Spanish High Schools Network Analysis

Luis Irisarri Galera

(Dated: February 19, 2024)

Instituto de Física Interdisciplinar y Sistemas Complejos IFISC (CSIC-UIB), Campus UIB, 07122 Palma de Mallorca, Spain.

Abstract

Network theory serves as a vital tool in understanding complex systems, transcending disciplinary boundaries to offer insights into various fields of research. High schools, as microcosms of society, present intriguing subjects for network analysis, reflecting broader societal structures and dynamics. In this study, we conduct a comprehensive structural analysis of a Spanish high school network, aiming to elucidate the implications of its structure on social interactions. Our analysis reveals how the network's topology intricately captures sociological information, reflecting the underlying social dynamics within the educational setting. Ultimately, this study underscores the relevance of network theory in unraveling the complexities of social structures, offering valuable unbiased insights into the fundamental dynamics of human social behaviour.

1 Introduction

In the quest to understand the intricate fabric of complex systems, *network theory* emerges as a pivotal tool. The importance of *network theory* transcends traditional disciplinary boundaries, serving as a bridge that connects various fields of research [1–3]. Of course, the study of networks is not the only aspect of the scientific endeavour, but it serves as a crucial tool. Indeed, the collaboration of researchers from all walks of academia is imperative. The world around us is a tapestry of interwoven elements, and to fully understand it, one must adopt a cross-disciplinary perspective.

The exploration of social networks through the lens of network theory sheds light on the dynamics and structures that underpin social interactions and organizational patterns. High schools, as microcosms of society, present an interesting case study for social networks. The interactions within these educational institutions mirror broader societal networks, making high schools a valuable subject for network analysis. By examining the networks within high schools, researchers can gain insights into social cohesion, peer influences, and the diffusion of ideas and behaviours among adolescents [4–7].

The aim of this work is to delve into the structural analysis of a Spanish high school network, introduced by Ruiz-García et al. This analysis seeks not only to study the structure of the network but also to understand the implications of it, thereby gaining insight into the complex phenomena of human interactions.

This project is structured into several sections, each dedicated to a different aspect of the analysis. First, in §2, we will introduce the fundamental definitions, models, and methods used in this work. Then, in §3 we will present the results of the analysis and discuss their implications. Finally, in §4 we will summarize the findings and outline potential future research directions.

2 Methods

In this section, we present the foundational tools and concepts drawn from network theory pertinent to our study (sections 2.1 to 2.6). Our primary sources of reference include [1–3, 8] for network theoretical principles and [9] for community detection. Additionally, we provide an overview of the computational methodologies employed and detail the dataset employed in this research (sections 2.7 and 2.8).

2.1 Fundamental Definitions

Definition 1 (Network)**.** A **Network** or **Graph** G, is a pair (V, E). V is the **Vertex Set** of G; its elements are the **Vertices** or **Nodes** of the network. $E = V \otimes V$ is the **Edge Set** of G; its elements are the **Edges** or **Links** of the network.

Notation. We will denote the number of vertices of a network G as $|V| = n$ and the number of edges as $|E| = m$.

Definition 2 (Subgraph)**.** $G_s = (V_s, E_s)$ is a **Subgraph** of $G = (V, E)$ if $V_s \subseteq V$ and $E_s \subseteq \cup E$.

Definition 3 (Undirected Graph)**.** A network $G = (V, E)$ is **Undirected** if E is *symmetric*.

Definition 4 (Directed Graph)**.** A network $G = (V, E)$ is **Directed** or a **Digraph** if E is *non-symmetric*.

Definition 5 (Bipartite Graph)**.** A network $G = (V, E)$ is **Bipartite** if the nodes can be divided into disjoint sets $V_1 \cup V_2$ such that $(u, v) \in E \;\Rightarrow\; u \in V_i, v \in V_j,\; i \neq j$.

Definition 6 (Proper Edge)**.** A **Proper Edge** is an edge that joins two distinct vertices.

Definition 7 (Self-Loop)**.** A **Self-Loop** is an edge that joins a single endpoint to itself.

Definition 8 (Degree)**.** The **Degree** of a vertex v in a graph G, denoted k_v, is the number of proper edges incident on v plus twice the number of self-loops.

The **indegree** (k^{in}) of a vertex v in a digraph is the number of arcs directed to v; the **outdegree** (k^{out}) of vertex v is the number of arcs directed from v.

Definition 9 (Adjacency Matrix)**.** Suppose $G = (V, E)$ is a simple network where $V = \{1, 2, \ldots, n\}$. For $1 \leq i, j \leq n$:

$$A_{ij} \coloneqq \begin{cases} 1 & \text{if } (i, j) \in E, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Then the square matrix $A = \left(A_{ij}\right)$ is called the **Adjacency Matrix** of G.

Definition 10 (Walk)**.** A **Walk** in a network is a series of edges (not necessarily distinct):

$$W \coloneqq (u_1, v_1), \ldots, (u_p, v_p), \tag{2}$$

such that $v_i = u_{i+1}$ for $i = 1, \ldots, p-1$, $\forall u_i, v_i \in V$.

Definition 11 (Length)**.** The **Length** of a walk is the number of p edge-steps in the walk sequence W.

Definition 12 (Distance)**.** The **Distance** $d(u, v)$ between two vertices u, v in a network G is the length of the shortest walk between u and v.

Definition 13 (Trail)**.** A **Trail** is a walk in which no edge is repeated.

Definition 14 (Path)**.** A **Path** is a trail in which all u_i are distinct.

Remark. Since a shortest walk contains no repeated vertices or edges [8], it is also a trail and a path [1]. Therefore, we extend the definition of distance to be the length of the shortest path between two vertices u, v in a network G.

2.2 General Network Metrics

Having established the fundamental definitions of networks, we proceed to introduce several general metrics commonly utilized for characterizing network structure.

We begin by defining the *Density* of a network, which gives us a measure of how many edges are present in the network relative to the maximum possible number of edges.

Definition 15 (Density)**.** The **Density** of a network G is defined as the ratio of the number of edges (m) in G to the maximum number of possible edges in G, that is:

$$\delta(G) \coloneqq \frac{2m}{n(n-1)}. \tag{3}$$

We delve now into the concept of *average path length*, which offers insights into the overall connectedness of a network

Definition 16 (Average Path Length)**.** The **Average Path Length** $\bar{l}$ of a graph G, is the average distance between all pairs of vertices in G. That is,

$$\bar{l} \coloneqq \sum_{x,y \in V} \frac{d(x, y)}{n(n-1)}. \tag{4}$$

Another valuable measure regarding the connectedness, is the *diameter*.

Definition 17 (Diamater)**.** The **Diameter** D of a graph G, is the maximum distance between two vertices in G. That is,

$$D \coloneqq \max_{x,y \in V} \{d(x, y)\}. \tag{5}$$

We now shift focus to examining the clustering characteristics of the network, beginning with the local measure known as the *Watts-Strogatz clustering coefficient*.

