures, such as number of connected components, existence of a giant component and component size distribution.

There's only 1 weakly connected component, so the network is fully connected (a giant component). The network is highly cohesive .

There are 4169 strongly connected components . The mutual reachability of the network is very limited, typical for social platforms like BlueSky , that a lot of asymmetric relationships are created .

There is a giant component. In a social network, the giant component indicates the largest group of users who can communicate or interact with each other. We have partitioned our graph in Strongly-Connected IDs to view this component better. Giant component is relatively small, so our network is fragmented.

Giant Component (pink color nodes) often contains the most important (influential) nodes of our network. Mark hamill is a part of it .

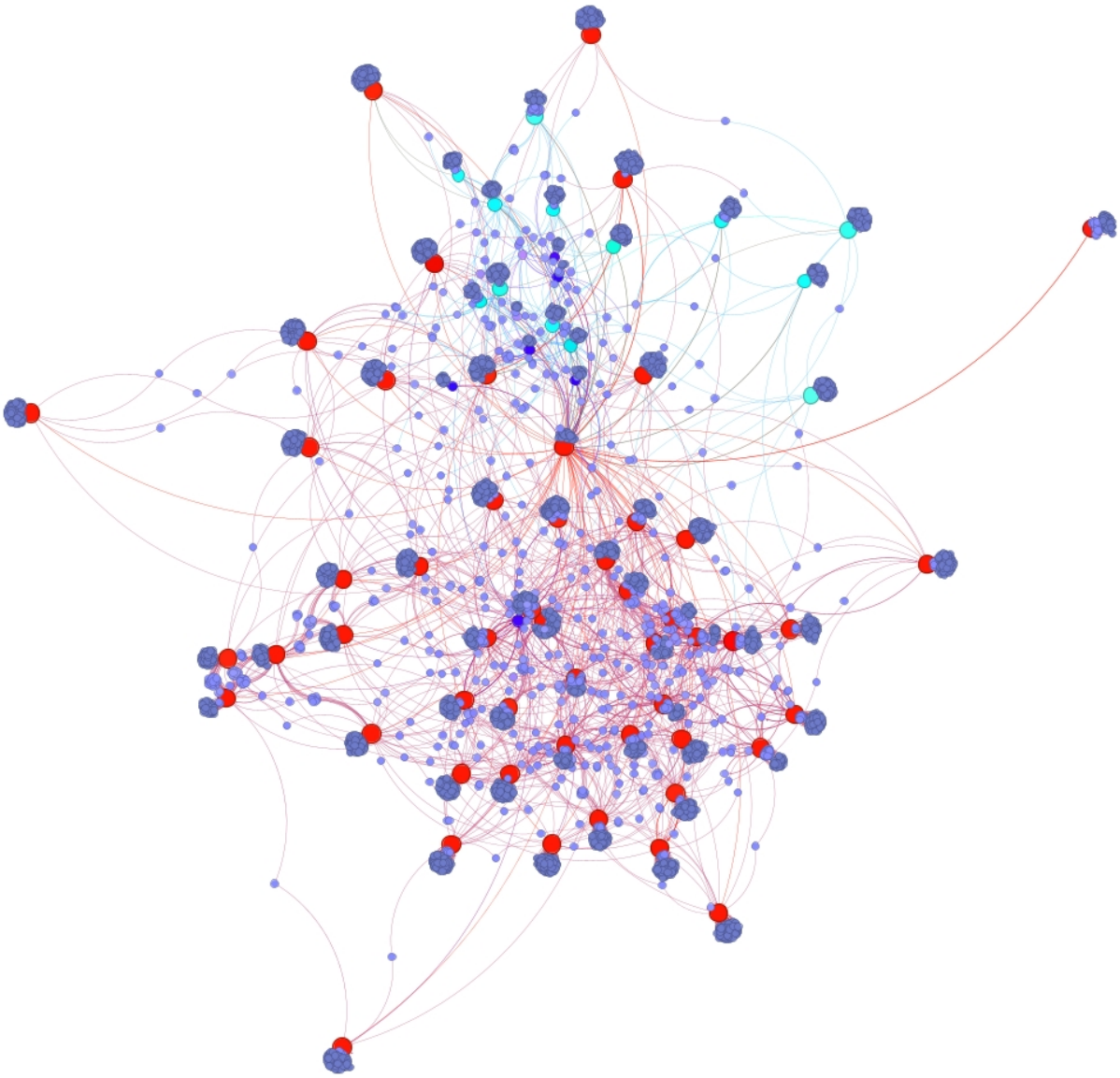


Degree measures, such as maximum and average node degrees, as well as degree distribution.

In our data we have 50 nodes with a degree of more than 100 each. These nodes are considered the possible influencers of our network . The maximum degree is 106 .

Average degree is 1.282. Average user (node) is low degree and is getting influenced by the higher degree users.

Bellow, the graph shows the degree distribution of the network we analyze .The nodes size is set according to the degree.With red colour are the users that are considered the influencers(degree of 100 +), with cyan nodes with degree of 50-99, with blue nodes with degree of 15-49 and with purple nodes with a degree of less than 15.

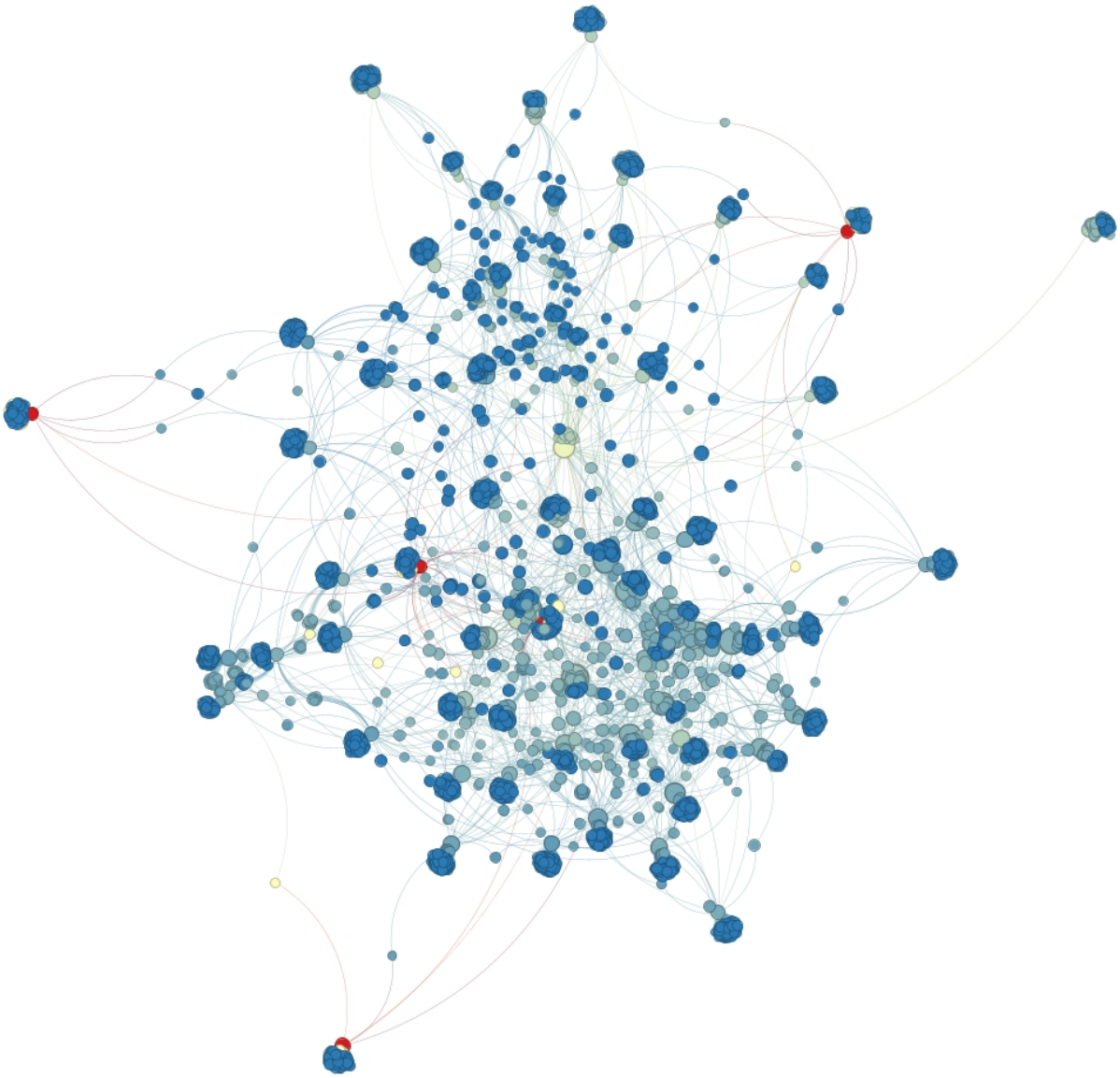


Centrality measures (degree, betweenness, closeness, eigenvector)

We used centrality measures for our social network. Initially, we present a graph where the node size depends on eigenvector centrality, and the node color is based on closeness centrality (blue to red).

