Social Network Analysis of Spanish High School Friendship Structures: A Topological Study

Aaron Salazar, LM in Artificial Intelligence, 0001122194
Davide Crociati, LM in Artificial Intelligence, 0001125001
Ivan Barona, LM in Artificial Intelligence, 0001109632

1. Introduction

This project explores the structure and dynamics of social relationships among students in 13 Spanish high schools using tools and techniques from social network analysis. By modelling friendships as directed, weighted networks derived from survey data, we aim to uncover patterns of cohesion, influence, and group formation within each school. Through both node-level and network-level analysis, we seek to understand how students connect, how social roles emerge, and how structural features of these networks shape the overall social environment. This approach provides a data-driven lens to examine peer interactions, offering insights that may inform educational strategies, social support initiatives, and further research into adolescent social systems.

2. Problem and Motivation

Understanding how students form and maintain social connections within schools is essential for identifying the underlying mechanisms of peer influence, group dynamics, and social integration. Adolescence is a critical period for social development, where friendships play a central role in shaping identity, behavior, and academic outcomes. Yet, the structure of these relationships is often complex and hidden from view. Traditional methods such as interviews or surveys offer limited insight into the broader patterns of connectivity and influence within a school setting.

This project seeks to address that gap by applying network science to the study of student relationships in 13 Spanish high schools. By modeling the social environment as a network of directed, weighted ties, we aim to move beyond anecdotal understanding and uncover systematic patterns of interaction, centrality, clustering, and fragmentation. Our motivation lies in the potential for these insights to inform educational policies and interventions—whether it be identifying socially isolated students, mapping influential peer groups, or understanding the resilience and vulnerability of social systems. Ultimately, this work contributes to a deeper, data-driven understanding of how adolescent social networks function and evolve within the school context.

3. Datasets

The Spanish High Schools (2023) dataset comprises 13 directed networks of friendships among over 3,000 students across 13 Spanish high schools, collected via social surveys as described in Ruiz-García et al. (2023). The data captures self-reported social relationships, with edge weights (-2, -1, +1, +2) indicating the nature and intensity of friendships and node attributes including age level, group, sex, and psychological traits (CRT and prosociality). The data was not digitized by the study team but obtained in its original digital form from the source, available publicly under a CC BY 4.0 license at https://doi.org/10.5281/zenodo.7647000. Data handling and storage were managed using Python, with NetworkX for network manipulation and Jupyter Notebook for workflow transparency. Network measures, such as triadic influence, degree centrality, and clustering coefficients, were computed using NetworkX, while visualizations were generated with Matplotlib.

4. Validity and Reliability

4.1. Validity

The Spanish High Schools dataset validly represents adolescent social dynamics, with edge weights reflecting relationship intensity and nature, aligning with social science concepts like homophily and triadic closure. Node attributes (age, sex, CRT, prosociality) enhance the model's ability to capture social compatibility, and triadic influence metrics mirror real-world network structures. However, reliance on self-reported survey data risks omission errors (missing external ties) and retrospective errors (inaccurate recall of interactions). Cross-checking responses during data collection mitigates these issues, ensuring the dataset measures intended social interactions accurately.

4.2. Reliability

The study ensures reliability through reproducible methods, using the original dataset without preprocessing due to its clean structure (no missing or duplicate edges). NetworkX and Matplotlib, executed in Jupyter Notebook, facilitated consistent computation of metrics like triadic influence and centrality. Deterministic algorithms and controlled randomness in neural network predictions ensure repeatability. The static 2023 dataset supports consistent replication, though applying the methodology to other contexts may introduce variability. Objective measures minimize subjective bias, addressing threats like data aggregation errors, ensuring robust results under similar conditions.

5. Measures and Results

5.1. Node-Level Measures

5.1.1. Degree Centrality

Degree centrality measures how connected and visible students are within their networks, based only on positive relationships (+1 or +2). It distinguishes outgoing (nominations made) from incoming (nominations received) ties, with values normalized for school size. School t11_7 showed the highest mean degree centrality (0.183), followed by t11_4 (0.153), indicating stronger social interaction, while t11_10 had the lowest (0.046), reflecting less engagement. In t11_6, the top student had an in-degree of 0.43, nominated by 43% of peers, but most had much lower scores.

Out-degree centrality also skewed, with a few highly sociable students nominating many friends, while most nominated few. This uneven distribution suggests socially dominant hubs exist within generally modestly connected networks. Overall, degree centrality reveals that while many students maintain few positive ties, a small group of well-connected individuals significantly influence the social structure.

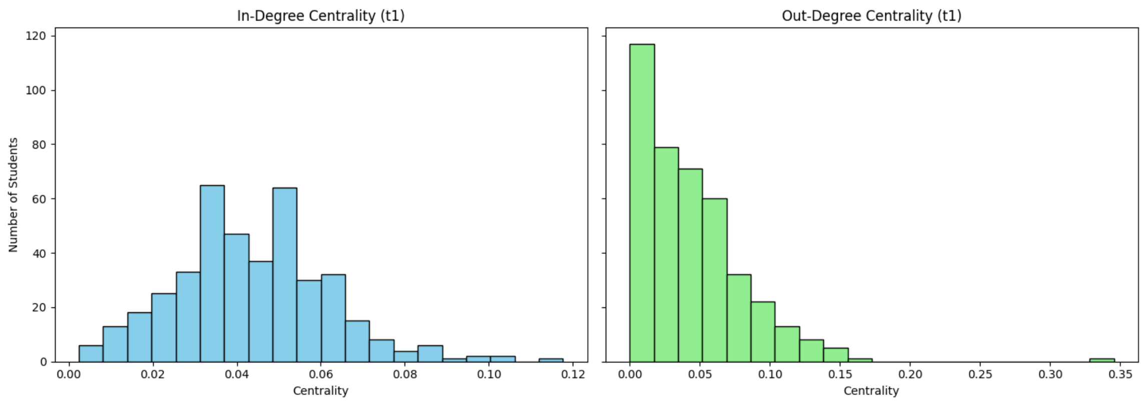

Figure 1 In and Out Degree Centrality for the Network t1

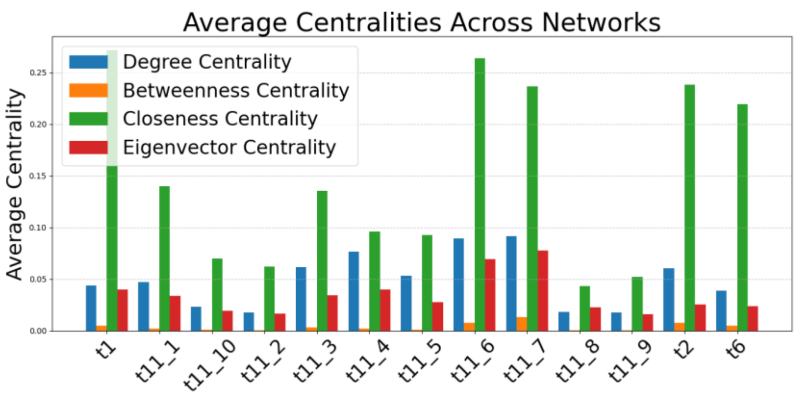

Figure 2 Average Centralities for each network

5.1.2. Betweenness Centrality

Betweenness centrality identifies students who act as social bridges—those connecting otherwise separate peer groups. School 11_7 had the highest mean betweenness (0.0116), with one student scoring 0.106, showing strong influence as a connector. In contrast, schools 11_2, 11_9, 11_8 showed low average betweenness suggesting more cohesive or compartmentalized structures with fewer intermediaries.