Definition 18 (Watts-Strogatz Clustering Coefficient)**.** The **Clustering Coefficient** $\bar{C}$ of a network G is defined as the average of the clustering coefficients C_i of all the vertices in G, that is:

$$\bar{C} \coloneqq \frac{1}{n} \sum_{i=1}^{n} C_i \,, \qquad\qquad C_i \coloneqq \frac{2t_i}{k_i(k_i - 1)}, \tag{6}$$

where t_i is the number of triangles attached to node i of degree k_i.

Regarding a global measure of how clustered the network is, we introduce the *transitivity index* or *Newman Clustering Coefficient*.

Definition 19 (Transitivity Index)**.** The **Transitivity Index** C of a network G is defined by:

$$C \coloneqq \frac{3|C_3|}{|P_2|}, \tag{7}$$

where $|C_3|$, $|P_2|$ are the total number of triangles and paths of length 2 in G, respectively.

Remark. Although both the clustering coefficient and the transitivity index measure the degree of clustering in a network, we stress that they are not the same. For example in [10], it is shown that a family of graphs named *Core-Satellite* graphs have a clustering coefficient of 0, but a transitivity index of 1.

In order to measure the tendency of vertices to connect to other vertices with similar degree, we can use the *Assortativity* of a network.

Definition 20 (Assortativity)**.** The **Assortativity** or **Degree–degree correlation** r of a network G is defined as:

$$r \coloneqq \frac{\frac{1}{m}\sum_{(i,j)\in E} k_i k_j - \left[\frac{1}{2m}\sum_{(i,j)\in E}\left(k_i + k_j\right)\right]^2}{\frac{1}{2m}\sum_{(i,j)\in E}(k_i^2 + k_j^2) - \left[\frac{1}{2m}\sum_{(i,j)\in E}\left(k_i + k_j\right)\right]^2} \tag{8}$$

In particular, we say that a network is **Assortative** if $r > 0$, **Disassortative** if $r < 0$ and **Neutral** if $r = 0$. Obviously, by the equivalence with the *Pearson's Correlation Coefficient*, we have that $r \in [-1, 1]$.

Finally, we consider how bipartite the network is. For this, we use the *Bipartivity Index*:

Definition 21 (Bipartivity Index)**.** The **Bipartivity Index** b_e of a network G is defined as:

$$b_e \coloneqq \frac{\operatorname{tr} e^{-A}}{\operatorname{tr} e^{A}}. \tag{9}$$

2.3 Degree Distribution

Delving deeper into the network's structure, an intriguing aspect lies within its Degree Distribution $p(k)$. This distribution unveils the frequency with which nodes possess certain degrees, offering profound insights into the network's organizational principles and underlying dynamics.

Definition 22 (Degree Distribution)**.** The **Degree Distribution** $p(k)$ of a network G is the probability that a randomly chosen vertex has degree k.

$$p(k) \coloneqq \frac{n_k}{n}, \tag{10}$$

where n_k is the number of vertices in G with degree k.

The degree distribution is a fundamental characteristic of a network, and it is often used to classify networks into different categories. For example, a network is said to be *scale-free* if its degree distribution follows a power-law $p(k) \sim k^{-\gamma}$ for some $\gamma > 0$.

2.3.1 Fitted Distributions

In order to deepen our understanding of the degree distribution, we can fit it to different distributions. In this work, we will consider the following distributions:

- **Dagum distribution** or **Mielke distribution** is given by:

$$f(x,k,s) = \frac{kx^{k-1}}{(1+x^s)^{1+k/s}}, \tag{11}$$

  for $x, k, s > 0$ [11].

- **Generalized Hyperbolic distribution** is given by:

$$f(x,p,a,b) = \frac{\left(a^2-b^2\right)^{p/2}}{\sqrt{2\pi}a^{p-1/2}K_p\left(\sqrt{a^2-b^2}\right)}e^{bx} \times \frac{K_{p-1/2}\left(a\sqrt{1+x^2}\right)}{\left(\sqrt{1+x^2}\right)^{1/2-p}}, \tag{12}$$

  for $x, p \in (-\infty;\infty), |b| < a$ if $p \geq 0, |b| \leq a$ if $p < 0$. K_p denotes the modified Bessel function of the second kind and order p [12].

- **Exponentially modified Gaussian distribution** is given by:

$$f(x,K) = \frac{1}{2K}\exp\left(\frac{1}{2K^2} - x/K\right)\operatorname{erfc}\left(-\frac{x-1/K}{\sqrt{2}}\right), \tag{13}$$

  where x is a real number and $K > 0$.

In order to test the statistical significance of the fitted distributions, we will use the *Kolmogorov-Smirnov* test. A statistical procedure designed to evaluate the null hypothesis that a sample is drawn from a given distribution.

2.4 Centrality Measures

Another valuable measure of the network structure is the *Centrality*, which gives us a measure of the "importance" of a vertex in the network. It turns out, that network centrality is a very broad concept and there are many different ways to measure it, each way capturing a different aspect of the importance of a vertex. We can divide the centrality measures into two categories: *Classical Node Centrality Measures* and *Spectral Node Centrality Measures* as it is done in *A First Course in Network Theory* by Estrada and Knight. *Classical Node Centrality* accounts for the short-range influence of the node inside the network, while *Spectral Node Centrality* accounts for the long-range influence of the node inside the network. Now, we list the centrality measures considered in this work.

2.4.1 Classical Node Centrality

- **Degree Centrality** (DC): simply corresponds to degree, and clearly measures the ability of a node to communicate directly with others.

- **Close Centrality** (CC): characterizes how close a node is from the rest of the nodes. This closeness is measured in terms of the shortest path distance. The closeness of the node i in an

undirected network G is defined as:

$$\mathrm{CC}(i) \coloneqq \sum_{j \in V} \frac{n-1}{d(i,j)}, \tag{14}$$

- **Betweenness Centrality** (BC): characterizes how important a node is in the communication between other pairs of nodes. That is, the betweenness of a node accounts for the proportion of information that passes through a given node in communications between other pairs of nodes in the network. The betweenness of the node i in an undirected network G is defined as:

$$\mathrm{BC}(i) \coloneqq \sum_{i} \sum_{k} \frac{\rho(j,i,k)}{\rho(j,k)}, \qquad i \neq j \neq k, \tag{15}$$

where $\rho(j,k)$ is the total number of shortest paths between j and k, and $\rho(j,i,k)$ is the number of these that pass through node i in the network.

2.4.2 Spectral Node Centrality

- **Katz Centrality** (KC): is a generalization of degree centrality. The Katz centrality of a node i in a network G is defined as:

$$\mathrm{KC}(i) \coloneqq \left[(\mathbb{I} - \alpha A)^{-1}\mathbf{e}\right]_i, \tag{16}$$

where the parameter α is constrained between $0 < \alpha < \lambda_1^{-1}$, being λ_1 the largest eigenvalue of the adjacency matrix A. Here, $\backslash I$ represents the identity matrix, and $\mathbf{e}$ signifies a vector consisting of all ones. The α restriction is crucial for ensuring convergence.

