Octavio Villalaz

Social Network Analysis

This report describes analysis of a benchmark of the social network dataset using Gephi. As our test case, we use a famous social network dataset called Karate Club graph. The analysis includes two separate layouts (spring and circular), many centrality and network metrics calculations (degree centrality, clustering coefficients, pagerank, and betweenness centrality), and community detection by assigning a modularity class to each node. Social network analysis (SNA) and link mining provide significant insights into the composition and dynamics of the complex network. By analyzing social networks, we can identify influential nodes, detect community structures, and understand the underlying relationship between institutions. In this assignment, we detect these concepts using a practical approach with Gephi.

The primary objectives are:

- Visualization: To represent the network using various layout techniques that highlight the structural properties.
- Metric Computation: To compute key metrics such as degree centrality, clustering coefficient, PageRank, and betweenness centrality.
- Community Identification: To divide the network into meaningful communities using the Louvain algorithm with different parameters.
- Documentation: To report on methods, algorithms and conclusions, including pseudocode for the pagerank algorithm.

The dataset used is the karate club graph, which is a classic benchmark dataset in social network research. The dataset captures social interactions between members of a karate club and is widely used to test community identification and network analysis techniques.

Two layout methods were employed:

- **Spring Layout (Fruchterman Reingold)**: A force-directed layout that positions nodes in such a way that all the edges are more or less equal length, while minimizing edge crossings. This layout helps to reveal clusters and hubs.
- **Circular Layout**: Places nodes in a circle, which can sometimes make the structure more apparent and highlights symmetries in the network.

Four centrality metrics were calculated for each node:

- **Degree centrality**: measures the number of direct connections relative to the total number of nodes.
- **Clustering Coefficient**: indicates how close it is to make a full graph to the neighbors of the node.

- **PageRank**: The importance of a node depends on the principle that high-importance contributes more to the connection scores from the nodes.
- **Betweenness Centrality**: Determines the number of times a node acts as a bridge with the smallest path between two other nodes.

These matrix provide a multidimensional view of the network structure, which help identify impressive nodes and community groups.

Community detection were implemented using the Louvain algorithm and providing 2 different parameters for the resolution:

- **0.5 Resolution**: Several nodes that share many connections are grouped into a smaller number of broader clusters.
- **1.0 Resolution**: the algorithm “tightens” the criteria for staying in the same community. As a result, you observe more, but smaller, clusters—some nodes that were previously in the same group now appear in different ones

PageRank Pseudo Code:

Initialization:

- For each node, initialize PageRank score(1/N) where N is the total number of nodes.

Iteration:

- For each node, update the PageRank score where d is the damping factor (typically set to 0.85).

Convergence:

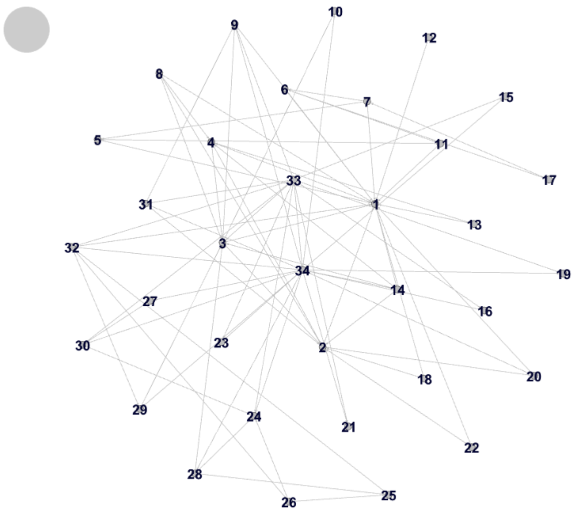
- Repeat until the change in PageRank scores between iterations is below a specified threshold.