Notably, students with the highest betweenness weren't always the most connected, emphasizing that strategic positioning, not just sociability, shapes influence. These findings underline the subtle yet significant roles some students play in maintaining the flow of information and connection within school networks.

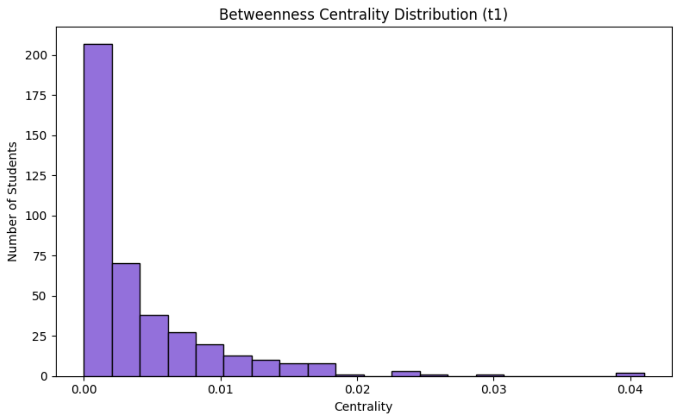

Figure 3 Betweenness Centrality Distribution for t1

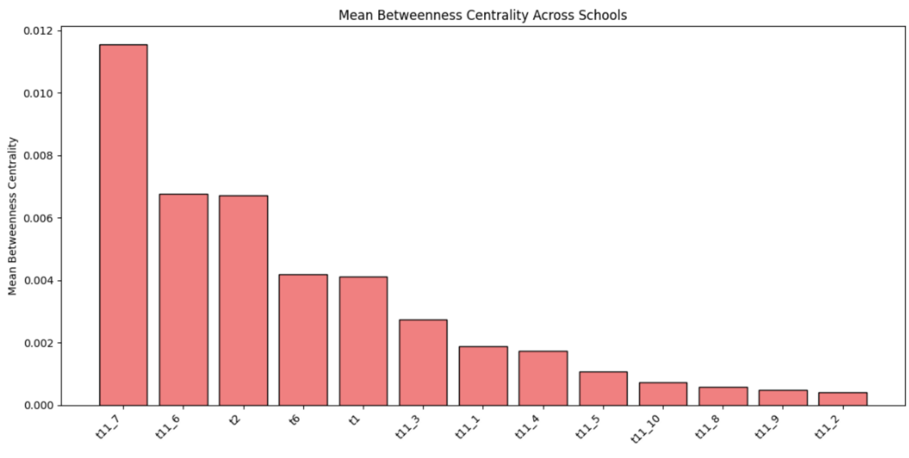

Figure 4 Average Betweenness Centrality per High School

5.1.3. Closeness Centrality

Closeness centrality reflects how efficiently a student can reach others in their school network, indicating their potential to access or spread information. Schools 1 (0.305), 2 (0.283), and 11_7 (0.282) had the highest mean closeness values, suggesting well-integrated networks with short paths between students. In contrast, school 11_4 had a standout student with a closeness score of 0.584, yet the school's average remained low (0.123), implying that most students are more peripheral.

Unlike popularity, closeness reflects structural efficiency—students who may not have the most ties but are strategically located to bridge across the group.

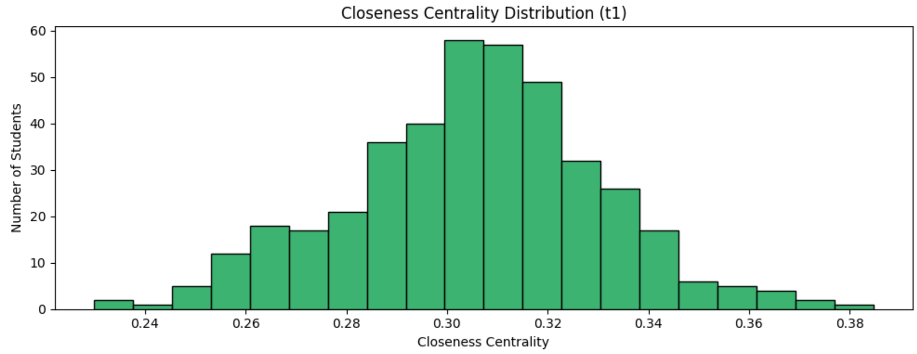

Figure 5 Closeness Centrality Distribution for t1

5.1.4. Eigenvector Centrality

Eigenvector centrality highlights students who are not just popular but are also connected to other influential peers. t11_5 had the highest average score (0.117), indicating a broad core of influential individuals. t11_4 showed the highest individual score (0.292), pointing to a tightly connected subgroup. Networks like t11_8 and t11_7 also had high values, suggesting cohesive clusters of central actors. In contrast, t1, though larger and sparser, still had a standout student (0.153), showing that strong influence can arise in less connected settings.

In contrast, 11_2 has the lowest averages (0.017), reflecting more fragmented structures. These results show that eigenvector centrality captures deeper patterns of social influence, identifying students who are central within influential groups—not just those with many direct connections.

5.2. K-Core Decomposition Analysis

K-core decomposition is used to identify the most connected substructures within a network. A k-core is a maximal subgraph where all nodes have at least k connections within that subgraph. Higher k-core levels represent more cohesive and resilient groups, often associated with leadership, influence, and structural stability.

5.2.1. Maximum and Average Core Numbers Across Schools

Figure 6 shows both the maximum core number and average core number for each school.

- School 11_2 and School 6 exhibit the highest maximum core numbers (26), as well as high average core numbers, indicating the presence of very dense and resilient social cores.
- School 1_9 has the lowest values, suggesting a fragmented or loosely connected social structure.
- The gap between maximum and average core values suggests a core-periphery structure.

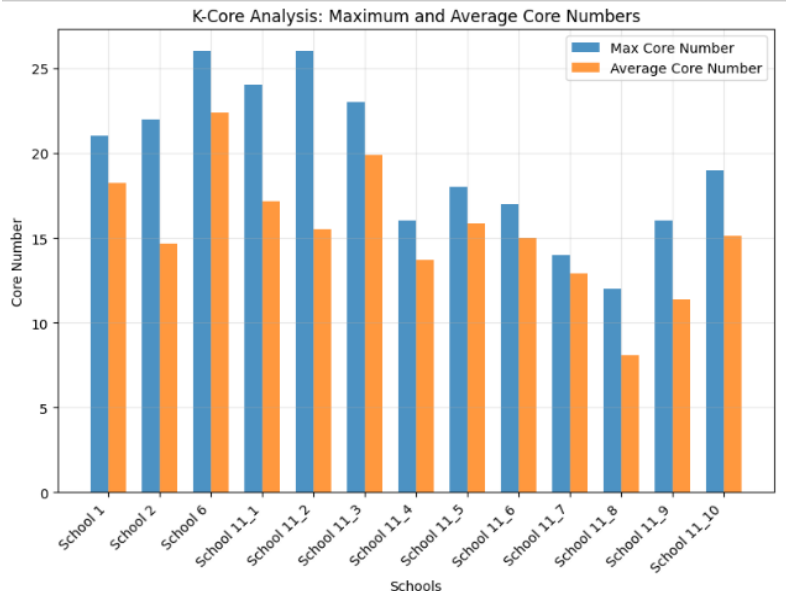

Figure 6 Maximum and Average Core Numbers Across Schools

5.2.2. Core size vs. Network

Figure 7 shows the number of nodes at each k-core level for School 6. A steep drop beyond core level 23 indicates that only a small elite group belongs to the innermost core. This hierarchy suggests the existence of key social actors embedded within a broader, moderately connected network.