- **Eigenvector Centrality** (EC): is conceptually similar to Katz centrality, but it is based on the eigenvector of the adjacency matrix. The eigenvector centrality of a node i is defined as:

$$\mathrm{EC}(i) \coloneqq \frac{1}{\lambda_1} \left(A\boldsymbol{q}_1\right)_i, \tag{17}$$

where λ_1 is the largest eigenvalue of the adjacency matrix A, and $\boldsymbol{q}_1$ is the corresponding eigenvector.

- **PageRank Centrality** (PC): is thought for directed graphs but when considering an undirected graph, we convert each edge in into two edges [13]. PC is closely related to eigenvector centrality, and it explicitly measures the importance of a node via the importance of other nodes pointing to it. The PC of a node i is defined as:

$$\mathrm{PC}(i) \coloneqq (P^T \overrightarrow{\mathrm{PR}})_i, \tag{18}$$

where $P \coloneqq \alpha S + (1-\alpha)/n\mathbf{e}\mathbf{e}^T$, $S \coloneqq H + (1/n)\tilde{\mathbf{k}}^{\mathrm{out}}\mathbf{e}^T$, $\tilde{\mathbf{k}}_i^{\mathrm{out}} \coloneqq 1$ if $k_i^{\mathrm{out}} = 0 \ \wedge \ 0$ if $k_i^{\mathrm{out}} > 0$, α is a parameter, $\overrightarrow{\mathrm{PR}}$ is the principal left-hand eigenvector of P, and:

$$H_{ij} \coloneqq \begin{cases} 1/k_i^{\mathrm{out}} & \text{if } (i,j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

- **Subgraph Centrality** (SC): is a generalized version of the Katz centrality; the idea behind

the SC is that we can characterize the importance of a node by considering its participation in all closed walks starting (and ending) at it. In particular, in this work we will consider the subgraph centrality defined as [14]:

$$\mathrm{SC}(i) \coloneqq \left(\sum_{l=0}^{\infty} c_l A^l \right)_{ii}, \tag{19}$$

where coefficients c_l are selected such that the infinite series converges.

2.5 Community Detection

Identifying communities within networks is a complex and widely debated task, primarily due to the absence of a universally accepted definition of what constitutes a community. To proceed, we define:

Definition 23 (Internal and External Degrees)**.** Given a subgraph $G_1(C, E_1) \subseteq G(V, E)$ with $n_C = |C|$ nodes, we define the **Internal** and **External** degrees to be the quantities:

$$k_i^{\mathrm{int}} = \sum_{i=\Gamma} A_{ij}, \qquad k_i^{\mathrm{ext}} = \sum A_{ij}, \tag{20}$$

respectively, where $\bar{C}$ is the complement of C and A is the adjacency matrix of G.

The number of links which connect nodes internally in the subgraph is given by:

$$m_C = \frac{1}{2} \sum_{i \in C} k_i^{\mathrm{int}}, \tag{21}$$

and the number of links which connect nodes in the subgraph with nodes outside the subgraph is given by:

$$m_{C-\bar{C}} = \sum_{i \in C} k_i^{\mathrm{ext}}. \tag{22}$$

Then, we define the *Internal* and *External Densities* as:

Definition 24 (Intra-Cluster and Inter-Cluster Density)**.** The **Intra-Cluster Density** δ_{int} and the **Inter-Cluster Density** δ_{ext} are defined as [9]:

$$\delta_{\mathrm{int}}(C) = \frac{2m_C}{n_C(n_C - 1)}, \qquad \delta_{\mathrm{ext}}(C) = \frac{m_{C-\bar{C}}}{n_C(n - n_C)}. \tag{23}$$

A critical characteristic of a community is its significantly higher intra-clustering density compared to both its inter-clustering density and the overall network's edge density. Consequently, the density of connections within the community should substantially surpass the network's average density. Given this premise, numerous methods and algorithms have been developed to identify communities by leveraging various network characteristics and analytical approaches. However, our goal transcends the mere application of diverse methodologies yielding various partitions; it's imperative to evaluate the effectiveness of the partition produced by a specific algorithm. This necessitates the introduction of *Quality Functions*.

Quality Functions serve to rate a network's partitioning by assigning a score, where a higher score indicates a superior partition quality. Among these, *Modularity* stands out as one of the most esteemed

Quality Functions.

Definition 25 (Modularity)**.**

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta \left(C_i, C_j \right), \tag{24}$$

where the sum runs over all pairs of vertices and $\delta(C_i, C_j)$ yields one if vertices i and j are in the same community ($C_i = C_j$), zero otherwise.

Despite its utility, Modularity is not without its flaws, such as the resolution limit issue, which can obscure the presence of smaller, distinct communities adjacent to larger ones. Other notable *Quality Functions* include *Performance*, which gauges the accuracy of node pair classifications within the same or different communities:

Definition 26 (Performance)**.**

$$P = \frac{|\{(i,j) \in E, C_i = C_j\}| + |\{(i,j) \notin E, C_i \neq C_j\}|}{n(n-1)/2}. \tag{25}$$

And *Coverage*, assessing the proportion of intra-community edges relative to the total edge count. The three *quality functions* are bounded between 0 and 1, where high scores suggest a well-defined community structure, whereas low scores indicate a lack of clear community delineation.

2.5.1 Community Detection Algorithms

In this work, we consider two algorithms for community detection. Both algorithms are based on the optimization of the modularity. In particular, we consider **Clauset-Newman-Moore Greedy Modularity Maximization** (see [15] for details) and **Louvain Community Detection** (see [16] for details).

2.6 Null Models

To deepen our understanding of the measurements obtained from the real network, it is beneficial to juxtapose these with characteristics observed in null models. Null models are specially designed networks that, while matching the actual network in terms of nodes and edges, differ in their degree distributions. This comparison aims to determine whether the network exhibits unique features—those not likely to arise merely by chance. In this context, we will specifically examine how our network aligns with or diverges from two established models: the *Erdös-Rényi* and the *Barabási-Albert networks*. Given the random nature of this models, we will perform 10 realizations of each model and compute the mean and standard deviation of the general network metrics.

2.6.1 Erdös-Rényi Model

In the model introduced by Erdös and Rényi in 1959, the process begins with n isolated nodes. Subsequently, each pair of nodes is considered, and a link is added between them with a probability p. In practice, a specific value of p is chosen to generate the network. For every pair of nodes, a uniform random number $u \in [0, 1]$ is generated. If $u \leq p$, a link is established between the nodes. A graph generated in this manner is referred to as an **Erdös-Rényi graph** G_{ER}. Due to their construction

method, Erdös-Rényi networks exhibit a degree distribution that adheres to a *Poisson distribution*:

$$p(k) = \frac{\bar{k}^k}{k} \exp(-\bar{k}) \tag{26}$$

where $\bar{k}$ is the average degree of the network.

