Network Analysis Project Relation

Emanuele Sinagra, Degree Programme

Abstract

This paper describes how to identify, quantify and approach the lack of integration in elementary school students using network science techniques.
We describe how was collected the data needed to study the integration, measured in terms of student interaction.
The data collected shape the network that we will study to identify the most interesting students and some network structure properties.
For instance, we identify the students who are most isolated, most popular, most frequently bridges for often students' potential interactions, and how much the network is heterophile/homophile.
We will provide a practical application to connect poorly integrated students with popular students and validate the results.

Context

The terms diversity, equity, and inclusion (DEI)[1] are words becoming very popular in many fields such as the workplace, academia, and medicine.
We know that diversity, considered a cultural advantage, is now an integral part of modern and industrialized societies.
While equity guarantees equal opportunities in the application context, we want to focus our study on inclusion which is the desired result.
The goal of inclusion is to create a united and cohesive environment where individuals are perfectly integrated.

Problem and Motivation

Inclusion is an important step from childhood when the child goes to primary school.
A lack of inclusion from childhood could lead to pathologies and poor prospects[2].
In academia, poor social inclusion could be detected by poor social interaction between schoolmates.
Our study aims to find a method to identify student inclusion shortages in a primary school network. Identifying an inclusion deficit is the first step in taking corrective action.

Datasets

We can think of a primary school as a network in which students and teachers are the nodes of the network and the interactions between students and teachers are the indirect edges.
Nodes have two attributes:

1.https://en.wikipedia.org/wiki/Diversity,_equity,_and_inclusion#Diversity,_Equity,_and_Inclusion_in_Academia
2.https://assr.regione.emilia-romagna.it/pubblicazioni/dossier/doss057/@@download/publicationFile/dossier%2057.pdf

1. *classname*, the school class name and its grade;
2. *gender*.

Edges have two weights associated with them:

1. *duration*, the cumulative time spent in daily interactions, measured in seconds;
2. *count*, for the number of times established during the school day.

We are aware that student interaction can differ from day to day, so the results should be averaged over multiple days of study, where each day is associated with a different social network.
Datasets are provided by Sociopatterns as a publicly available resource.
This resource is under creative commons attribution non-commercial ShareAlike.
The datasets were originally used to study transmission opportunities of respiratory infections in the paper 'High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School'[3].
Data interactions were collected in a primary school in Lyon (France) on 1st and 2nd October 2009 using proximity-sensing badges based on radio frequency identification (RFID) devices that will trigger an interaction at a close range of about 1-1.5m.
It records 77,602 contact events between 242 individuals (232 children and 10 teachers).
The datasets format is two gexf files, one per day of the study, so Gephi and Networkx Python library will be perfect tools to handle, manipulate and compute metrics on the data.

Validity and Reliability

Data collection was performed by offering all participants Radio-Frequency Identification (RFID) badges (nRF24L01 Single Chip 2.4GHz Transceiver).
These badges were worn on the chest throughout the school day, except during sports activities.
RFID badges register face-to-face interaction within approximately 1-1.5m with a success rate of more than 99% in a 20-second interval.
The registration infrastructure receives from the badge the start of an interaction which counts as 20 seconds, this interaction can be renewed by the device within 20 seconds, otherwise, it is interrupted.
This data collection methodology sets up 3 scenarios that impact the validity of our study:
- Lost Interaction: An interaction shorter than 20 seconds is not recorded.
- Fictitious interaction: An interaction a very short or accidental could be recorded as an interaction of at least 20 seconds;
- Unexpected interaction: An interaction during sports activities is not recorded because badges are not worn.

We believe it is appropriate for the study to cover many days, minimizing the possibility of a significant impact on the described scenarios.
This will also prevent students' behaviours from being influenced by the presence of badges.
To reproduce results and avoid random effects, you need to set these metric parameters:

- Eigenvector centrality: max iteration = 100, weight = None
- Betweenness centrality: k = 236, seed = 42, normalized = True, weight = None

3.https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0023176

In the Eigenvector centrality[4], we choose max interaction of 100 because it guarantees to find a solution in the computation of values and weight equals None to calculate results in terms of different interactions.
In the Betweenness centrality[5], we choose the maximum possible value k (equal to the number of nodes) to have a better result value approximation, a seed value of 42 to avoid randomness, normalized equals True to normalize the result in a range of [0,1], and weight equals None to calculate results in terms of different interactions.