Eigenvector Centrality: We derive significant insights into our social network. This measure indicates which and how many important nodes in our network are more strongly connected to other important nodes. In other words, it helps us understand which users have a higher degree of significant connections.

Closeness Centrality: A high Closeness score means the node is near the center of the network, while a low score means it is far from most nodes. Nodes with high scores can efficiently spread information to the entire network (low shortest path distances) , while those with low scores are hard to reach the others.



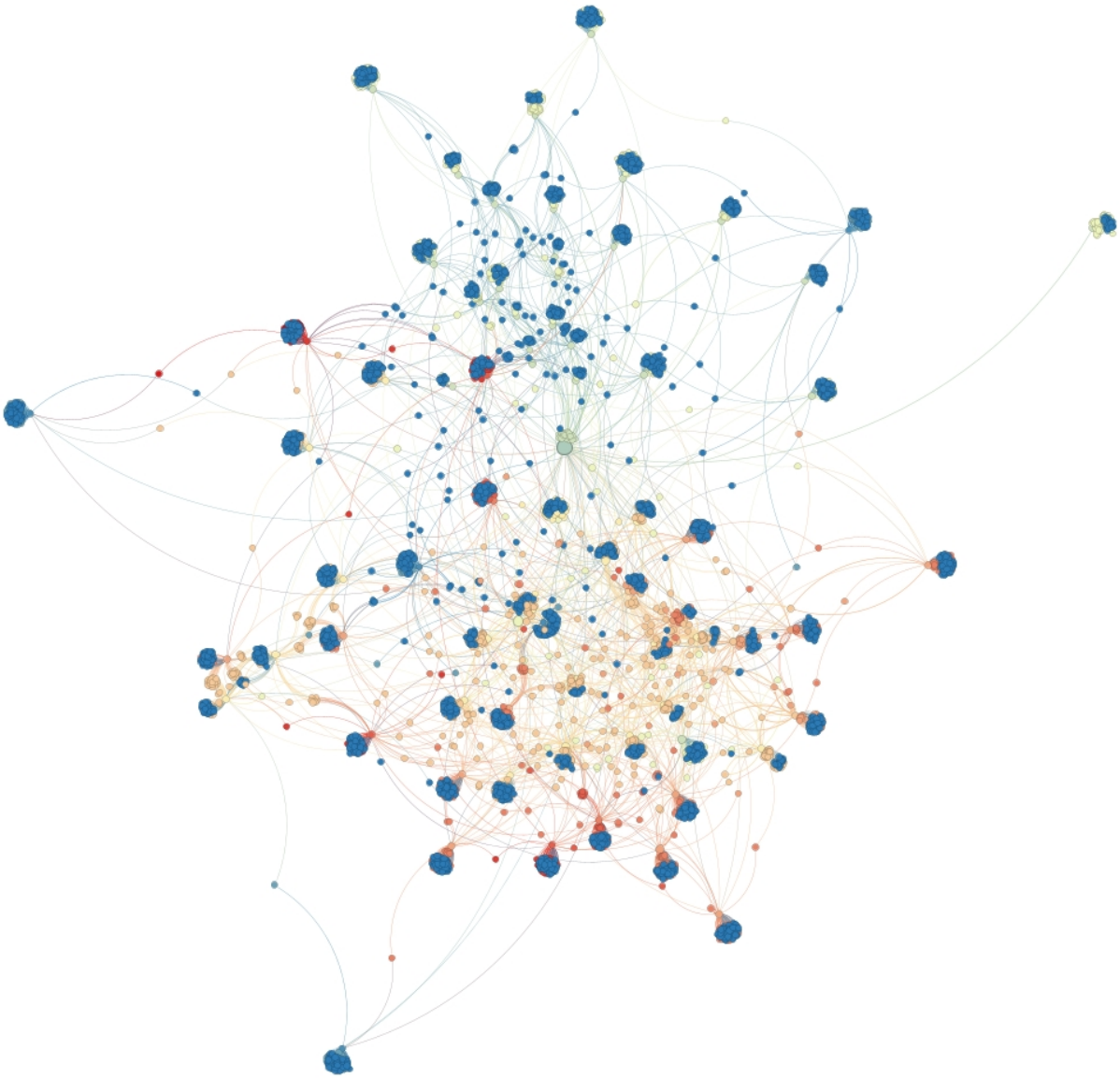
We observe that we have:

- Large and red nodes: These are the most important nodes in the network for information transfer.
- A single large but yellow node: This node has a medium closeness centrality but remains crucial for information transfer due to its high eigenvector centrality. This node represents **Mark Hamill**.
- Small and yellow nodes: These nodes have low influence but play a significant role in spreading information.
- Large but blue nodes: These nodes are influential but may be located more peripherally within the network.
- Small and blue nodes: These nodes have low influence and are distant from the network's core.

Next, we use betweenness centrality and eccentricity measures. We present a graph where the node size represents betweenness centrality, and the node color is based on eccentricity (blue to red).

Betweenness Centrality: This measures how often a node appears on the shortest path between two other nodes. Nodes with high betweenness control the flow of information and act as bridges between different parts of the network or between communities.

Eccentricity Measures: Nodes with low eccentricity are centrally located and can reach all other nodes quickly, whereas those with high eccentricity are more isolated.



Core nodes (blue, large) are well-connected, while peripheral nodes (red, small) are more isolated. Some bridging nodes (yellow/orange, medium size) connect different communities. The network relies on the core nodes for efficient communication.

Mark Hamill(green,large) is well-connected to other influencers but not at the absolute shortest reach to everyone.

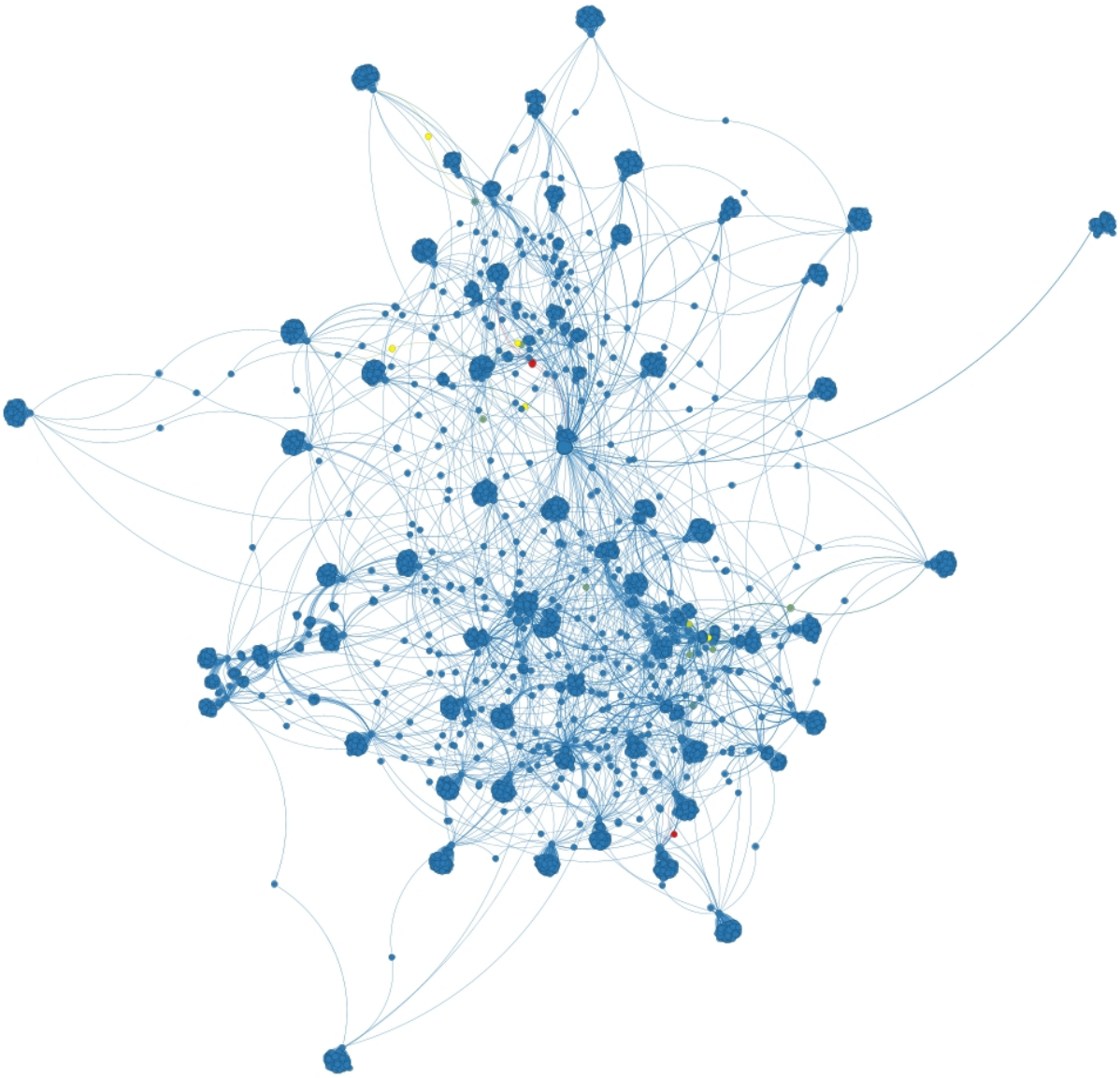
Clustering effects in the network: average clustering coefficient, number of triangles, clustering coefficient distribution, existence of the triadic closure phenomenon in the friendship neighborhood.