2.6.2 Barabási-Albert Model

The Barabási-Albert model, introduced in 1999, is based on the principle of *preferential attachment*. This model begins with m_0 nodes, each connected to at least one other node. Subsequently, new nodes are added to the network, with each node forming $m \leq m_0$ links to existing nodes. The probability of a new node u connecting to an existing node $v \in V$ is proportional to the degree of the latter (k_v). A graph generated in this manner is referred to as a **Barabási-Albert graph** G_{BA}. This model generates networks with a degree distribution that adheres to a *power-law distribution*:

$$p(k) = \frac{2d(d-1)}{k(k+1)(k+2)} \approx k^{-3}. \tag{27}$$

2.7 Computing Specifics

For the numerical analysis of the network, we employed the `Python` programming language, specifically utilizing the `NetworkX` library for network analysis. This section outlines the computational tools and functions used to measure general network metrics, analyze degree distributions, compute centrality measures, detect community structures, and generate null models.

We computed the network's foundational metrics using predefined functions in `NetworkX`, as summarized in tab. 3. For metrics not directly supported by `NetworkX`, such as the Bipartivity Index (9), we developed custom functions to align with our specific definitions and requirements[1].

Metric	Python Function (`NetworkX`)
δ	`nx.density`
$\bar{l}$	`nx.average_shortest_path_length`
D	`nx.diameter`
$\bar{C}$	`nx.average_clustering`
C	`nx.transitivity`
r	`nx.degree_assortativity_coefficient`

Table 1: Python functions used from `NetworkX` to compute the general network metrics [eqs. (4) to (8)].

To fit the degree distribution, we leveraged the `scipy` library, utilizing the `curve_fit` function and the `.fit` method. Inspired by [17], we developed a function to fit the degree distribution to all continuous distributions available in the `scipy.stats`. In order to asses the statistical significance of the fitting, we used the `kstest`.

[1] The source code of the function `nx.spectral_bipartivity` does not adjust to the definition 21.

For the centrality measures, we used the functions listed in tab. 2. In particular, for the eq. (16), we used the `nx.katz_centrality` function with the parameter `alpha` set to 0.001 to ensure convergence, while for the eq. (18), we used the `nx.pagerank` function with the default parameter `alpha` set to 0.85.

Centrality Method	**Python Function (`NetworkX`)**
DC	`nx.degree_centrality`
CC	`nx.closeness_centrality`
BC	`nx.betweenness_centrality`
KC	`nx.katz_centrality`
EC	`nx.eigenvector_centrality`
PC	`nx.pagerank`
SC	`nx.subgraph_centrality`

Table 2: Python functions used from `NetworkX` to compute the centrality measures.

For the community detection, we used the functions from `nx.community`. In particular, regarding the algorithms, we used `greedy_modularity_communities` and the `louvain_communities`, while for the quality functions, we used the `modularity` and the `partition_quality` functions.

Finally, for the null models, we used the `nx.erdos_renyi_graph` with $p = \delta(G)$ and the `nx.barabasi_albert_graph` with $m_0 = m/n$ functions to generate the networks.

The codes used to perform the analysis are available at https://github.com/liris8.

2.8 Network Specifics

This study focuses on the "Spanish High Schools" (SHS) Networks repository, in particular on the "6" network within this collection. The SHS network maps the social interactions among high school students in Spain, where each node represents a student, and each edge denotes a social relationship between two students. Originally, this network is characterized as weighted and directed, capturing the intensity and direction of social interactions. However, for the purposes of our analysis, we simplify this model to treat the network as undirected and unweighted, streamlining our focus on the presence of social connections irrespective of their strength or directionality.

Furthermore, the nodes are associated with multiple attributes: *name*, *Curso*, *Grupo*, *Sexo*, *prosocial*, *crttotal* and *_pos*. For the purposes of this study, we will specifically consider the attribute related to course enrolment, identifying students by their course numbers: 1, 2, 3, or 6. This decision allows us to examine the network's structure with a focus on the academic grouping of students. For a comprehensive understanding of the network's compilation and the rationale behind these attributes, we direct readers to the "Materials and Methods" section in "Triadic Influence as a Proxy for Compatibility in Social Relationships".

Regarding the treatment of directed edges in our transition to an undirected network model, we adopt a broad approach by converting all directed edges into undirected ones. This method ensures that any form of reported social relation is recognized, facilitating a more inclusive analysis of the network's social fabric.

3 Results & Discussion

This section unveils the key findings derived from the analysis of the SHS Network. We commence with an overview of the general network metrics (§3.1), followed by an exploration of the degree distribution (§3.2). Subsequently, we delve into the outcomes of centrality measures (§3.3), and finally, community detection (§3.4).

3.1 General Network Metrics

The SHS network under study comprises $n = 534$ nodes and $m = 9527$ edges, resulting in a density of $\delta = 0.067$. General network metrics, including average path length ($\bar{l}$), diameter (D), clustering coefficient ($\bar{C}$), transitivity index (C), assortativity (r), and bipartivity index (b_e) [eqs. (4) to (9)], are computed for both SHS and the null models (G_{ER}, and G_{BA}). The results are summarized in tab. 3.

	$\bar{l}$	D	$\bar{C}$	C	r	b_e
G	2.65	5	0.52	0.46	-0.06	$1.83 \cdot 10^{-15}$
G_{ER}	2.018(7)	3.0(0)	0.0674(8)	0.0673(7)	-0.007(9)	$1.5(6) \cdot 10^{-10}$
G_{BA}	2.079(4)	3.0(0)	0.132(3)	0.1270(15)	-0.023(8)	$3(1) \cdot 10^{-15}$

Table 3: General Network Metrics [eqs. (4) to (9)] for SHS (G) and the null models ($G_{\mathrm{ER}}, G_{\mathrm{BA}}$). Values in parentheses represent the standard deviations of the mean after 10 realizations of the null models.

Our analysis reveals that the metrics of the SHS network align with those reported in the repository [18]. Notably, considering the network as undirected and unweighted, we encountered a lack of bibliographic references explicitly providing these structural measures.

To gauge the relative density of our network, we compare its density with that of other real-world networks, such as those examined in [19]. Interestingly, we find that our network's density exceeds that of most networks, with the exception of a network representing a mouse retina, which exhibits a density of $\rho = 0.157$. Thus, our analysis confirms that our network can be classified as moderately dense.

Besides the assortativity (r) and bipartivity index (b_e), which are negligible across all networks, the SHS network demonstrates considerably higher values for general network metrics compared to both null models. This observation suggests that the network's structural properties are unlikely to emerge by chance alone.

The average path length and diameter of the SHS network surpass those of both null models, indicating a less interconnected network. This aligns with the network's embedded structure, which limits connections between students in different courses and classes. Notably, despite the network's large size, the average path length required to connect any two points is remarkably short, typically spanning only 2 to 3 nodes and never exceeding 5 nodes. This level of connectivity is higher than that observed in other student networks of similar size, as reported in [20].

We also find that the network exhibits significant clustering, both locally and globally, with clustering coefficients of $\bar{C} = 0.52$ and $C = 0.46$, respectively. This level of clustering is higher than other student networks of similar size, as reported in [20]. These features resemble those of a small-world network

[21]. Computing the *small-world coefficient* $\bar{C}\bar{l}_{\mathrm{ER}}/\bar{l}\bar{C}_{\mathrm{ER}}$ we obtain a value of $5.88 > 1$ indicating that the network exhibits small-world.