Measures

We define a graph as $G = (V, E)$ with $|V|$ vertices and $|E|$ edges.
In the study application context vertex are students or teachers and edges are interactions weighted by count and cumulative time of interactions.
We define weights for i-th interaction E_i as t_i for the cumulative time consumed in interactions and c_i for the count of the number of interactions.
The first simple metric we want to use is degree centrality which counts the number of edges for a vertex. The degree centrality of a vertex v, for a given graph G, is defined as:

$$C_D(v) = \deg(v) \text{ where } \deg(i) = \sum_{j=i}^{n} E_{ij}$$

Considering the weighted degree centrality, we use the formula:

$$deg_t(i) = \sum_{j=i}^{n} E_{ij} \cdot p_i$$

Where p_i is t_i for weight as cumulative time or c_i for the count of interactions.
By ranking the node's degree centrality ascending or setting a cut-off we can detect inclusion troubles.
A good inclusion has the highest values in deg_t and deg_c, low values otherwise.
We prefer longer fewer interactions than frequent shortest interactions.
The Dunbar[6] number set the upper bound to 150 as the maximum limit of significant relationships that an individual can maintain.
The second useful metric is Eigenvector centrality (using c as weight) to identify important students in terms of interaction with students having more different interactions. We use the formula:

$$x_i = \left(\sum_{j \in neighbours(i)} \deg(j) \right) - \deg(i)$$

By finding more important students in terms of inclusion we can evaluate how to connect these students with other students with a lack of inclusion.
The third metric we use is Betweenness centrality to quantify the number of times a node acts as a bridge along the shortest path between two other nodes.
In an undirected network with at most one shortest path between any pair of nodes and let be 1 if the node *i* lie on the shortest path from the source *s* to the destination *d*, 0

4.https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.centrality.eigenvector_centrality.html#networkx.algorithms.centrality.eigenvector_centrality
5.https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.centrality.betweenness_centrality.html#networkx.algorithms.centrality.betweenness_centrality
6.https://it.wikipedia.org/wiki/Numero_di_Dunbar

otherwise.
The betweenness centrality is given by the formula: $x_i = \sum_{sd} n_{sd}^i$ for all shortest paths.
The nodes on "trafficked" shortest paths have a more central role in the network.
Our study will consider the nodes with higher betweenness centrality to choose the best walk from the lower node's degree centrality to the higher node's eigenvector centrality walking by a node with higher betweenness centrality.
For instance, we want to connect an isolated student with another well-included student. We can do it by passing different students, but we will choose the student that acts as a walking point more times for all the shortest paths because it can open new shortest links.
To verify if the chosen destination node with higher eigenvector centrality is a choice in line with the network's nature, we can inspect the Homophily degree modularity using the assortativity coefficient.

$$r = \frac{\sum_{ij}\left(A_{ij} - \frac{d_i d_j}{2m}\right) x_i x_j}{\sum_{ij}\left(d_i \delta_{ij} - \frac{d_i d_j}{2m}\right) x_i x_j}$$

Where d_x is the degree of node x, m is the number of edges, and $\delta_{kl} = \begin{cases} 1 \; if \; k = l \\ 0 \; otherwise \end{cases}$
A positive value indicates that the index is homogeneous by degree, negative otherwise.
Because a social network tends to maintain its natural characteristics, it can use to check if this connection method is in line with the nature of the network.
It is conciliable if the network has a non-positive degree assortativity coefficient.
The fourth metric we use is k-core used to find a set of nodes where each one is reachable to at least k of the others.
We can set k to a high value to find the best reachable nodes so that we can remove selected nodes. The remaining isolated nodes are potentially the nodes that can have trouble in inclusion, missing more connected nodes.
This technique can be used to take preventive action for students connected with other students by a few entry points.

Results

We start by giving some information about the network in Figure 1.
The day one dataset records 37,351 contact events (5899 different) between 236 individuals (226 children and ten teachers).
The network diameter is 3 and the average path length is 1.86.
Each student has an average of 323 daily contacts with other schoolmates (of which 49,99 are distinct interactions), leading to an average daily interaction time of 10560 seconds.