Output:

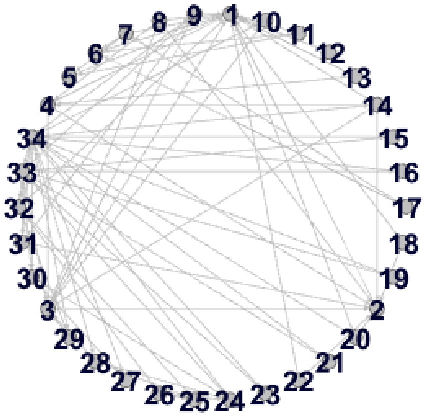
- Return the converged PageRank scores.

Spring layout revealed the overall structure and clustering of the karate club network. Nodes that are more central or connected to cluster together, reflecting the impact of the social hub in the network. The force-guided nature of the layout effectively reduces the least edge, allowing a clear identification of community groups. In contrast, the circular layout provided a more similar and aesthetic balanced view of the network. Although it did not clearly expose the cluster as a spring layout, the circular system made it easy to see the overall connectivity and symmetry of the network.

Sprint Layout



Cicular Layout



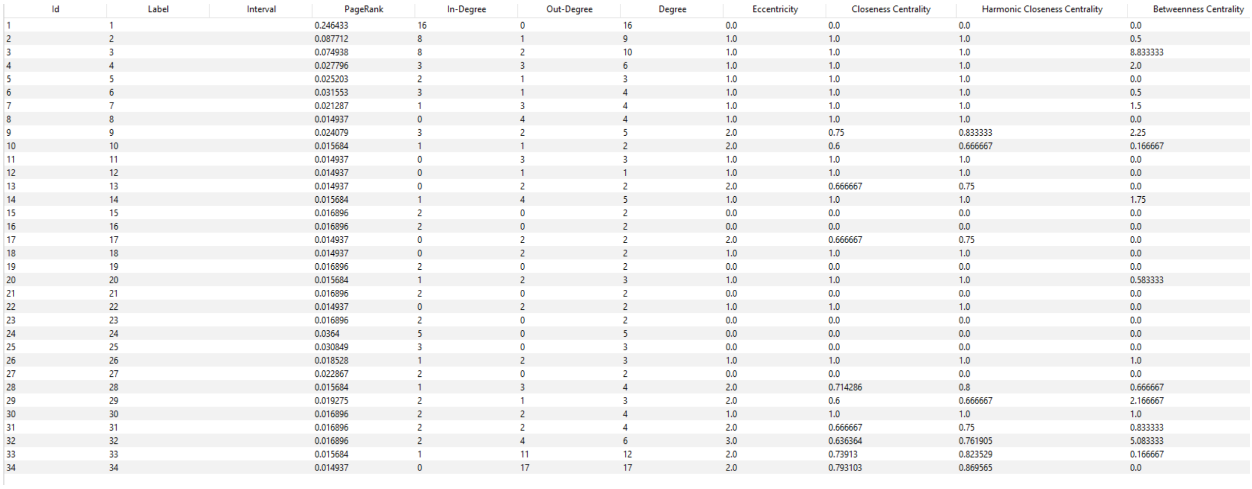
Metric Computation

Degree centrality: Nodes with high degree centrality were identified as prominent players in the network, with more direct relations. This is important in understanding the immediate effect and access of metric nodes.

Clustering Coefficient: The clustering coefficients vary between the nodes, forming a tight-sore group with the neighbors of some nodes. This provides insight into the local density of metric connection, suggests that nodes with high clustering coefficients are part of strong local communities.

PageRank: The results of the pagerank underlined the importance of nodes not only by their connections but also from the quality of those connections. Other high-ranking nodes got a boost in their score. This metric is especially useful in ranking nodes according to the overall effect within the network.

Betweenness Centrality: Nodes with higher centrality served as bridges between various groups, playing an important role in the flow of information. These nodes can be necessary for network connectivity and affect the spread of information or trends in communities.