On the other hand, we observe that the network exhibits low assortativity, a common feature found in student networks, as reported in [20]. Furthermore, we appreciate the neutrality of the network by comparing it with other real-world networks, as shown in [22]. This suggests that students do not preferentially attach to others with similar or different degrees, reflecting a lack of degree-based social preference within the network. Finally, the network's non-bipartiteness is reflected in the low value of the bipartivity index $b_e = 1.83 \times 10^{-15}$, significantly smaller than that of other social networks, as demonstrated in [23]. This indicates that students are not divided into two distinct groups within the network.

3.2 Degree Distribution

We proceed by analysing the degree distribution of the SHS network, comparing it to the degree distributions of the null models. The main results are summarized in tab. 4 and figs. 1 and 2.

For the SHS network, we find an average degree of $\langle k \rangle = 36 \pm 15$ slightly higher than the average class size of 30. This aligns with the common observation that students tend to connect more within their class but also with a few students from other classes.

We fit the probability density function (PDF) and cumulative distribution function (CDF) of the SHS network to several distributions, finding that the Mielke distribution provides the best fit with a p-value of 0.5592. The Generalized Hyperbolic distribution and Exponentially Modified Gaussian distribution also exhibit good fits with p-values of 0.4209 and 0.3637, respectively. It is noteworthy that these distributions are not commonly found in social networks. Refer to tab. 4 for the fitted parameters and p-values.

Distribution	**Parameters**	**p-value**
Mielke	$k = 4.36$, $s = 5.14$	0.5592
Genhyperbolic	$p = -2.27$, $a = 0.98$, $b = 0.98$	0.4209
Exponnorm	$K = 1.69$	0.3637

Table 4: Fitted parameters and p-values for the PDF and CDF of the SHS network. The p-values are obtained from the Kolmogorov-Smirnov test.

On the other hand, considering the null models G_{ER} and G_{BA}, the average degree are $\langle k_{\mathrm{ER}} \rangle = 35.7 \pm 5.8$ and $\langle k_{\mathrm{BA}} \rangle = 32 \pm 22$, respectively. The PDF and CDF are shown in figs. 2a and 2b respectively. However, none of the null models provide good fits for the degree distribution of the SHS network. This discrepancy underscores the unique structural characteristics of the SHS network.

We attempted to fit the theoretical Poisson and Power-law distributions to the null models, resulting in parameters consistent with expectations for the ER model with $\bar{k} = 36.4(3)$, but not for the BA model. This discrepancy may be attributed to the sensitivity of the Power-Law fitting method, as discussed in [24]. Overall, none of the null models examined are able to replicate the features exhibited by the SHS network, which suggests the presence of underlying social structures that differ from pure randomness.

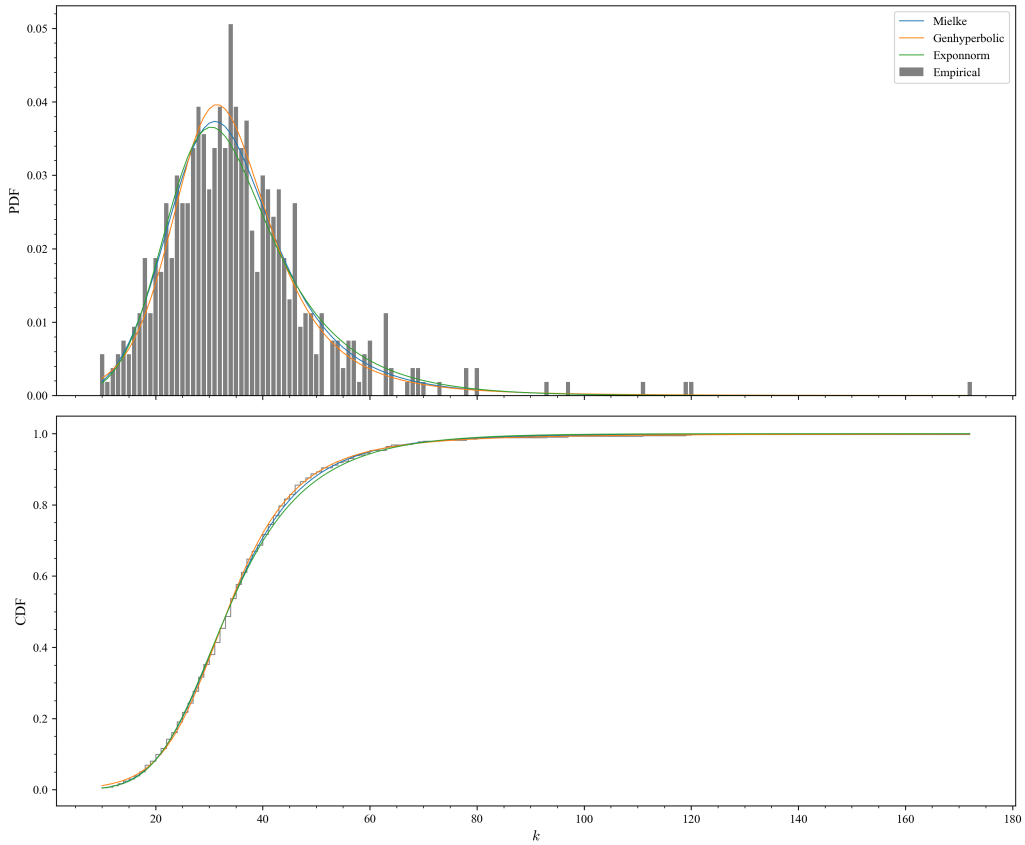


Figure 1: Fittings of the PDF and CDF of the SHS network. Both the PDF and CDF are fitted to a Mielke distribution, Burr distribution, and Generalized Hyperbolic distribution [eqs. (11) to (13)]. The fitted parameters are given in tab. 4.

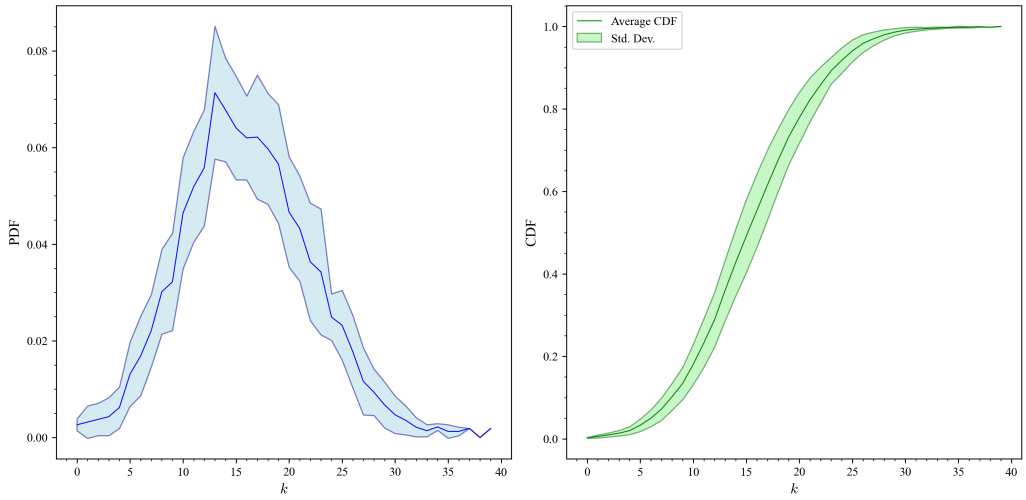
3.3 Centrality Measures

We present the results of the centrality measures for the SHS network in section 3.3. The 25 most central nodes are ranked according to each centrality measure.