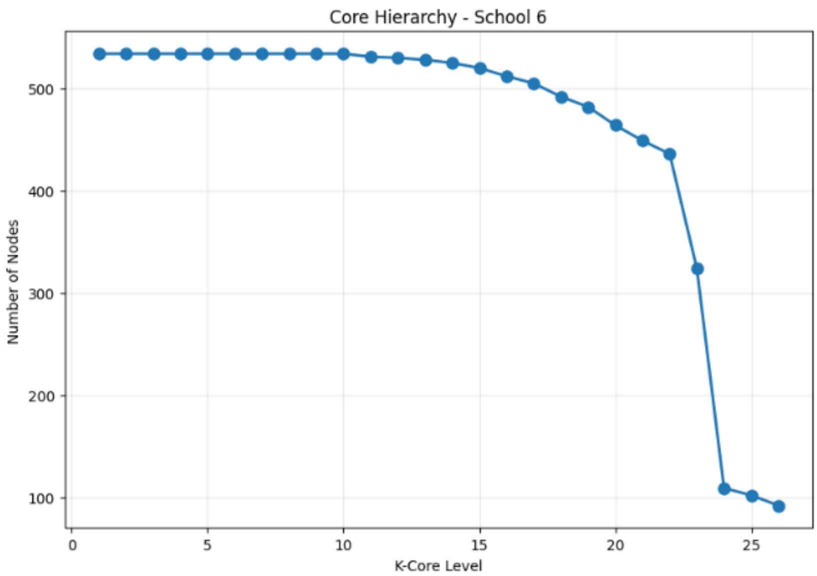

Figure 7 Core Hierarchy in School 6

5.2.3. Core Size vs. Network Size

In Figure 8, we observe a moderate positive correlation (r = 0.407) between network size and the number of nodes in the maximum core. This suggests that larger networks may support larger dense cores, but this is not a linear relationship—structural factors such as network density also play a crucial role.

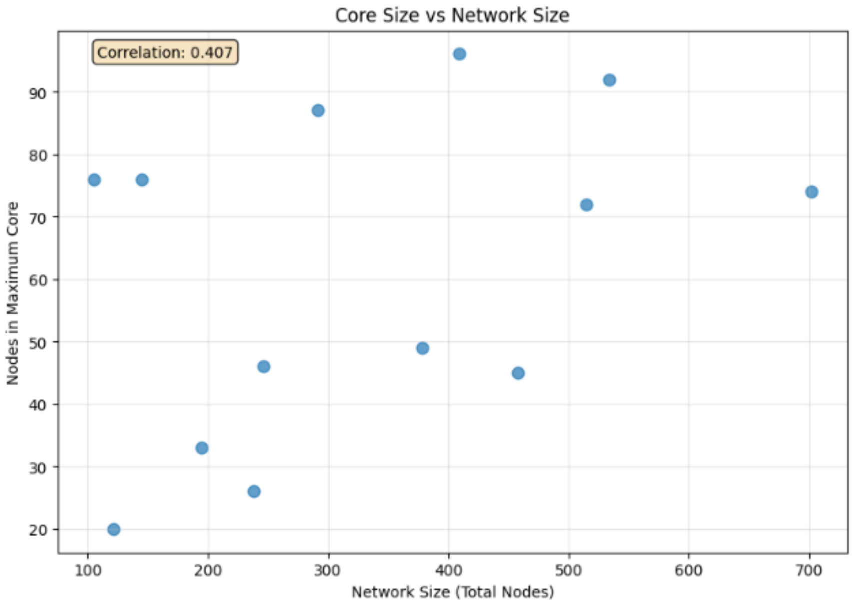

Figure 8 Nodes in Maximum Core vs Network Size

5.2.4. Core Visualization – School 6

Figure 9 visualizes the core structure of School 6. The highest-core individuals (yellow nodes) are embedded at the intersection of multiple clusters, potentially acting as social bridges. This visualization highlights the modular and hierarchical nature of the network.

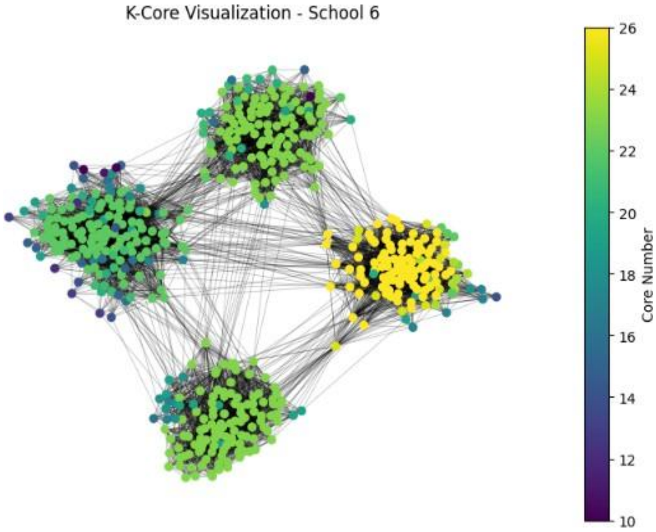

Figure 9 K-Core Visualization of School 6 (Color-coded by Core Level)

5.3. Clique Analysis

Clique analysis identifies fully connected subgroups—cliques—providing insight into the tightest social clusters within a network. These groups often represent strong ties and high trust environments.

5.3.1. Clique Size Distribution Across Schools

Figure 10 presents the distribution of clique sizes using boxplots:

- Schools like 11_3, 6, and 2 show larger median clique sizes and more outliers, indicating multiple large, tightly-knit friendship groups.
- In contrast, School 1 and School 11_1 have smaller clique sizes.
- This implies different levels of micro-level cohesion across school contexts.

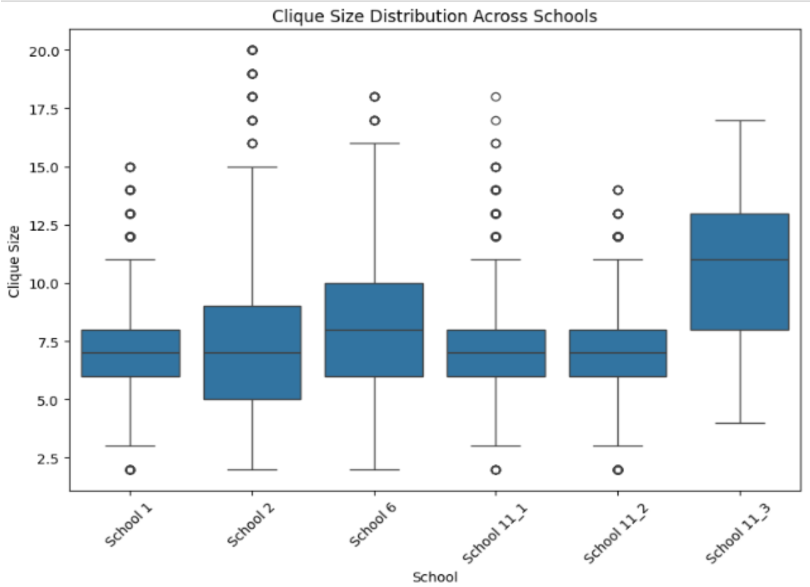

Figure 10 Clique Size Distribution Across Schools

5.3.2. Max Clique Size vs Network Characteristics

Figure 11 shows that clique size is not strictly a function of network size. Larger cliques tend to appear in denser networks, regardless of size. This indicates that structural cohesion, not population size, drives clique formation.