After analyzing the clustering coefficient, we obtained the following results:

- Average Clustering Coefficient: 0,001
- Number of triangles: 42 Number of paths (Length 2): 249025 Value of Clustering Coefficient: 5.059732939116657E-4

A clustering coefficient close to 0 suggests that the network does not form many tightly-knit groups. Most nodes are connected to others, but their neighbors do not interact much with each other.

Triadic closure occurs when two nodes with a mutual friend also become friends. This rarely happens in our network. But, the large value of number of paths indicates that most nodes are still connected indirectly.



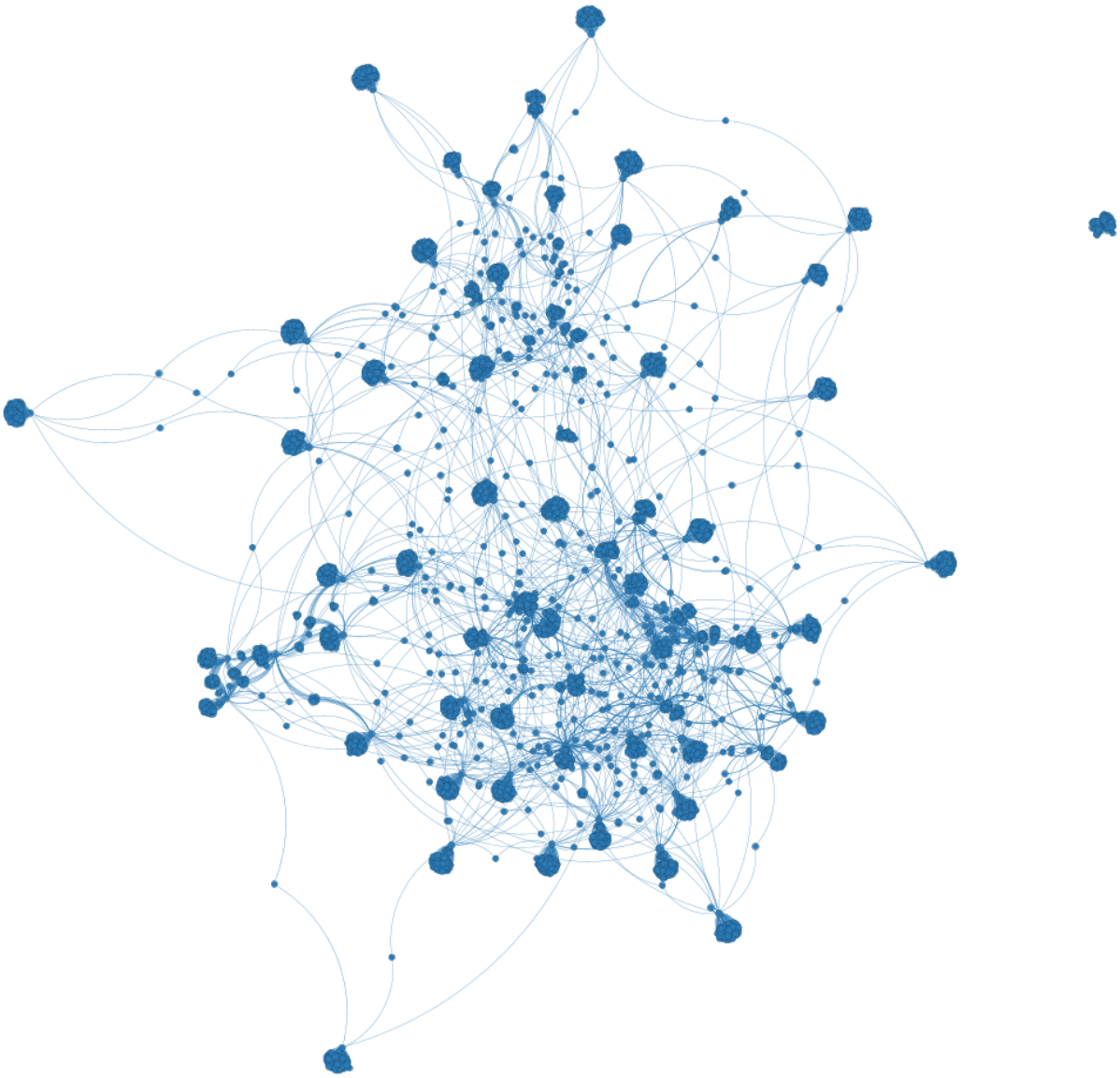
Many nodes are blue, indicating low local clustering and fewer closed triangles.
A small number of red/yellow nodes are present, showing some highly clustered sub-networks.

Bridges and local bridges.

Local bridges exist when two connected nodes have no mutual neighbors. Since the clustering coefficient is so low, many nodes do not share common connections, making local bridges more common in your network.

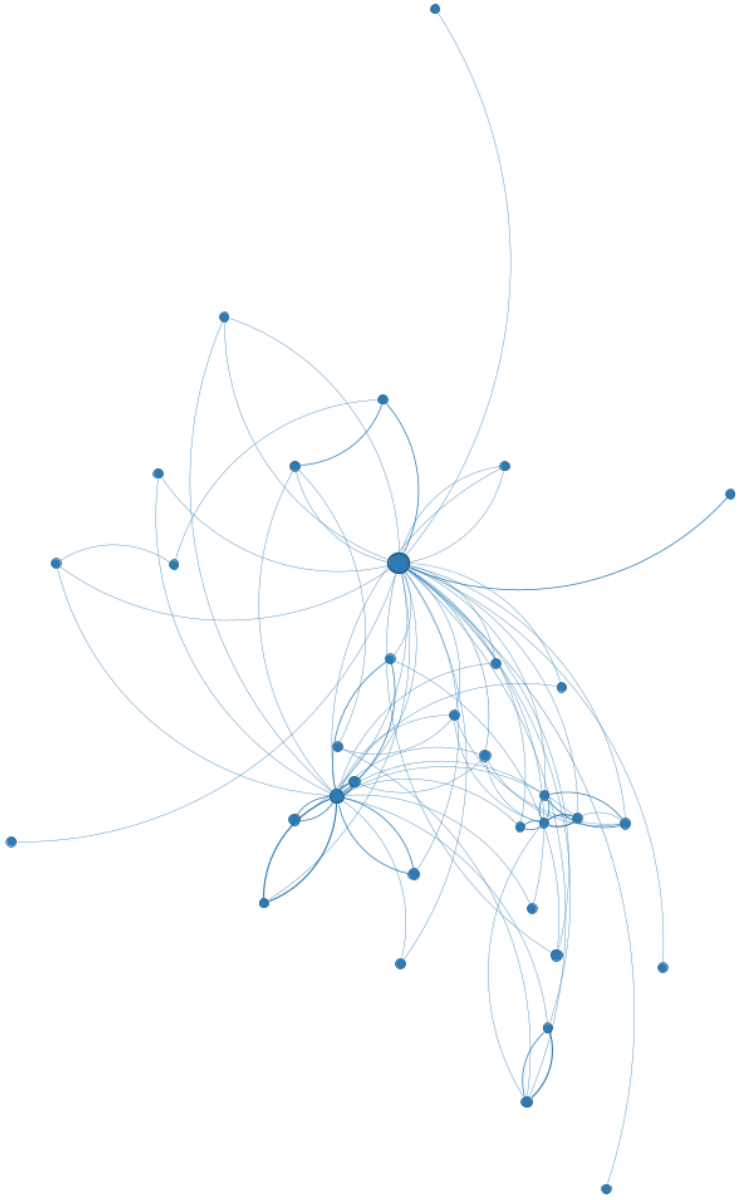
To identify potential local bridges, we applied a filter to display nodes with the lowest possible clustering coefficient values (filter → range → clustering coefficient), based on the rationale that these nodes are likely to be less embedded and may function as local bridges.