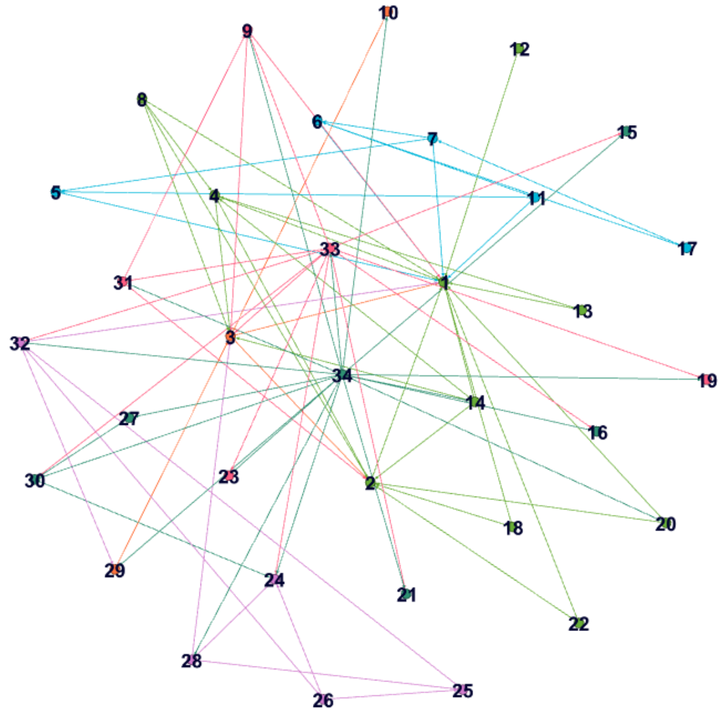


Id	Label	Interval	PageRank	In-Degree	Out-Degree	Degree	Eccentricity	Closeness Centrality	Harmonic Closeness Centrality	Betweenness Centrality
1	1		0.246433	16	0	16	0.0	0.0	0.0	0.0
2	2		0.087712	8	1	9	1.0	1.0	1.0	0.5
3	3		0.074938	8	2	10	1.0	1.0	1.0	8.833333
4	4		0.027796	3	3	6	1.0	1.0	1.0	2.0
5	5		0.025203	2	1	3	1.0	1.0	1.0	0.0
6	6		0.031553	3	1	4	1.0	1.0	1.0	0.5
7	7		0.021287	1	3	4	1.0	1.0	1.0	1.5
8	8		0.014937	0	4	4	1.0	1.0	1.0	0.0
9	9		0.024079	3	2	5	2.0	0.75	0.833333	2.25
10	10		0.015684	1	1	2	2.0	0.6	0.666667	0.166667
11	11		0.014937	0	3	3	1.0	1.0	1.0	0.0
12	12		0.014937	0	1	1	1.0	1.0	1.0	0.0
13	13		0.014937	0	2	2	2.0	0.666667	0.75	0.0
14	14		0.015684	1	4	5	1.0	1.0	1.0	1.75
15	15		0.016896	2	0	2	0.0	0.0	0.0	0.0
16	16		0.016896	2	0	2	0.0	0.0	0.0	0.0
17	17		0.014937	0	2	2	2.0	0.666667	0.75	0.0
18	18		0.014937	0	2	2	1.0	1.0	1.0	0.0
19	19		0.016896	2	0	2	0.0	0.0	0.0	0.0
20	20		0.015684	1	2	3	1.0	1.0	1.0	0.583333
21	21		0.016896	2	0	2	0.0	0.0	0.0	0.0
22	22		0.014937	0	2	2	1.0	1.0	1.0	0.0
23	23		0.016896	2	0	2	0.0	0.0	0.0	0.0
24	24		0.0364	5	0	5	0.0	0.0	0.0	0.0
25	25		0.030849	3	0	3	0.0	0.0	0.0	0.0
26	26		0.018528	1	2	3	1.0	1.0	1.0	1.0
27	27		0.022867	2	0	2	0.0	0.0	0.0	0.0
28	28		0.015684	1	3	4	2.0	0.714286	0.8	0.666667
29	29		0.019275	2	1	3	2.0	0.6	0.666667	2.166667
30	30		0.016896	2	2	4	1.0	1.0	1.0	1.0
31	31		0.016896	2	2	4	2.0	0.666667	0.75	0.833333
32	32		0.016896	2	4	6	3.0	0.636364	0.761905	5.083333
33	33		0.015684	1	11	12	2.0	0.73913	0.823529	0.166667
34	34		0.014937	0	17	17	2.0	0.793103	0.869565	0.0