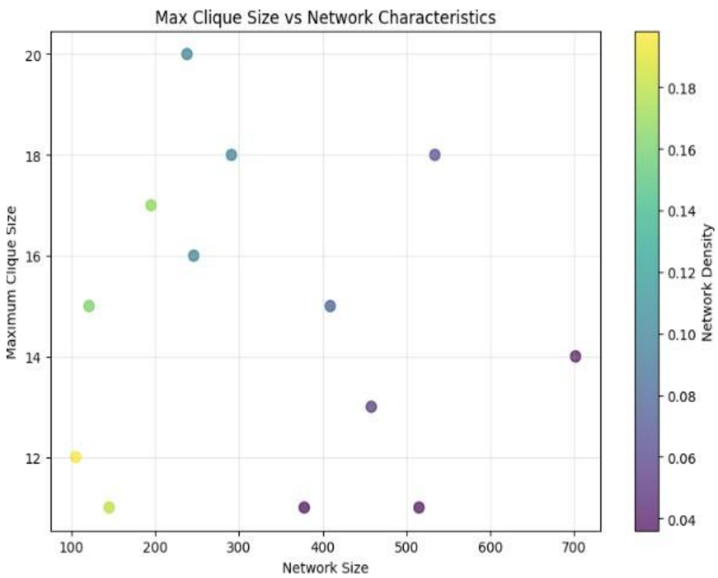

Figure 11 Max Clique Size vs Network Size and Density

5.3.3. Clique Membership Distribution – School 2

As seen in Figure 12, most students are members of few cliques, while a small subset belongs to over 100. These central actors likely hold significant social capital and play key roles in integrating various subgroups.

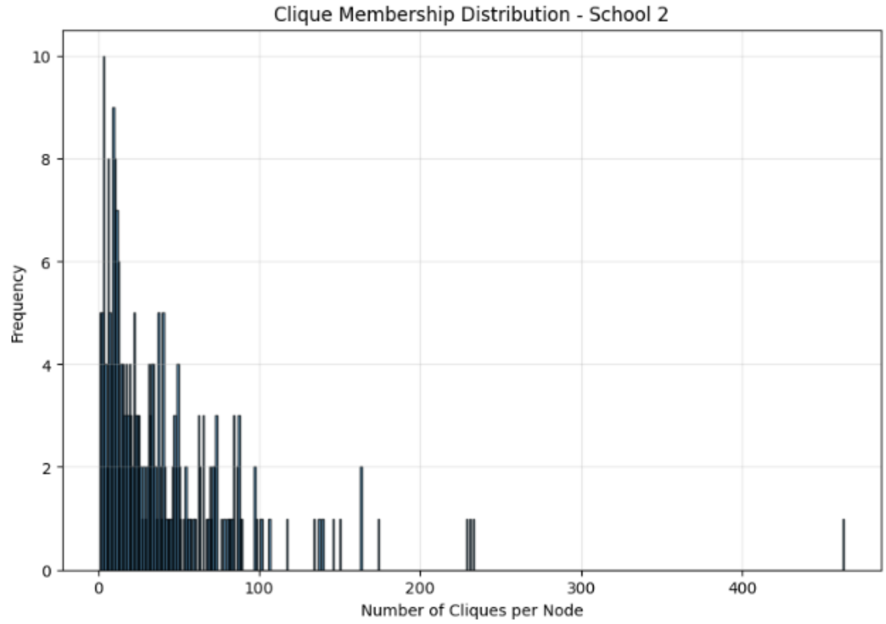

Figure 12 Clique Membership Distribution in School 2

5.3.4. Largest Clique Visualization – School 2

Figure 13 displays the largest identified clique in School 2. The complete interconnectivity among members suggests a dominant social group likely influential in shaping social norms and peer behavior.

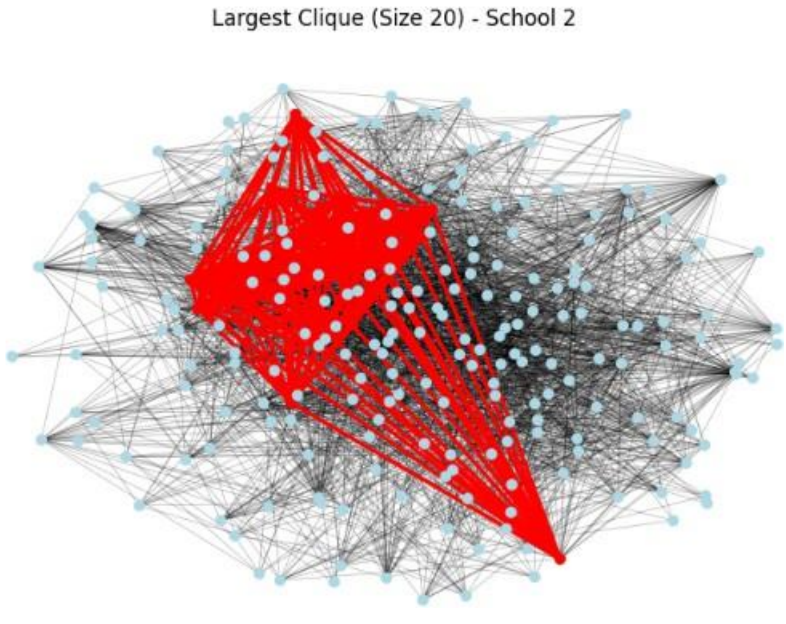

Figure 13 Largest Clique (Size 20) in School 2

5.4. Multi Category Nominal Scale Analysis

5.4.1. Student Category Distribution

The Student Category Distribution classifies each student into four groups based on their degree centrality: high-degree (top 10%), moderate-degree (10–50%), low-degree (bottom 50%), and isolated (no reported positive ties). This framework allows a comprehensive view of social connectivity patterns within and across schools. Across all networks, the majority of students fell into the low-degree and moderate-degree categories, showing that most students maintain limited social ties. As seen in Figure 14, low-degree students account for the largest portion, followed by the moderate group together including the bulk of the social structure.

Approximately 10% of students belong to the high-degree group, emerging as socially central figures likely to influence and connect peer clusters. Remarkably, only one student across all schools was identified as fully isolated, meaning nearly every student had at least one positive relationship. Figure 14 visually reinforces this distribution: a small slice of highly connected individuals shapes the network's cohesion, while most students are in more modest social roles. These patterns are consistent with dynamics in social systems and show the importance of supporting students in a periphery while recognizing the connective power of those in the network core.

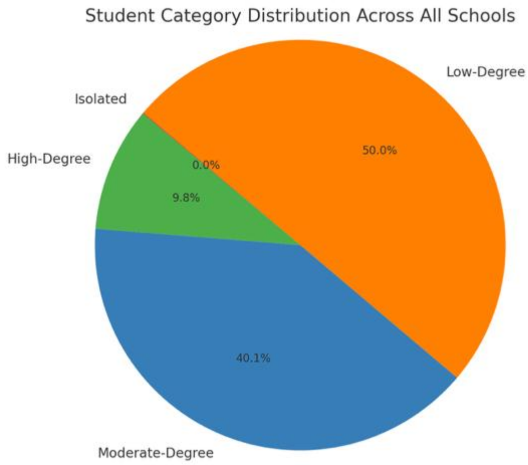

Figure 14 Student Category Distribution

5.4.2. Scalar Network Analysis

Scalar network analysis captures high-level structural properties that characterize the overall shape and organization of each school's social network. Among the schools examined, t1 and t2 emerged as the largest networks with 409 and 238 students, respectively. Both exhibited relatively high average degrees 17.85 for t1 and 14.27 for t2 indicating an active exchange of positive relationships among students. In contrast, smaller networks like t11_7 and t11_4 had lower average degrees, though their densities were higher (e.g., t11_7 at 0.091), suggesting that students in these schools tend to interact more extensively within tighter-knit communities.