The edges displayed in the diagrams represent potential local bridges.



Following a similar approach, to identify potential bridges, we applied a filter to display nodes with the top 1% highest betweenness centrality values (filter → range → Betweenness Centrality).

The most important nodes will be crucial for the network's structure. If their connections are severed, it is likely that subnetworks will form.



Gender and homophily.

Initially, we apply filters to generate visual representations of:

- Connections only between females
- Connections only between males
- The overall network

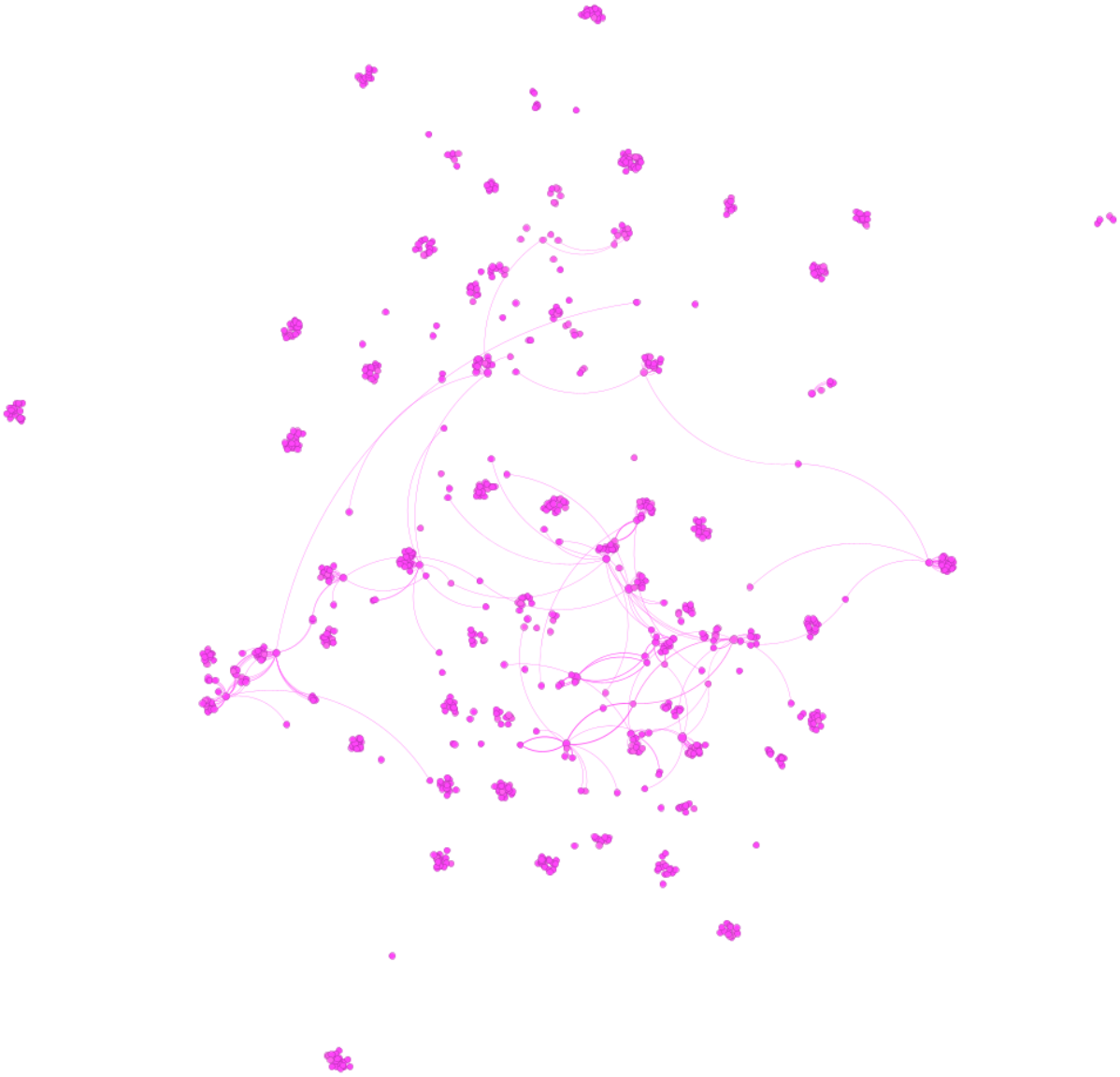
Color coding:

- Females are distinguished by pink or purple
- Males are represented in blue

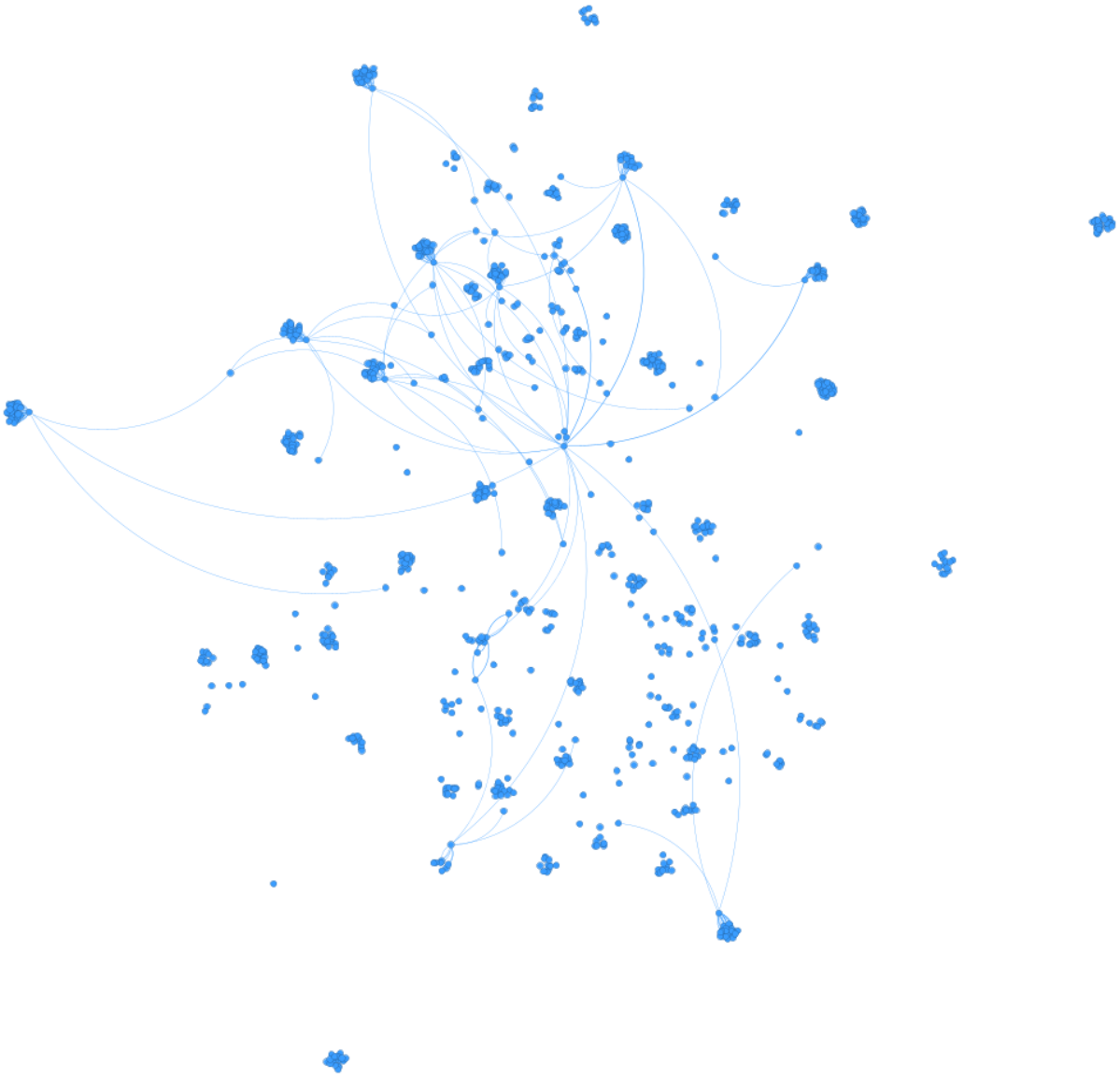
This approach helps us visualize the gender distribution within the network and analyze their connections.