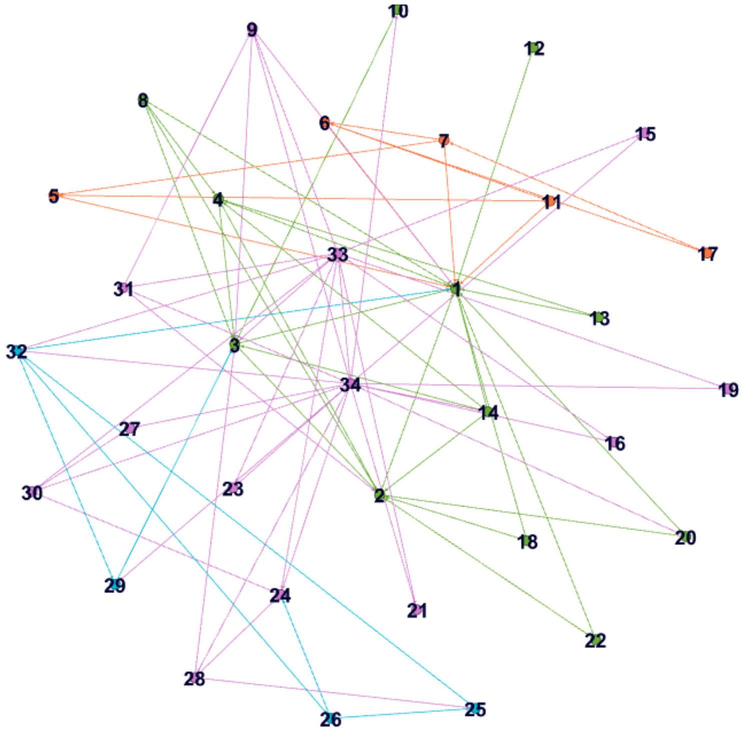
This capture show that node 34 is the standout in terms of overall influence, showing the highest degree centrality and the highest betweenness centrality, indicating that it not only has the most direct connections but also acts as a critical bridge between different parts of the network. Node 1 follows closely in importance due to its relatively high degree and notable betweenness centrality, making it another significant hub for information flow. We also see that nodes 31 and 32 have higher betweenness than many others, suggesting they are key intermediaries even if their degree is more moderate. Meanwhile, nodes 26 and 27 stand out with a clustering coefficient of 1.0, meaning they each form a perfect clique (triangle) with their immediate neighbors, though their overall influence (as measured by degree or betweenness) is lower. Collectively, these observations show that while a few nodes (like 34, 1, 32, and 33) are crucial bridges and hubs, others can be highly clustered locally (like 26 and 27) but less impactful on global connectivity.

Octavio Villalaz

0.5 Resolution



1.0 Resolution



When you run a modularity (Louvain) algorithm in Gephi with various resolution parameters, you are effectively zoom in or out of the community structure of the network. A low resolution value (eg, 0.5) usually makes the algorithm more inclined to merge the nodes into large, coarse communities, while a high resolution (eg, 1.0) can divide the network into smaller communities.

This behavior shows the so-called "resolution range" of modularity-based methods, depending on the resolution parameter, the same network can be divided into more or more communities. The best option of resolution often if you want a comprehensive observation of the network structure, a low resolution may be sufficient; If you want to highlight small subgroups or more fine differences, a high resolution can be more informative.

The report shows that effective social network analysis depends on visualization, metric computation and combination of community identification techniques. Various layouts, such as spring and circular, affect how network structures are considered, clusters and circular layouts with spring layouts are highlighted which offer balanced overall ideas. The local properties measured by the degree centrality and clustering coefficients, with global insights from the centrality, identify important nodes for network harmony and information flow. Additionally, Louvain community detection algorithm suggest that the boundaries of the community may vary depending on the chosen parameters. Finally, the provided pagerank pseudocode and modular python codes underline these analytical methods recurrence and scalable nature, allowing them to suit more complex datasets.