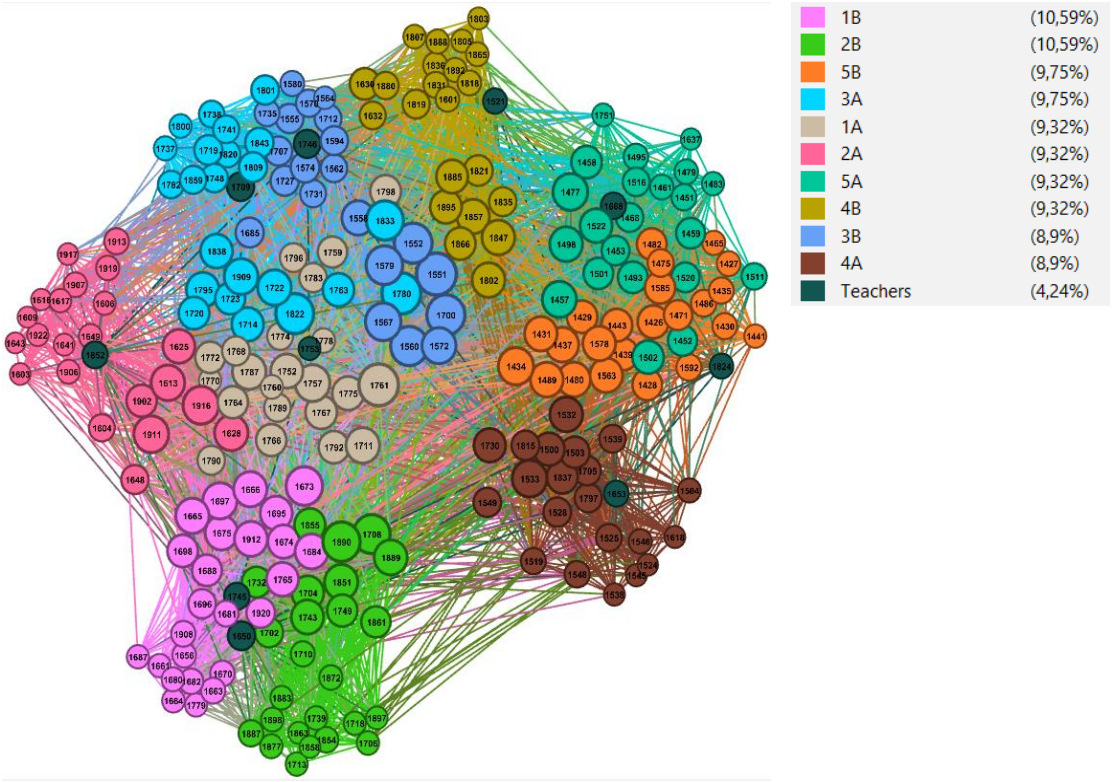

Figure 1 - Graph nodes have id as the label, the degree as dimension, and the class as colour.

Degree Centrality

We found the more isolated student using weighted degree centrality by count and duration (Table 1).

Student ID	Weighted degree centrality by count	Weighted degree centrality by duration
1483	67	1480
1753	70	1620
1441	81	2040
1521	96	2280
1913	94	2340
1465	106	2580
1805	124	2860
1609	99	2920
1710	103	2960
1917	102	3140

Table 1 - The first ten nodes with weighted degrees ordered by duration (ascending)

We sort by duration because we prefer longer fewer interactions than frequent shortest interactions.

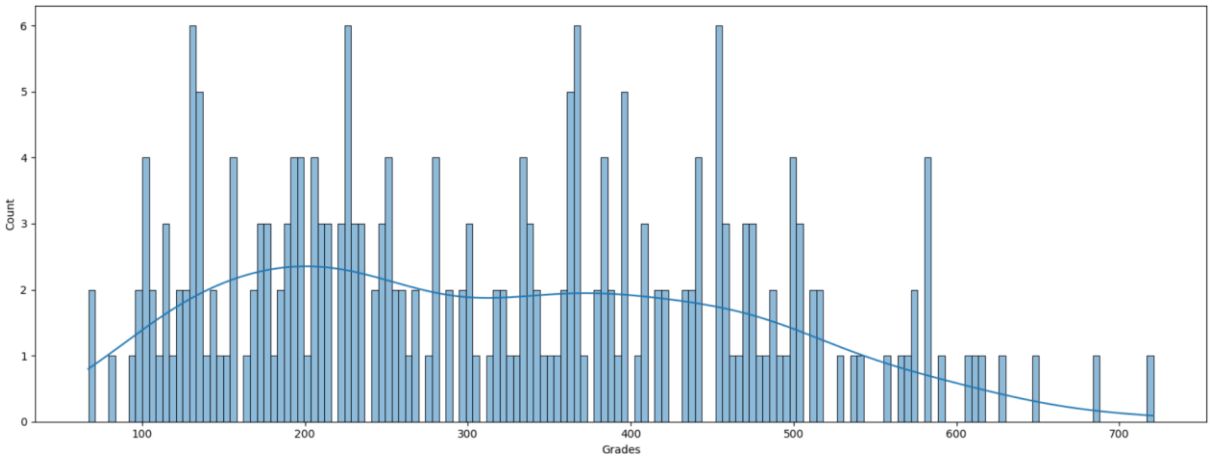

Figure 2 - Histogram of binning=$\sqrt{n}$ weighted degrees by count with distribution