The filter used: Equal (gender) .
(Filters → Attributes → Equal → gender)

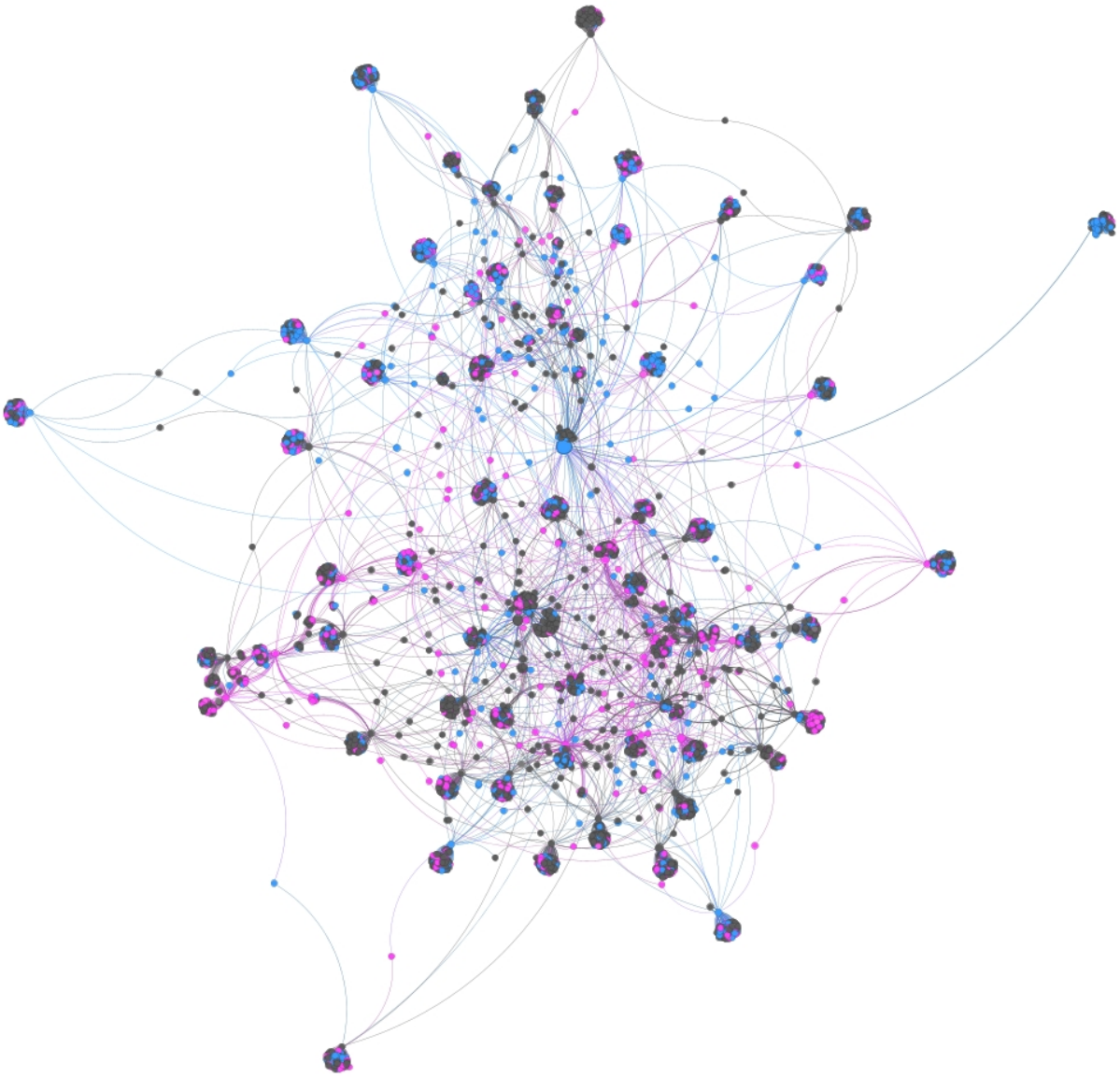
Representation of female-to-female connections:



Representation of male-to-male connections:



Overall graph, where node color represents gender (pink for female, blue for male) and node size is proportional to betweenness centrality :



Below, the exact number of connections has been calculated for:

- Male-Male Edges: 308
- Female-Female Edges: 362
- Male-Female Edges: 651

The network shows some homophily (more Male-Male and Female-Female connections) but also a healthy number of Male-Female interactions. It is almost balanced, as the numbers of same-gender and cross-gender connections are very close.

Graph density

Gephi calculates the network density as 0.000, while the exact value is 0.000537. This indicates that we have a very sparse network, where the vast majority of possible connections do not exist.

This suggests that we likely have highly connected influencers, while many other users remain isolated. A very low density is an indication that we may later identify several separate communities, where users will interact more within their own clusters rather than across the entire network.

On Bluesky, it is expected to observe a social network with very low density, as users primarily seek out, follow, and are influenced by influencers. It is a platform where extensive connections between low-interest nodes are less common. Moreover, Bluesky is still developing, so users are more likely to follow well-known figures rather than forming highly interconnected networks.

Community structure (modularity)

After using the modularity statistic, we present the graph where:

- Node size is based on eigenvector centrality, allowing us to identify the most influential nodes within their communities.
- Node color is determined by the modularity class, enabling us to observe the different communities.