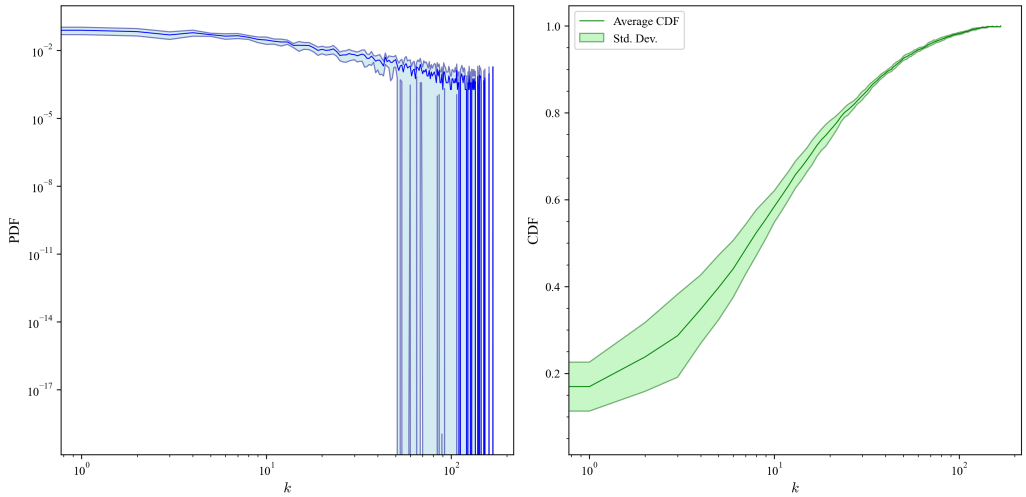
To compare the centrality measures, we computed the number of common nodes between all pairwise comparisons of the methods, as shown in fig. 3. We found that, on average, around 11 nodes are common between the pairwise comparisons, representing nearly half of the ranking. This consistency aligns with previous findings indicating that the most central nodes tend to align across different centrality measures [25]. Despite these similarities, the ranking order may vary, with nodes ranking differently across measures.

Notably, the pair (DC, KC) share the 25 ranked nodes, albeit in different orders. This observation is intriguing, given that Katz centrality is a generalization of degree centrality, suggesting distinctions in the local and global influence of nodes. Similarly, the (EC, SC) pair shares the 25 ranked nodes, despite the SC's introduction to differentiate node rankings [14]. Additionally, we observed that (DC, PC) and (KC, PC) share 21 nodes, while (CC, BC) share 15 nodes. Conversely, the most divergent rankings are (CC, EC) and (CC, SC), with only 2 common nodes.

On the other hand, all rankings only share the 225 node. We could argue that the most central nodes, on average, are 145, 288, and 530. We find that these nodes represent students from different courses (1, 2, and 3 respectively) and have high "prosociality" according to the data.



(a)



(b)

Figure 2: PDF and CDF of the degree distribution for the null models. (a) G_{ER} and (b) G_{BA}. The actual degree distribution of the SHS network is not included for clarity purposes.

3.4 Community Detection

We conducted community analysis using both the greedy optimization and Louvain algorithms, resulting in four communities identified by each algorithm. The quality functions, as shown in tab. 6, indicate similar results between the two algorithms, with the Louvain algorithm demonstrating slightly better quality functions.

Algorithms	**Modularity**	**Coverage**	**Performance**
Greedy Optimization	0.7124	0.9641	0.8109
Louvain	0.7138	0.9654	0.8113

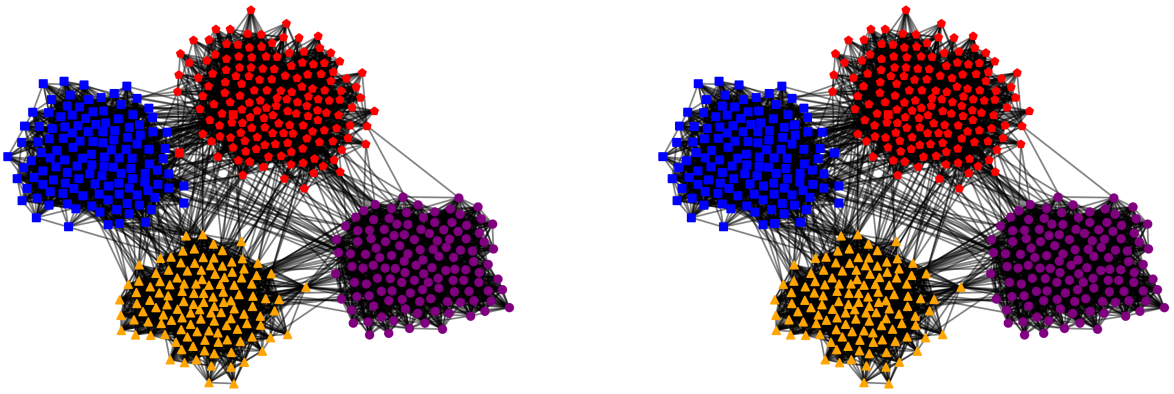
Table 6: Community detection quality functions for the Greedy Optimization and Louvain algorithms.

The resulting communities, depicted in fig. 4, predominantly align with the courses taken by the students. However, notable differences are observed. The Greedy Optimization algorithm fails to capture a specific node (marked with a red square) in the blue community, which the Louvain algorithm

Rank	DC	CC	BC	KC	EC	PC	SC
1	288	288	288	288	145	288	145
2	145	299	299	145	198	530	198
3	530	357	153	530	227	145	227
4	198	289	468	198	242	198	242
5	227	468	99	227	231	299	231
6	299	225	133	299	153	227	153
7	71	348	225	153	225	468	225
8	153	283	223	71	159	350	159
9	468	133	95	350	161	71	161
10	350	380	237	468	149	498	149
11	242	223	163	242	212	153	212
12	498	454	51	498	181	133	181
13	28	467	530	231	250	456	250
14	231	132	454	28	213	454	213
15	133	163	357	225	165	28	165
16	225	190	126	133	223	45	223
17	45	418	467	45	237	342	237
18	159	530	359	159	192	242	192
19	161	137	9	161	221	374	221
20	29	350	348	223	142	463	142
21	342	23	289	212	150	29	150
22	212	475	283	29	197	63	197
23	223	359	63	342	209	528	209
24	456	138	12	454	220	231	220
25	454	131	115	456	147	225	147

Table 5: The 25 most central nodes according to each centrality measure. In KC we have considered $\alpha = 0.001$ to ensure convergence. In PC, we have used the default parameter of the `NetworkX` library, that is $\alpha = 0.85$.

successfully identifies. Additionally, both algorithms fail to capture a square node found in the red community, potentially indicating a student who has repeated a course.