One particularly insightful scalar metric is degree assortativity, which measures whether students with many connections tend to link with others who are similarly sociable. While t1 and t2 showed slightly positive assortativity values (0.06 and 0.03), suggesting mild homophily in connection patterns, schools like t11_10 and t11_7 exhibited negative assortativity (−0.042 and −0.063), implying a hub-and-spoke structure where a few highly connected individuals link to many less-connected peers.

To support and visualize these findings, we generated a degree centrality distribution plot (see Figure 15), combining data from all 13 schools. Plotted on a log-log scale, the graph reveals a skewed distribution: the majority of students hold low centrality, while a small number exhibit significantly higher centrality values. This confirms the heterogeneous nature of social visibility and influence only a few students occupy structurally dominant roles, while most participate from the periphery. Together, these scalar metrics and distributions reveal a wide spectrum of network architectures, from dense egalitarian clusters to hierarchical, hub-driven systems.

These structural patterns are consistent with findings by Brewe et al. (2012), who used network metrics like density, degree, and assortativity to analyze peer interaction in academic environments. Their study confirms that such measures offer valuable insights into the cohesiveness and engagement levels within educational settings.

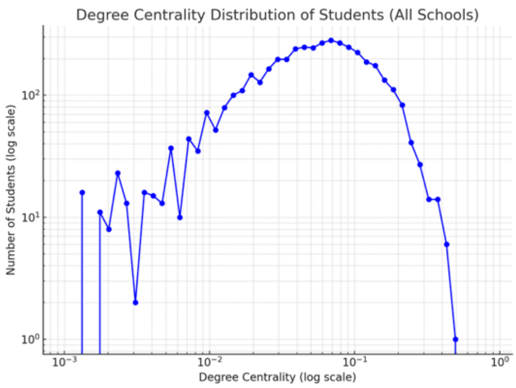

Figure 15 log-log plot of the Degree Centrality Distribution

5.4.3. Small World Network Analysis

Small world networks show a balance between local clustering and global efficiency, enabling quick access to others while maintaining close-knit groups. To evaluate small world structure in each school, we compared the actual network's clustering coefficient and average path length to those of a random graph with the same number of nodes and edges. The result is the small world coefficient (σ), where values greater than 1 indicate small world properties.

In this analysis, schools like t1 (σ = 2.87) and t2 (σ = 2.29) demonstrated strong small world structure, with high clustering and efficient paths, suggesting an ideal balance for social cohesion and communication. Networks like t11_10 (σ = 1.46) and t11_7 (σ = 1.34) showed moderate small world characteristics, implying that while they maintain reasonable clustering and reachability, the structure is not as tightly integrated as in t1 and t2.

On the other end, t11_4 stood out for its very high clustering coefficient (0.66), but it produced a low small world coefficient (σ = 0.75). This indicates that although students in t11_4 tend to form tightly interconnected peer groups, the network lacks the global efficiency seen in a typical small world structure. In essence, the network is cohesive but relationships are dense locally, and access across broader peer groups may be limited. This contrast illustrates how a network can be richly clustered without achieving the balance necessary for small world efficiency.

5.5.Network Diameter Analysis

The network diameter refers to the longest among all the shortest paths between any two nodes in a network. It indicates the maximum distance required for information to travel from one point to another, thus reflecting the network's overall connectivity and efficiency. The network radius represents the minimum eccentricity across all nodes, capturing how close the most central node is to the farthest node in the network.

School	Diameter	Radius
1	4	3
11_4	2	1
6	5	3
11_8	2	1
11_6	2	1
2	4	3
11_10	2	1
11_2	2	1
11_5	19	10
11_3	2	1
11_7	2	1
11_9	2	1
11_1	2	1

Table 1 Diameter and Radius of Networks

For the school networks, we observe the following:

- The majority of schools, especially those labelled with "11_”, show very low diameters and radius (typically 2 and 1), suggesting tight social groups where students are closely connected and no one is far from others in terms of social ties.
- Schools 1, 2, and 6 exhibit slightly more dispersed networks, with diameters ranging from 4 to 5. This suggests that while students are still connected, there are more degrees of separation between some individuals.
- School 11_5 is a notable outlier, with a diameter of 19 and radius of 10, indicating a highly fragmented or loosely connected network. In this school, some students may be much more socially isolated, requiring many steps to reach others through mutual connections.