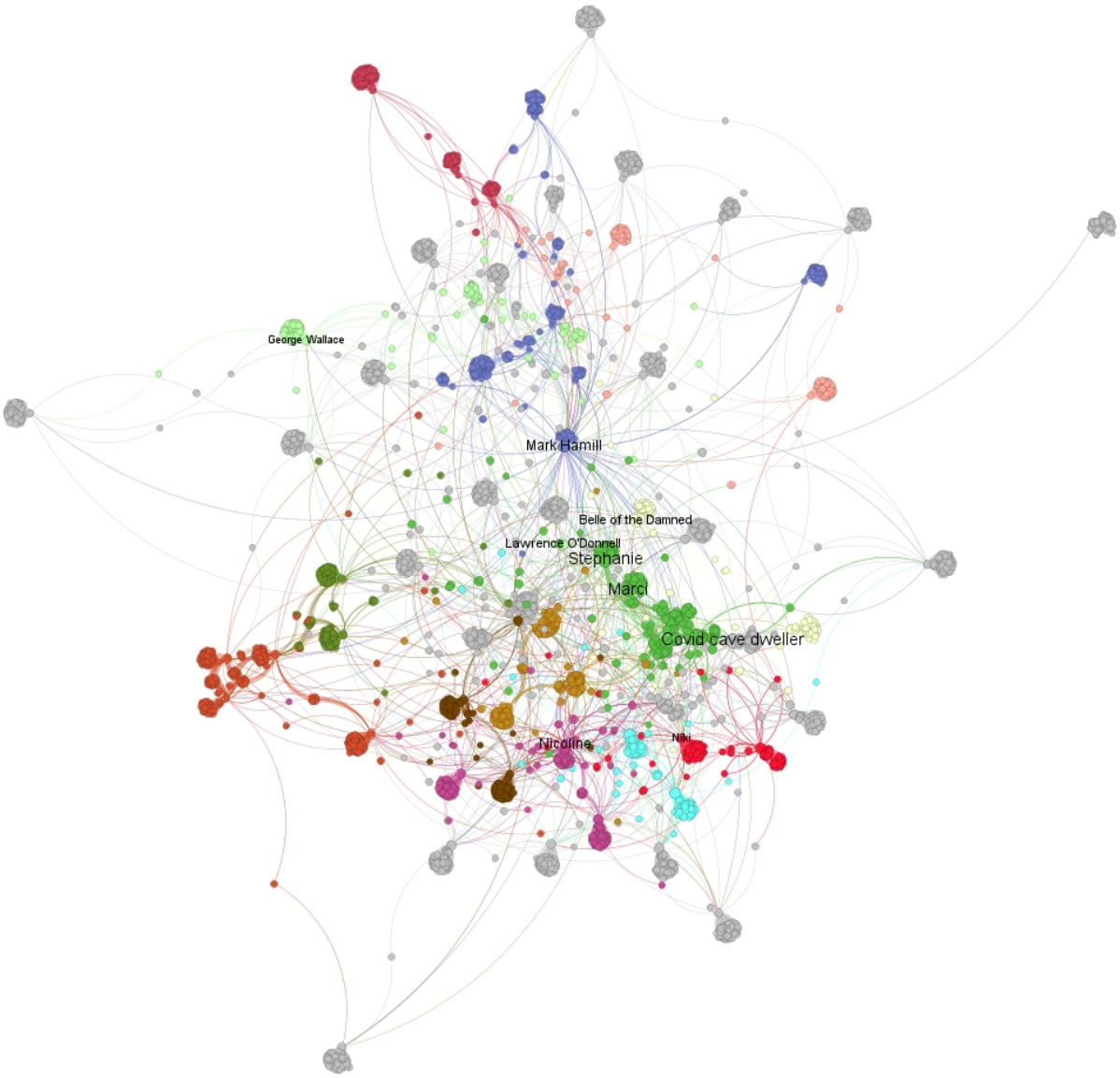
As modularity statistic uses Louvain Modularity algorithm to optimise community assignments. We run Modularity statistic for 10 times to identify the most stable and meaningful communities.

We obtain the following results:

- Modularity: 0,853
- Modularity with resolution: 0,853
- Number of Communities: 39

As expected, based on the previous conclusions, we observe a very strong community structure. The users are tightly grouped into 39 well-defined communities. High modularity suggests that people rarely interact outside their groups.

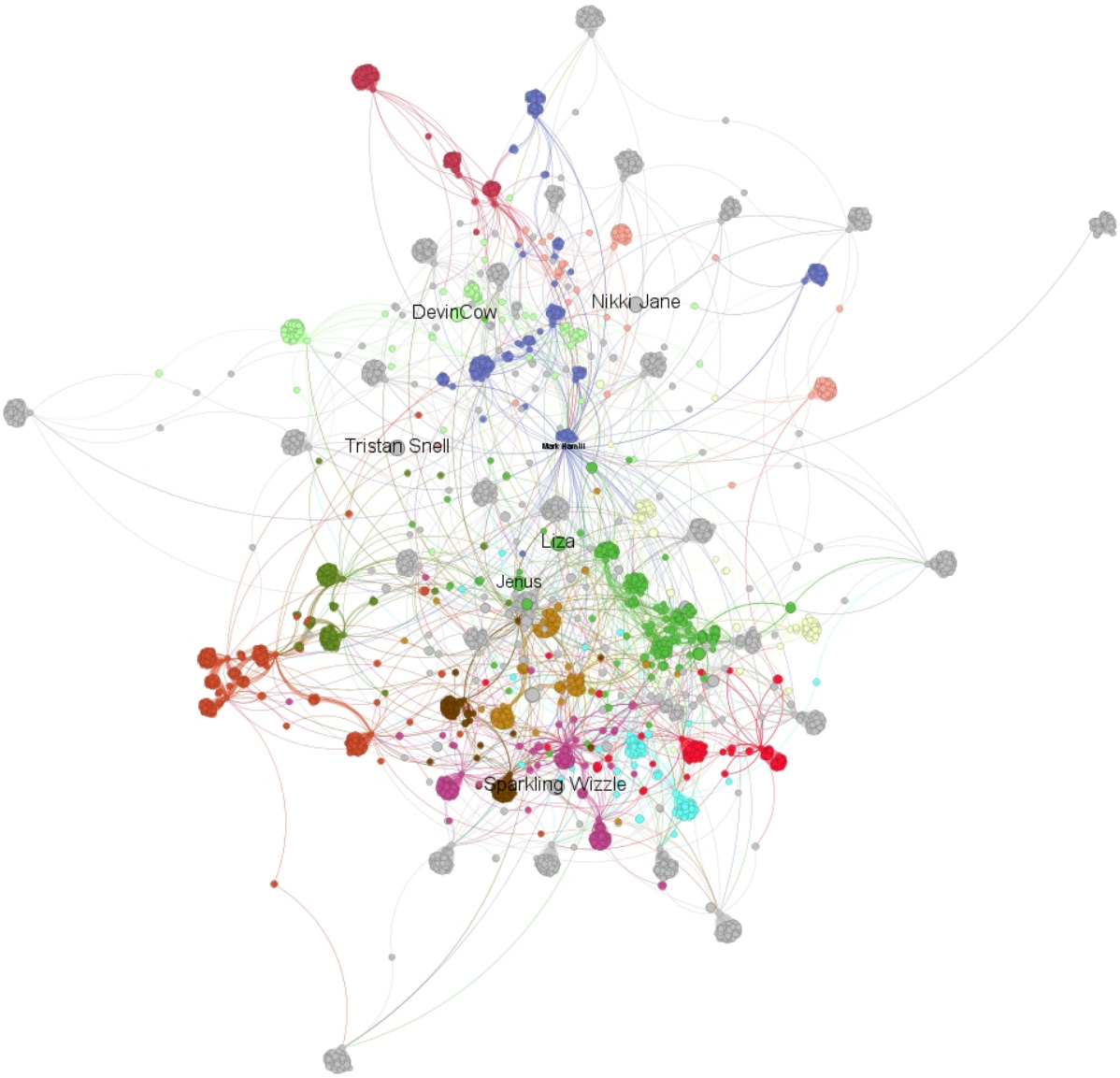
It is evident that our network is fragmented into many different communities, where each community has one or more hubs that exert strong influence. Our network does not have a single absolute central hub. Additionally, some important hubs within their respective communities are labeled in the graph.



Then, we are using the bridging centrality statistic, to present a graph where:

- Node size is based on bridging centrality, allowing us to identify the information bridging nodes between communities.
- Node color is determined by the modularity class, enabling us to observe the different communities.

We also highlight some of the nodes with the highest values, which play the most crucial role in transferring information between communities.