(a) (b)

Figure 4: Community detection results. (a) Greedy optimization and (b) Louvain algorithm. We denote the school courses with different shapes; triangles: Course 1, squares: Course 2, pentagons: Course 3, and circles: Course 6.

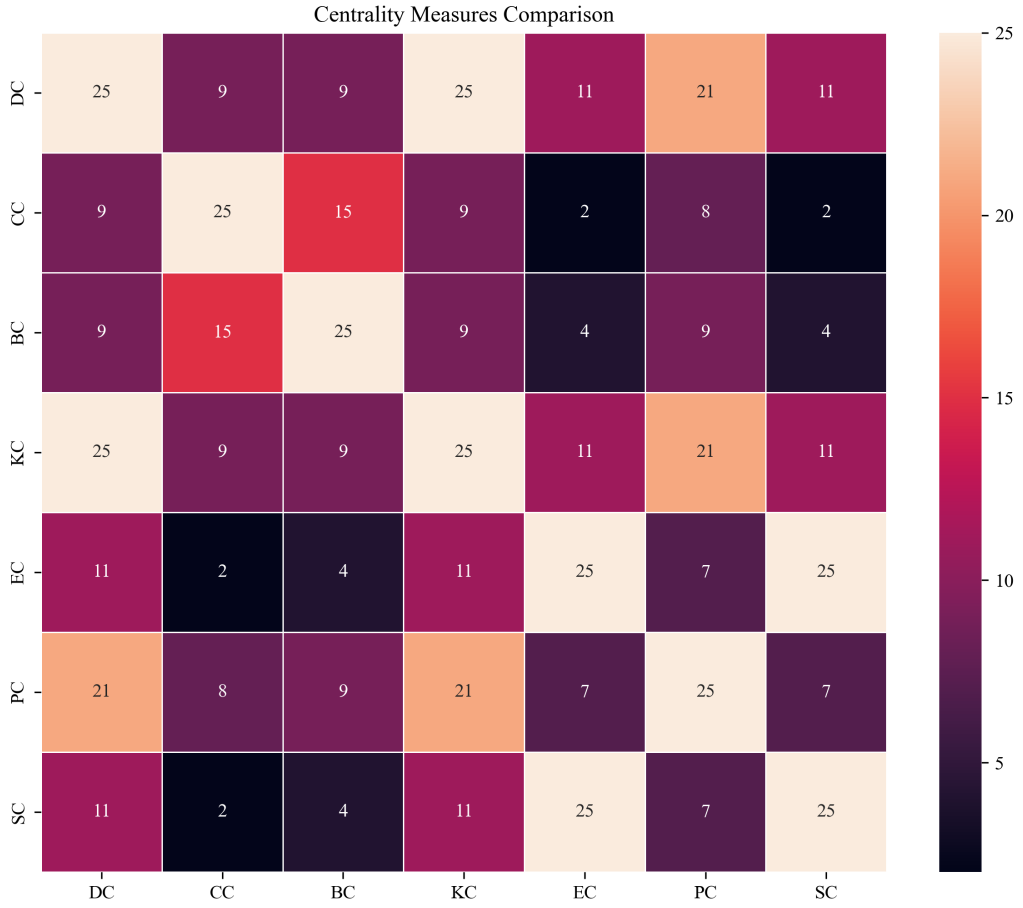


Figure 3: Centrality measures pairwise comparison.

4 Conclusions

In conclusion, our analysis of the SHS network reveals several key insights into its structural properties. Firstly, the agreement between the general network metrics computed in this study and those available in the data repository underscores the reliability of our analysis. Additionally, we observe that the average degree of the network slightly exceeds the average class size, suggesting that students tend to form connections not only within their own classes but also with a few individuals from other classes. Furthermore, our investigation into the degree distribution reveals that the Mielke distribution provides the best fit, indicating the presence of non-trivial structural patterns in the network.

Moreover, our analysis of centrality measures highlights the significance of certain nodes, particularly those from different courses exhibiting high prosociality. Community detection algorithms further support the notion that the network's structure reflects the courses taken by students. Across these analyses, a consistent theme emerges: the network's topology intricately captures sociological information, reflecting the underlying social dynamics within the educational setting.

Looking ahead, there are several promising avenues for future research. Firstly, delving into the heterogeneity index [26] could yield deeper insights into the network's diversity and unveil any underlying structural patterns. Exploring alternative null models, such as the Watts-Strogatz model, may offer valuable comparisons and shed light on the distinctive features of the SHS network. Regarding centrality measures, it would be interesting to explore the social significance of the most central nodes by conducting further research. Additionally, conducting a finer-grained examination of the network

within individual courses could offer valuable insights into its structure. Analyzing the network at this level may uncover unique characteristics, like clustering tendencies or influential nodes, unique to each course. This finer-grained analysis would illuminate how course structures and academic environments shape social interactions among students, enriching our understanding of the network's dynamics.

Moreover, a meta-analysis involving more networks could offer a broader understanding of the SHS network's structure and potentially uncover general patterns akin to previous studies [4, 5]. Finally, considering a different symmetrization approach for the network, focusing only on interactions between students with bi-directional connections, may offer fresh perspectives. Comparing the results obtained from this symmetrization method with those from our current study could provide valuable insights into the network's behaviour under different conditions.

In essence, our study highlights the relevance of network theory in understanding complex social structures. By uncovering the intricate interplay between individual interactions and broader societal influences, network analysis offers a powerful framework for exploring the underlying mechanisms shaping social networks. As we continue to delve deeper into the complexities of network structures, we gain valuable insights into the fundamental dynamics of human social behaviour.

References

[1] Ernesto Estrada and Philip A. Knight. *A First Course in Network Theory*. First edition. Oxford, United Kingdom: Oxford University Press, 2015. 254 pp. ISBN: 978-0-19-872645-6 978-0-19-872646-3 (cit. on pp. 1–3, 5).

[2] Ernesto Estrada. *The Structure of Complex Networks: Theory and Applications*. New York: Oxford university press, 2012. ISBN: 978-0-19-959175-6 (cit. on pp. 1, 2).

[3] Mark Newman. *Networks*. Vol. 1. Oxford University Press, Oct. 18, 2018. ISBN: 978-0-19-880509-0. DOI: 10.1093/oso/9780198805090.001.0001. URL: https://academic.oup.com/book/27884 (visited on 02/17/2024) (cit. on pp. 1, 2).

[4] Miguel Ruiz-García et al. "Triadic Influence as a Proxy for Compatibility in Social Relationships". In: *Proceedings of the National Academy of Sciences* 120.13 (Mar. 28, 2023), e2215041120. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2215041120. URL: https://pnas.org/doi/10.1073/pnas.2215041120 (visited on 02/05/2024) (cit. on pp. 1, 10, 17).

[5] Miguel A. González-Casado et al. "Towards a General Method to Classify Personal Network Structures". In: (Feb. 18, 2024). DOI: 10.31235/osf.io/23efd. URL: https://osf.io/23efd (visited on 02/18/2024) (cit. on pp. 1, 17).