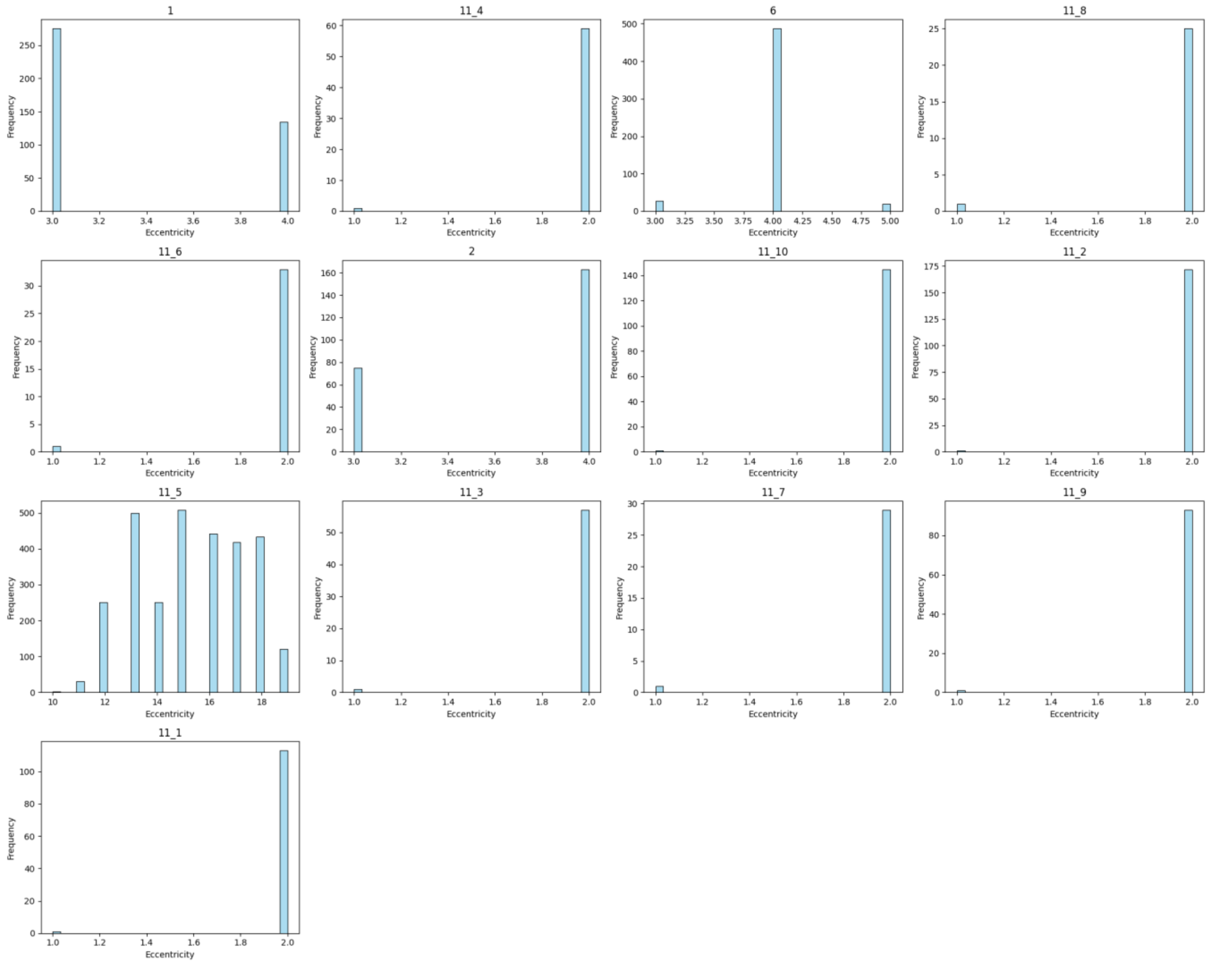

Figure 16 Eccentricity histograms for each school

As shown in the Figure, in many schools, most students have either low eccentricity (indicating they are socially central) or high eccentricity (indicating they are more socially peripheral), with relatively few students falling in between. This bimodal pattern suggests a core-periphery social structure, where a central group of well-connected students serves as the hub through which more isolated students maintain indirect connections to the rest of the network.

5.6.Community Structure Analysis

The Louvain community detection algorithm revealed substantial variation in the structure of student social networks across schools. Some schools, such as School 1 (5 communities), School 6 (4 communities), and School 2 (6 communities), showed relatively low numbers of detected communities, indicating more cohesive networks with larger, tight social groups.

In contrast, other schools displayed extreme fragmentation. For example, School 11_2 had 702 communities, School 11_10 had 458, and School 11_9 had 515, suggesting highly divided networks, where students form numerous small isolated clusters.

The Figure presents the community structure across schools, with distinct colors denoting different student groups:

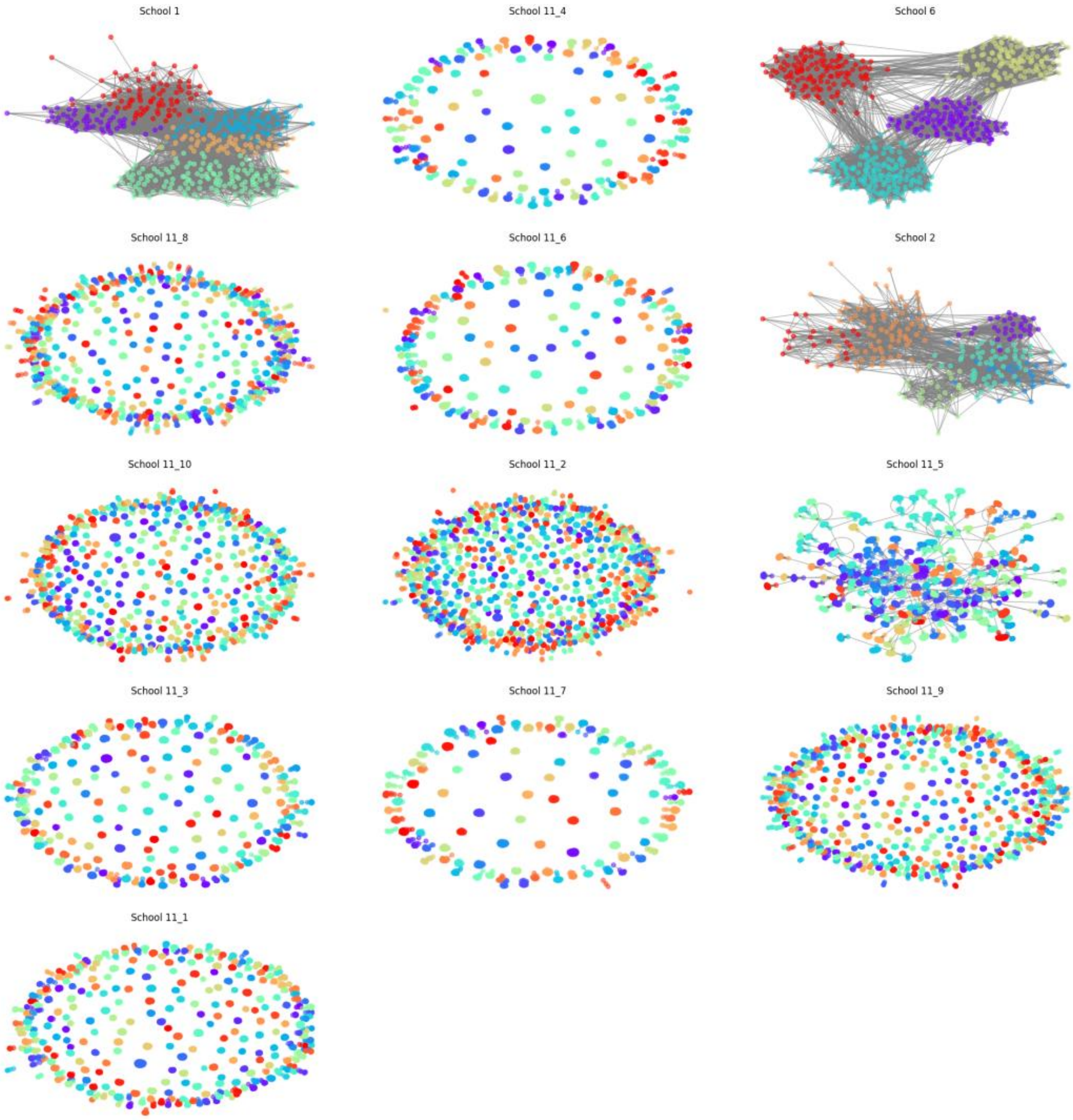

Figure 17 View of the communities detected with the Louvain algorithm

The bar plots for each school illustrate the distribution of student counts across detected communities. A consistent pattern emerges in most schools: rather than having one or a few dominant communities, the sizes decrease gradually. This lack of a single standout large community suggests that student social networks are often evenly fragmented into many similarly-sized groups.

In schools with low numbers of communities (e.g., School 1, School 2, and School 6), the communities are large and distinct.

In contrast, schools with high numbers of communities (e.g., Schools 11_2, 11_4, 11_5, 11_10, and 11_9) exhibit a smooth, descending distribution of community sizes. While some communities are slightly larger than others, there's no dramatic drop or spike, indicating an absence of a central or dominant group. Instead, we observe a long tail of smaller communities, reflecting a highly modular structure.