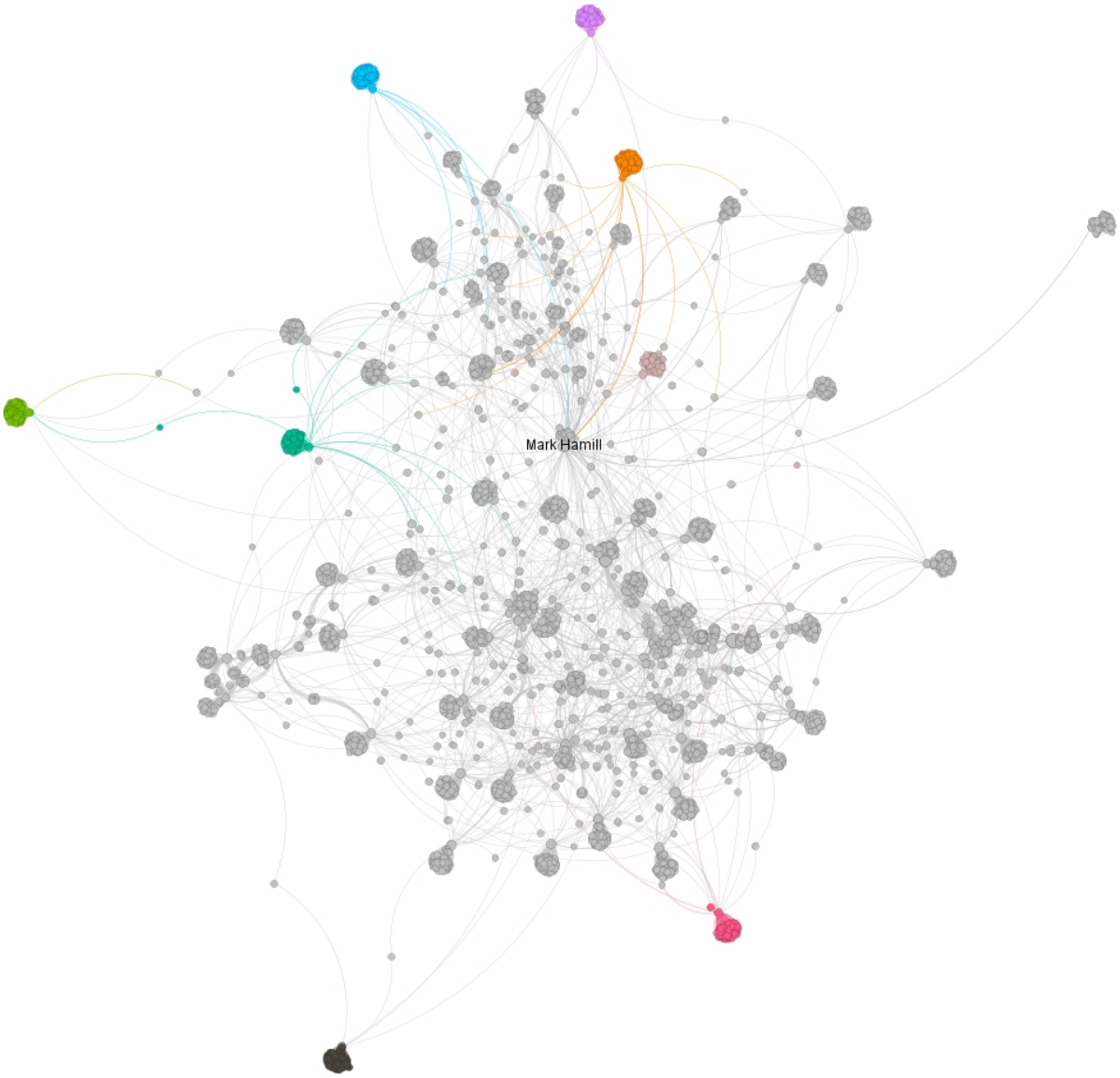


Other community detection algorithms

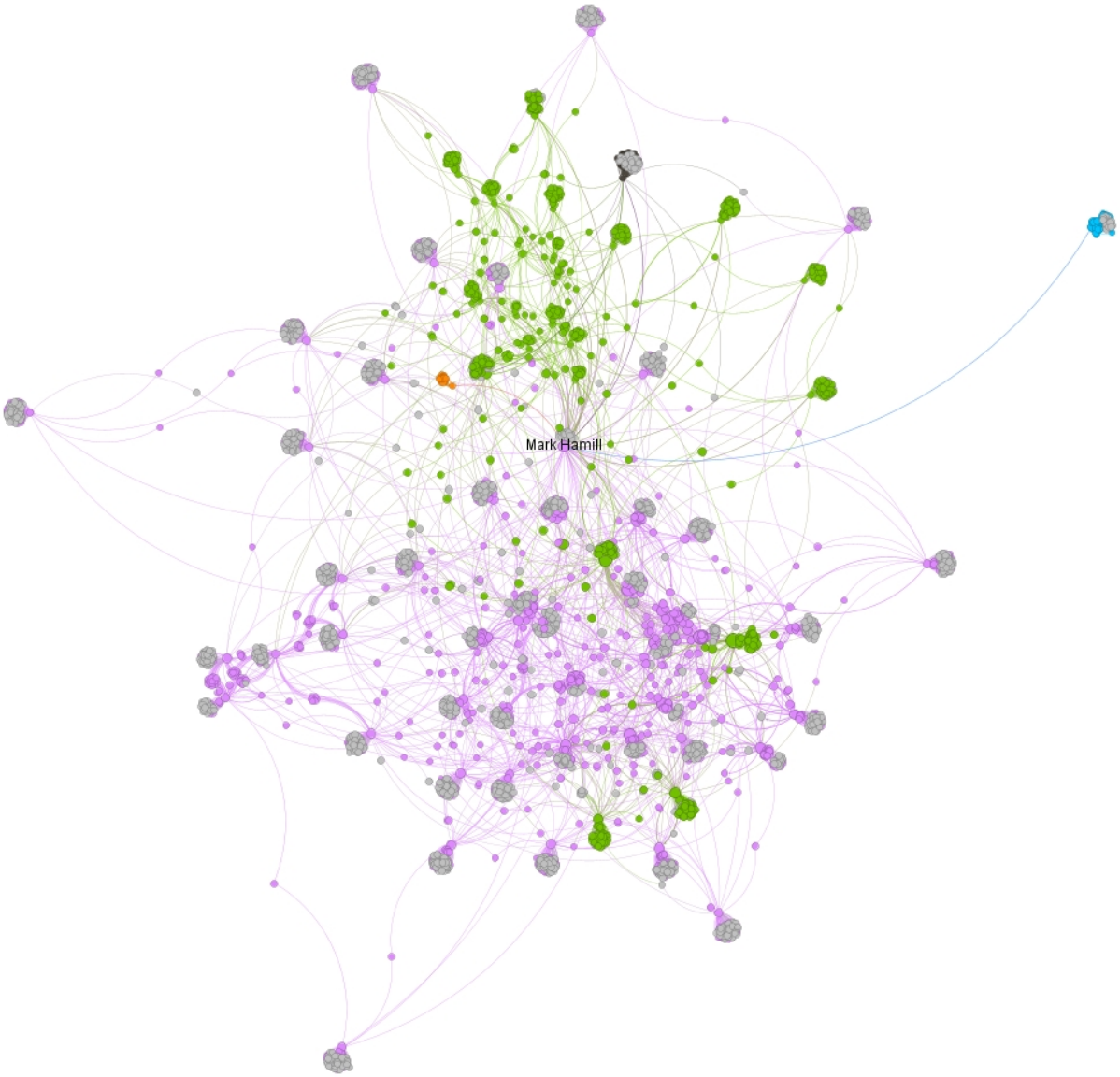
We will use two more algorithms, Leiden and Girvan-Newman Clustering.

- Node size is based on Eigenvector centrality
- Node color is determined by the modularity class