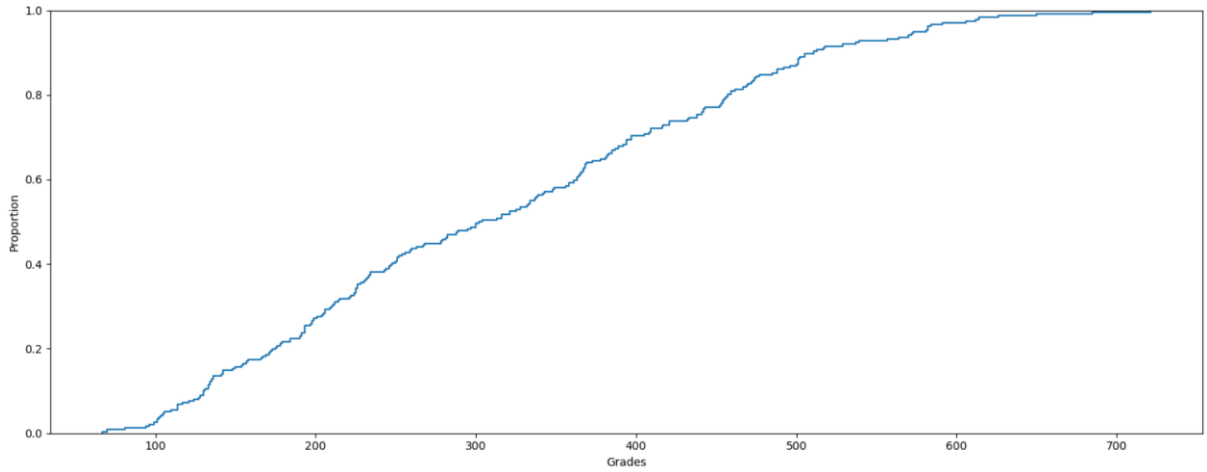

Figure 3 - Weighted degrees by count cumulative distribution function

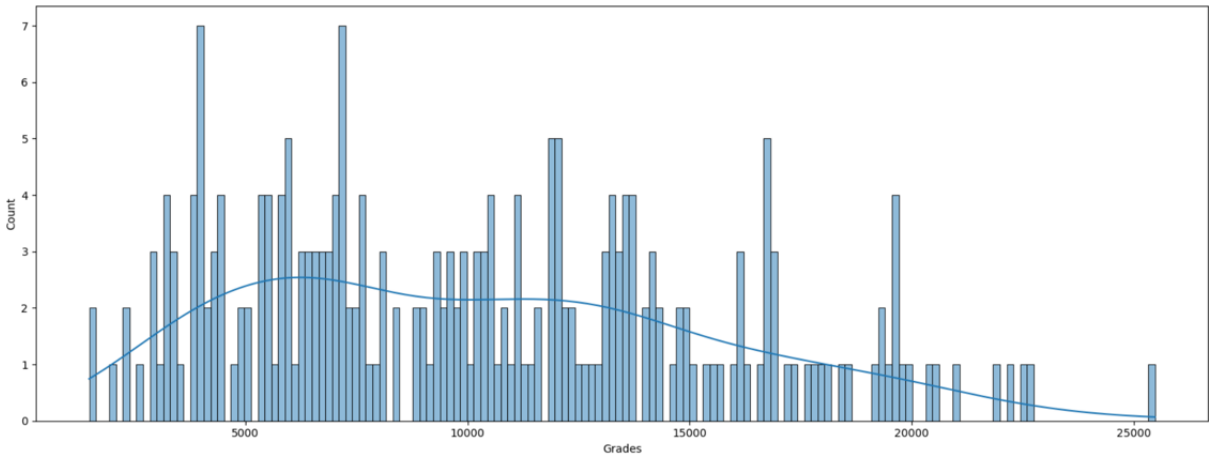

Figure 4 - Histogram of binning=$\sqrt{n}$ weighted degrees by duration with distribution