[6] Diego Escribano et al. "Evolution of Social Relationships between First-Year Students at Middle School: From Cliques to Circles". In: *Scientific Reports* 11.1 (1 June 3, 2021), p. 11694. ISSN: 2045-2322. DOI: 10.1038/s41598-021-90984-z. URL: https://www.nature.com/articles/s41598-021-90984-z (visited on 02/18/2024) (cit. on p. 1).

[7] Diego Escribano et al. "Stability of the Personal Relationship Networks in a Longitudinal Study of Middle School Students". In: *Scientific Reports* 13.1 (1 Sept. 4, 2023), p. 14575. ISSN: 2045-2322. DOI: 10.1038/s41598-023-41787-x. URL: https://www.nature.com/articles/s41598-023-41787-x (visited on 02/18/2024) (cit. on p. 1).

[8] Jonathan L Gross, Jay Yellen, and Mark Anderson. *Graph Theory and Its Applications* (cit. on pp. 2, 3).

[9] Santo Fortunato. "Community Detection in Graphs". In: *Physics Reports* 486.3-5 (Feb. 2010), pp. 75–174. ISSN: 03701573. DOI: 10.1016/j.physrep.2009.11.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S0370157309002841 (visited on 02/17/2024) (cit. on pp. 2, 7).

[10] Ernesto Estrada and Michele Benzi. "Core–Satellite Graphs: Clustering, Assortativity and Spectral Properties". In: *Linear Algebra and its Applications* 517 (Mar. 2017), pp. 30–52. ISSN: 00243795. DOI: 10.1016/j.laa.2016.12.007. URL: https://linkinghub.elsevier.com/retrieve/pii/S0024379516305869 (visited on 10/09/2023) (cit. on p. 4).

[11] *Scipy.Stats.Mielke — SciPy v1.12.0 Manual*. URL: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mielke.html#rfff07386050d-2 (visited on 02/17/2024) (cit. on p. 5).

[12] Ernst Eberlein and Karsten Prause. "The Generalized Hyperbolic Model: Financial Derivatives and Risk Measures". In: *Mathematical Finance — Bachelier Congress 2000*. Ed. by Hélyette Geman et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 245–267. ISBN: 978-3-642-08729-5 978-3-662-12429-1. DOI: 10.1007/978-3-662-12429-1_12. URL: http://link.springer.com/10.1007/978-3-662-12429-1_12 (visited on 02/17/2024) (cit. on p. 5).

[13] *Pagerank — NetworkX 3.2.1 Documentation*. URL: https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.link_analysis.pagerank_alg.pagerank.html (visited on 02/17/2024) (cit. on p. 6).

[14] Ernesto Estrada and Juan A. Rodríguez-Velázquez. "Subgraph Centrality in Complex Networks". In: *Physical Review E* 71.5 (May 6, 2005), p. 056103. DOI: 10.1103/PhysRevE.71.056103. URL: https://link.aps.org/doi/10.1103/PhysRevE.71.056103 (visited on 02/17/2024) (cit. on pp. 7, 13).

[15] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. "Finding Community Structure in Very Large Networks". In: *Physical Review E* 70.6 (Dec. 6, 2004), p. 066111. DOI: 10.1103/PhysRevE.70.066111. URL: https://link.aps.org/doi/10.1103/PhysRevE.70.066111 (visited on 02/18/2024) (cit. on p. 8).

[16] Vincent D. Blondel et al. "Fast Unfolding of Communities in Large Networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (Oct. 2008), P10008. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/P10008. URL: https://dx.doi.org/10.1088/1742-5468/2008/10/P10008 (visited on 02/18/2024) (cit. on p. 8).

[17] s_sherly. *Fitting Empirical Distribution to Theoretical Ones with Scipy (Python)?* Stack Overflow. Nov. 6, 2021. URL: https://stackoverflow.com/q/6620471/20569862 (visited on 02/17/2024) (cit. on p. 9).

[18] Tiago de Paula Peixoto <tiago@skewed.de>. *Spanish_highschools — Spanish High Schools (2023).* URL: https://networks.skewed.de/net/spanish_highschools (visited on 02/18/2024) (cit. on p. 11).

[19] Cristopher G. S. Freitas et al. "A Detailed Characterization of Complex Networks Using Information Theory". In: *Scientific Reports* 9.1 (Nov. 13, 2019), p. 16689. ISSN: 2045-2322. DOI: 10.1038/s41598-019-53167-5. pmid: 31723172 (cit. on p. 11).

[20] M. E. J. Newman. "The Structure and Function of Complex Networks". In: *SIAM Review* 45.2 (Jan. 2003), pp. 167–256. ISSN: 0036-1445, 1095-7200. DOI: 10.1137/S003614450342480. arXiv: cond-mat/0303516. URL: http://arxiv.org/abs/cond-mat/0303516 (visited on 02/19/2024) (cit. on pp. 11, 12).

[21] Duncan J. Watts. *Small Worlds: The Dynamics of Networks between Order and Randomness.* 8. print., 1. paperback print. Princeton Studies in Complexity. Princeton, N.J.: Princeton University Press, 2004. 262 pp. ISBN: 978-0-691-11704-1 978-0-691-00541-6 (cit. on p. 12).

[22] M. E. J. Newman. "Assortative Mixing in Networks". In: *Physical Review Letters* 89.20 (Oct. 28, 2002), p. 208701. DOI: 10.1103/PhysRevLett.89.208701. URL: https://link.aps.org/doi/10.1103/PhysRevLett.89.208701 (visited on 02/18/2024) (cit. on p. 12).

[23] Ernesto Estrada and Juan A. Rodríguez-Velázquez. "Spectral Measures of Bipartivity in Complex Networks". In: *Physical Review E* 72.4 (Oct. 7, 2005), p. 046105. DOI: 10.1103/PhysRevE.72.046105. URL: https://link.aps.org/doi/10.1103/PhysRevE.72.046105 (visited on 02/18/2024) (cit. on p. 12).

[24] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. "Power-Law Distributions in Empirical Data". In: *SIAM Review* 51.4 (2009), pp. 661–703. ISSN: 0036-1445. JSTOR: 25662336. URL: https://www.jstor.org/stable/25662336 (visited on 02/18/2024) (cit. on p. 12).

[25] Lembris Laanyuni Njotto. "Centrality Measures Based on Matrix Functions". In: *Open Journal of Discrete Mathematics* 08.04 (2018), pp. 79–115. ISSN: 2161-7635, 2161-7643. DOI: 10.4236/ojdm.2018.84008. URL: http://www.scirp.org/journal/doi.aspx?DOI=10.4236/ojdm.2018.84008 (visited on 02/18/2024) (cit. on p. 13).

[26] Ernesto Estrada. "Quantifying Network Heterogeneity". In: *Physical Review E* 82.6 (Dec. 2, 2010), p. 066102. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.82.066102. URL: https://link.aps.org/doi/10.1103/PhysRevE.82.066102 (visited on 10/15/2023) (cit. on p. 16).