Graph for Leiden algorithm

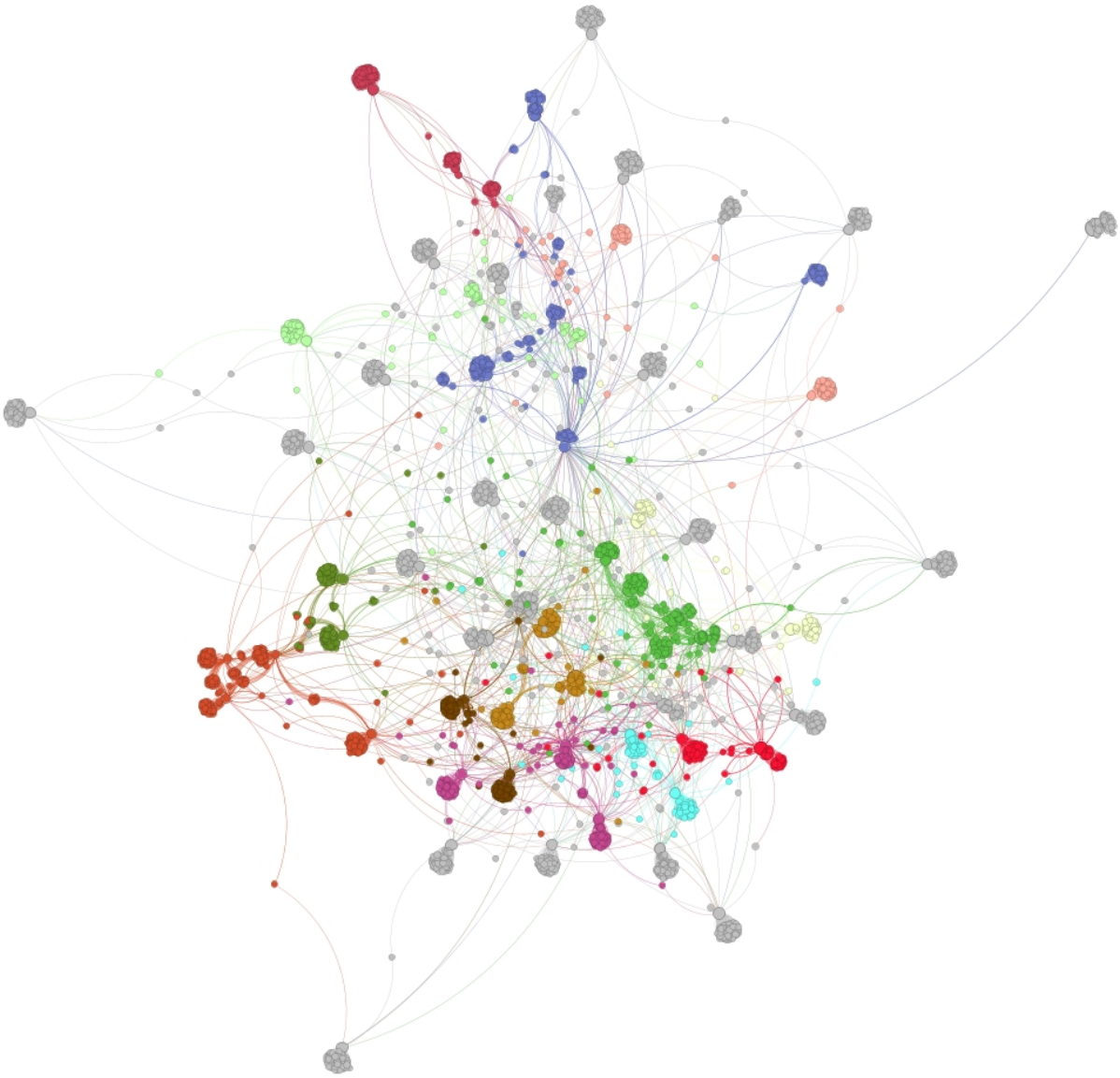


Graph for Girvan-Newman Clustering

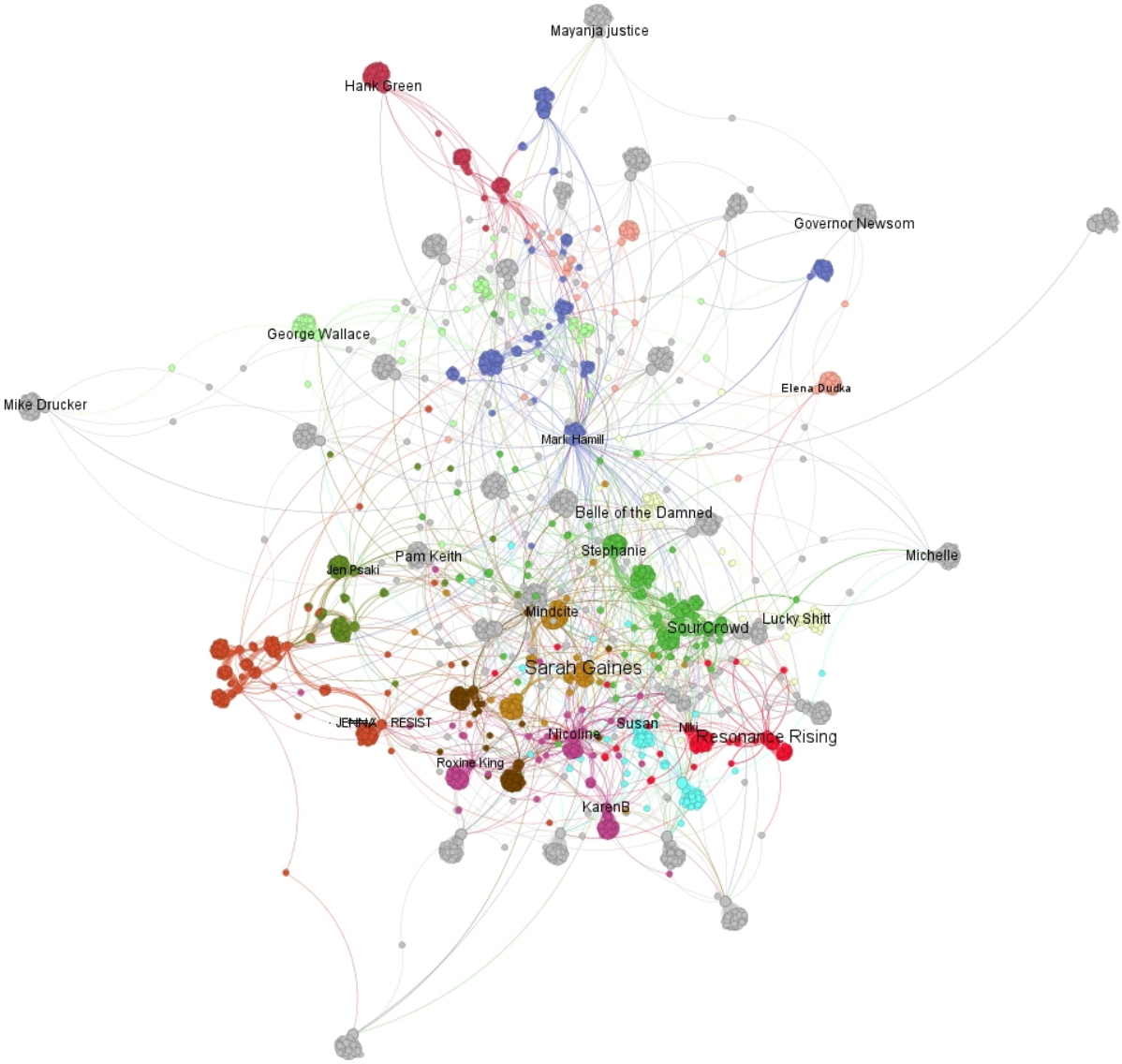


PageRank

We used PageRank metric for our social network. We present a graph where the node size depends on PageRank, and the node color is based on modularity classes.



Then we present the graph once more, with the names of the nodes with the higher pagerank values .



PageRank measures how authoritative a user is based on connections and references from others. The PageRank score is influenced by how well a user is connected to other highly important users.

- A user with high PageRank is authoritative; the information he spreads has greater credibility and he is considered a high-trust node.
- When a user has a lower PageRank, it means that they are either not well connected to other highly important nodes or that the majority of their followers have low value in terms of influence.
- A user with low PageRank, is not important.

Mark Hamill, while exerting massive influence in the network and being highly connected to important users, has a majority of followers who are regular users with low PageRank. This factor shapes his overall PageRank to be at moderate to high levels rather than extremely high.

In our analysis, we have identified users with much lower Eigenvector Centrality and Betweenness Centrality, yet they are considered more trusted and authoritative within our network.

Conclusion of analysis

Our social network is sparse, with low density (graph density ≈ 0) and an almost zero clustering coefficient, indicating that users do not form densely connected groups. The Giant Component is small, and only 50 nodes have a high degree, while the average degree is 1.2.

From a homophily perspective, the network shows minimal homophily, as it is generally well-balanced.

Mark Hamill serves as a major influencer and hub of the network, as he has the 5th highest Eigenvector and the highest Betweenness Centrality. This means that he is not only connected to other influential nodes, but he also has a strategic role in the network's flow of information, as he controls how information moves between different parts of the network. Many users rely on him to reach other users.

However, there are other nodes with a more critical role in facilitating information dissemination between different communities within our social network. (higher Bridging Centrality)

His PageRank value is moderate to high. This is very important for a profile with such a large number of followers. However, there are nodes that serve as higher-ranked sources and act as more trusted influencers in our network, with very high local influence. Mark is like the linking chain between the crowd and the experts, regarding mostly political information , science and art .

His Closeness and Eccentricity values are moderate, which indicates that some nodes have faster access to the entire network.

The high modularity (modularity = 0.85) and the presence of 39 distinct communities confirm that the network is fragmented, with relatively independent groups that have limited interconnections. In each community, there is one or more hubs with very high local influence.