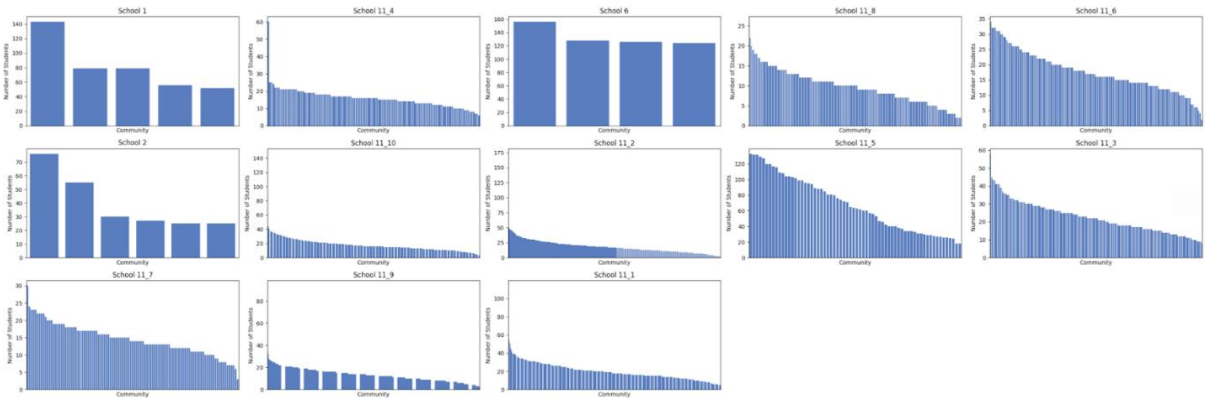

Figure 18 Distribution of the communities' sizes

5.7. Network Robustness

To assess the resilience of the networks, we conducted systematic simulations simulating network degradation under two failure scenarios: random removals and targeted removals of highly connected students (hubs).

Under random removals, most networks demonstrated good robustness. However, in some instances, removing a specific node caused a sudden and significant drop in the size of the largest connected component. These cases often corresponded to points where targeted removals had a similar impact, but much earlier in the removal process.

Schools 1, 2 and 6 have an almost linear decrease of nodes remaining in the largest component. This pattern suggests that these are relatively homogeneous, with connections distributed more evenly among students rather than concentrated around a few key individuals. As a result, the removal of any single student tends to have a proportional and predictable impact on overall network connectivity, reducing the risk of sudden fragmentation.

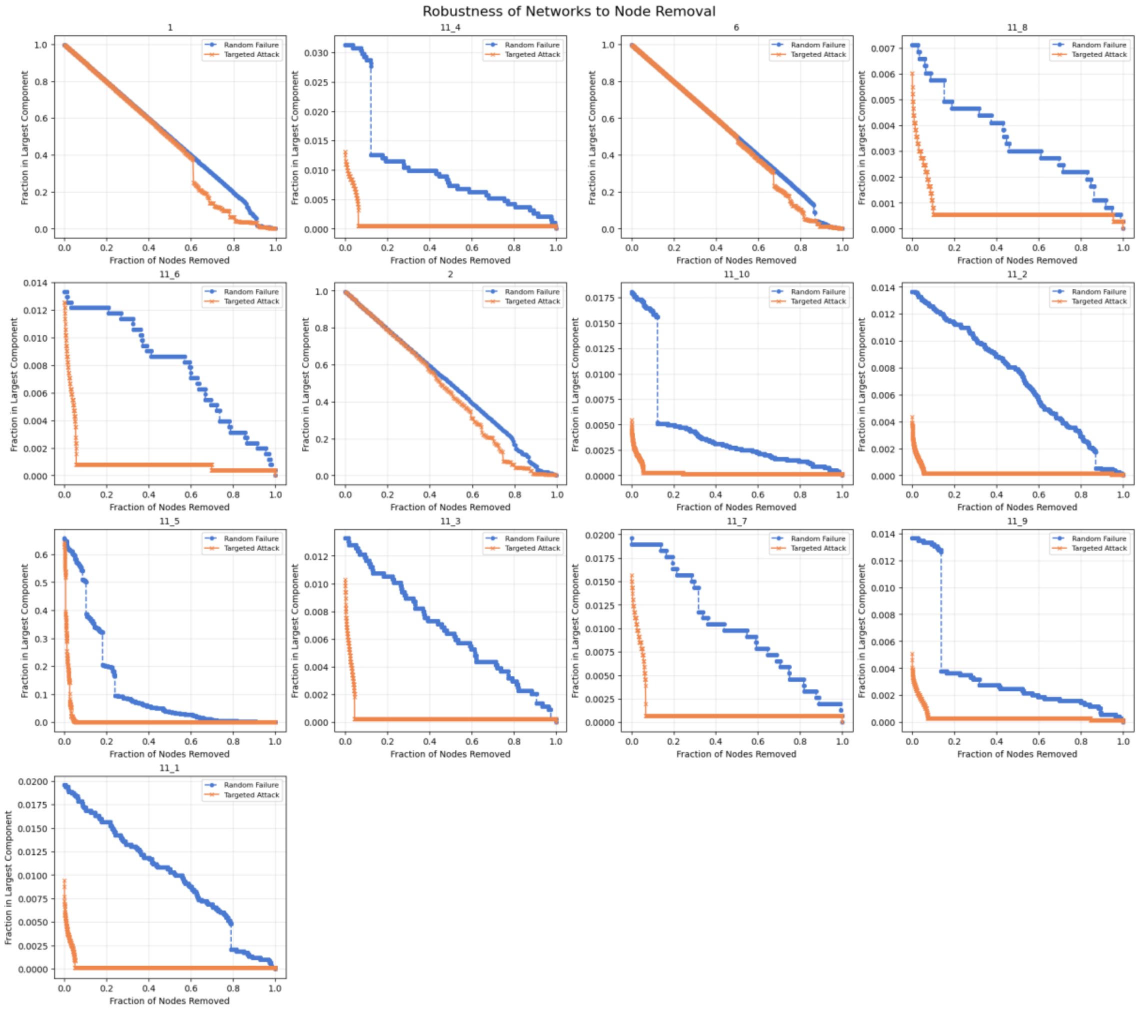

Figure 19 Network resilience with random failures (blue) versus targeted attacks (orange). On the x-axis the number of nodes removed, while on the y-axis the fraction of nodes left in the largest connected component.

5.8. Network-Level Structure and Cohesion Measures

5.8.1. Density

Density captures how many actual connections exist relative to the total number of possible connections. In directed graphs, it's calculated as:

$$Density = \frac{Number\ of\ actual\ edges}{n(n-1)}$$

This metric reflects overall connectivity and cohesion—higher density indicates more widespread interaction among students, while lower density implies fragmentation or limited engagement.

Network density acts as a proxy for social cohesion—higher values suggest more frequent interaction and peer engagement. Among the 13 schools, t11_7 and t11_4 had the highest densities (0.091 and 0.076), indicating highly interactive, close-knit student groups. In contrast, larger schools like t1 and t11_10 had much lower densities (0.044 and 0.023), suggesting more fragmented or exclusive social networks.

These patterns imply that larger student populations do not guarantee stronger connectivity. Instead, cohesion is shaped by internal interaction norms and relational dynamics, not just group size. Dense networks may reflect inclusive environments, while sparse ones may point to social divisions or logistical barriers.

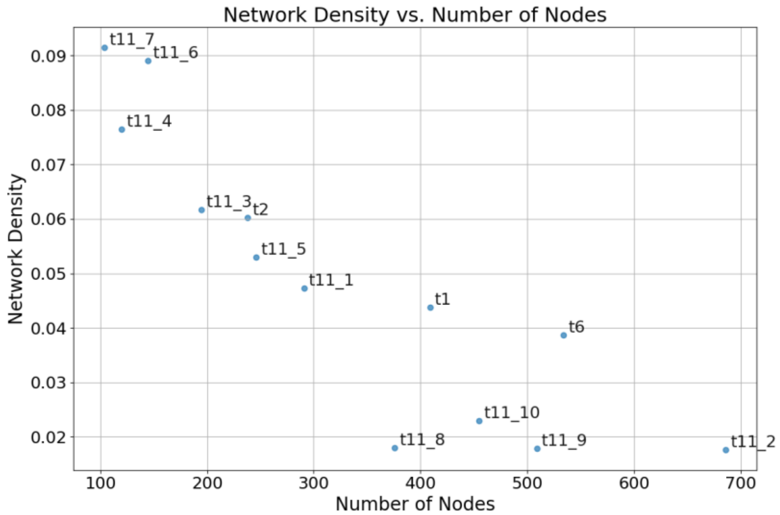

Figure 20 Network Density versus Number of Nodes

5.8.2. Average Path Length

Average path length reflects how efficiently information or influence spreads through a network by measuring the typical steps needed to connect any two students. It was calculated using each school's largest strongly connected component, ensuring all included students were mutually reachable.

School t11_4 had the shortest path length at 1.92, suggesting a tightly-knit, highly communicative group of 28 students. Larger networks like t2 and t1, with 195 and 360 students respectively, still showed low path lengths around 2.90, indicating strong cohesion despite size. Other schools such as t11_10 (2.42) and t11_7 (2.59) had slightly longer paths, suggesting more fragmentation or indirect links between subgroups.

These findings suggest that even large student networks maintain short paths for communication, supporting the idea of small-world structures in high school social networks.

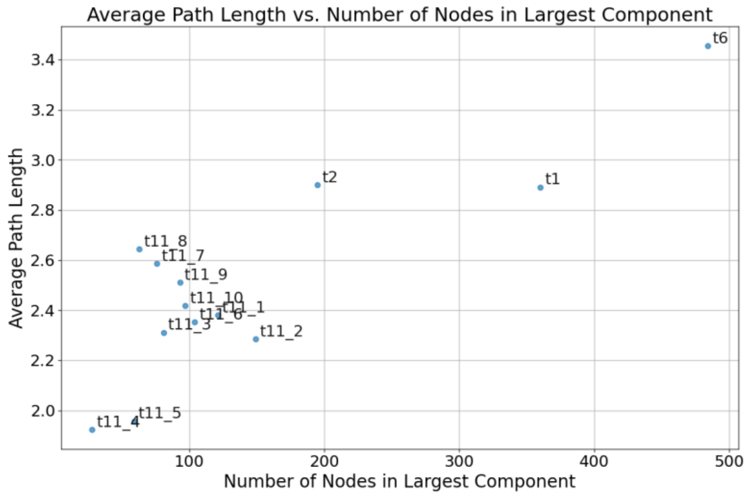

Figure 21 Average Path Length versus Number of Nodes in Largest Component

5.8.3. Modularity

Modularity measures the strength of division within a network, indicating how well it separates into distinct communities such as groups of students with dense internal connections and sparse external ties. A high modularity value suggests clear subgroup structures, while lower values indicate more blended, interconnected networks. In this analysis, t11_10 and t11_4 demonstrated the most modular structure, with values of 0.72 and 0.69, respectively as can be seen in Figure 22. Each network contained 4 communities, indicating that students in these schools tend to cluster into tight social circles with limited interaction between groups.

In contrast, schools such as t1 and t2 exhibited moderate modularity scores (~0.51) with 3 to 4 communities, reflecting a balance between subgroup formation and broader integration. Notably, t11_7 had the lowest modularity (0.40), suggesting a more cohesive and interconnected network with fewer internal divisions. This aligns with earlier metrics like high density and short path lengths, which point to an overall integrated structure. These findings highlight how modularity can uncover latent social groupings within schools offering insight into the informal social architecture that shapes student interactions.

This use of modularity to detect tightly-bound peer groups aligns with the approach of Pérez-Flores et al. (2022), who analyzed triadic influence in Spanish high schools and emphasized the importance of community structures in understanding compatibility and social cohesion within adolescent networks.

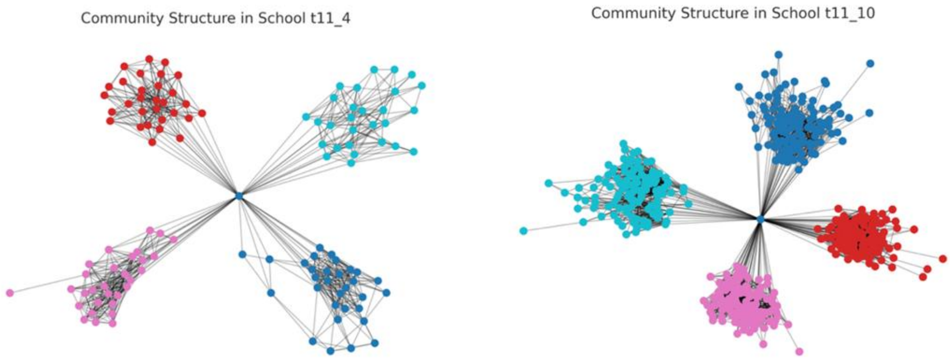

Figure 22 Community Structures of t1_4 and t11_10

5.8.4. Scale Freeness

Scale free networks are characterized by a degree distribution that follows a power law, where a few nodes (students) accumulate many connections, while most have relatively few. This structure typically indicates the presence of highly influential "hubs" that dominate network connectivity. To evaluate scale-freeness in the school networks, we analyzed the degree distributions using a log linear regression approach and examined both the slope fit (R^2) of each school's distribution.

Among the networks, t11_10 showed the strongest signs of scale-freeness, with a slope of −0.68 and an R^2 of 0.36, indicating a moderate fit to the power law model. This aligns with earlier findings of high modularity and structural separation, suggesting a few socially dominant individuals anchor multiple smaller communities. t11_7 and t1 also displayed negative slopes (−0.32 and −0.28, respectively), but their low R^2 values (0.13 and 0.06) suggest that the power law fit is weak and potentially inconsistent across the degree range.

On the other hand, t11_4 notably deviates from scale free behavior, with a positive slope and minimal explanatory power ($R^2 = 0.002$), indicating a more uniform distribution of social ties without dominant hubs. Overall, while some schools exhibit partial scale free characteristics, the majority reflect more evenly distributed social connectivity, highlighting diverse patterns of peer influence rather than centralized structures.

6. Conclusion

This

The combination of k-core and clique analysis reveals nuanced insights into student social structures:

- K-core analysis reveals the hierarchical and resilient substructures, identifying central individuals and the layered organization of social groups.
- Clique analysis highlights micro-communities with strong internal ties and identifies highly connected individuals with cross-group influence.

These insights are critical for designing evidence-based interventions and educational policies that foster inclusive, connected school environments. Specifically, schools with frag- mented or low-cohesion cores might benefit from programs aimed at increasing cross-group collaboration, while those with dominant cliques may require monitoring to ensure inclusivity and prevent exclusionary behavior.

7. Critique

While

References

[1] Brewe, E., Kramer, L. H., & Sawtelle, V. (2012). *Investigating student communities with network analysis of interactions in a physics learning center*. Physical Review Special Topics - Physics Education Research, **8**(1), 010101. https://doi.org/10.1103/PhysRevSTPER.8.010101
[Available at: https://arxiv.org/abs/1109.4139]

[2] Pérez-Flores, A., Ruiz-García, M., & Borge-Holthoefer, J. (2022). *Triadic influence as a proxy for compatibility in social relationships: The case of Spanish high schools*. PLOS ONE, **17**(12), e0279423. https://doi.org/10.1371/journal.pone.0279423
[Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9731467]