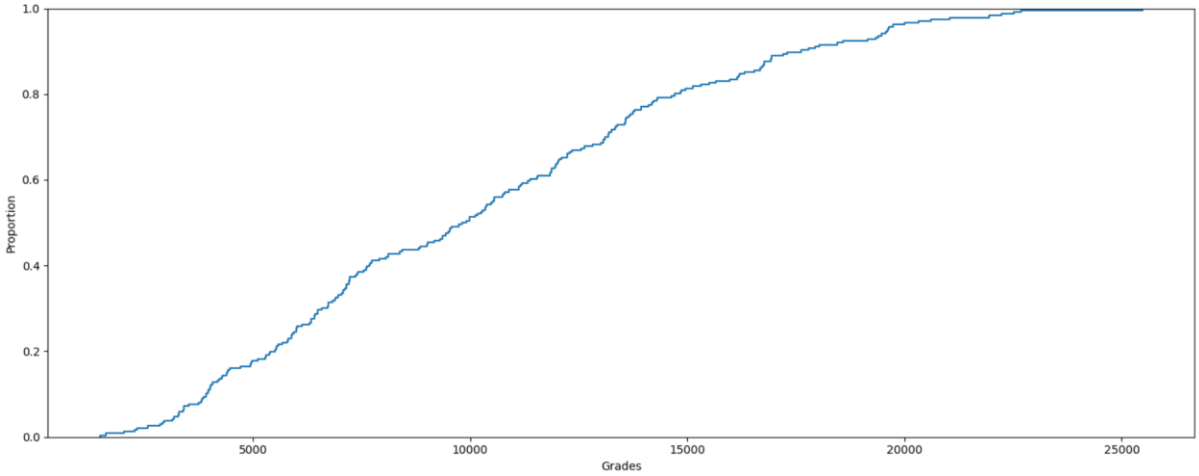

Figure 5 - Weighted degrees by duration cumulative distribution function

We identify id 1483 as the most isolated student having the lowest number of 67 interactions, 27 different connections, and the lowest cumulative time interactions of 1480 seconds.
The most connected student id 1551 has 455 interactions and 13700 seconds of cumulative time interactions, with his 98 distinct connections don't exceed Dunbar's number.
Endly, the most interactive student id 1695 has 721 interactions and 25480 seconds of cumulative time interactions, with 67 distinct connections.

Eigenvector Centrality

Looking for the most popular student we use the eigenvector centrality metric.
We select the most popular student (ID 1761) not already connected with ID 1483.

Student ID	Eigenvector centrality
1761	0.12513797172920654
1551	0.12288257130824916
1780	0.12215021259081889
1822	0.1191970158509773
1700	0.11877538874329832
1552	0.11422668541577652
1560	0.11313101998084242
1833	0.11075998812694977
1673	0.10908045809130279
1579	0.10802797740238816

Table 2 – The first ten nodes with eigenvector centrality (descending)

We want to connect the student IDs 1483 and 1761 to improve the lower student inclusion.

Betweenness Centrality

We have identified 11 shorter paths, with a length of two, that connect the two students: ['1483', '1498', '1761'], ['1483', '1522', '1761'], ['1483', '1837', '1761'], ['1483', '1727', '1761'], ['1483', '1585', '1761'], ['1483', '1500', '1761'], ['1483', '1501', '1761'], ['1483', '1443', '1761'], ['1483', '1562', '1761'], ['1483', '1426', '1761'], ['1483', '1452', '1761'].

To choose the best walk we use Betweenness centrality and select the first student in the walking path with the highest value (ID 1500) not already connected to the student with id 1483, but already connected to the student with id 1761 (Table 3).

ID	Betweenness Centrality
1551	0.01442753753914393
1890	0.014355310462114464
1552	0.013228395966694962
1916	0.012587128821227372
1761	0.012478531230613099
1911	0.011655308893925416
1708	0.01161330285219455
1780	0.011169336771375889
1700	0.010851085124673872
1613	0.010349996832618148
1673	0.009983603268689193
1851	0.009790377881789011
1560	0.00961616634075579
1579	0.009596573316647369
1833	0.009565259722307172
1628	0.0094982442508829
1889	0.00913619008143413
1533	0.008815296880302874
1477	0.008808771639816925
1434	0.008655608408545888
1822	0.008637971355759956
1457	0.008631666595276498
1787	0.008329048524755942
1666	0.008268265754603902
1650	0.008177762024682003
1665	0.008107045216494198
1567	0.008076795009649342
1895	0.007563644317057511
1847	0.007520076171951845
1480	0.007437976242473744
1625	0.007137055065473852
1684	0.007093108570775236
1500	0.007051624321271914
1802	0.007000875948405944
1866	0.006932494275125863

Table 3 – First 35 nodes with betweenness centrality (descending)

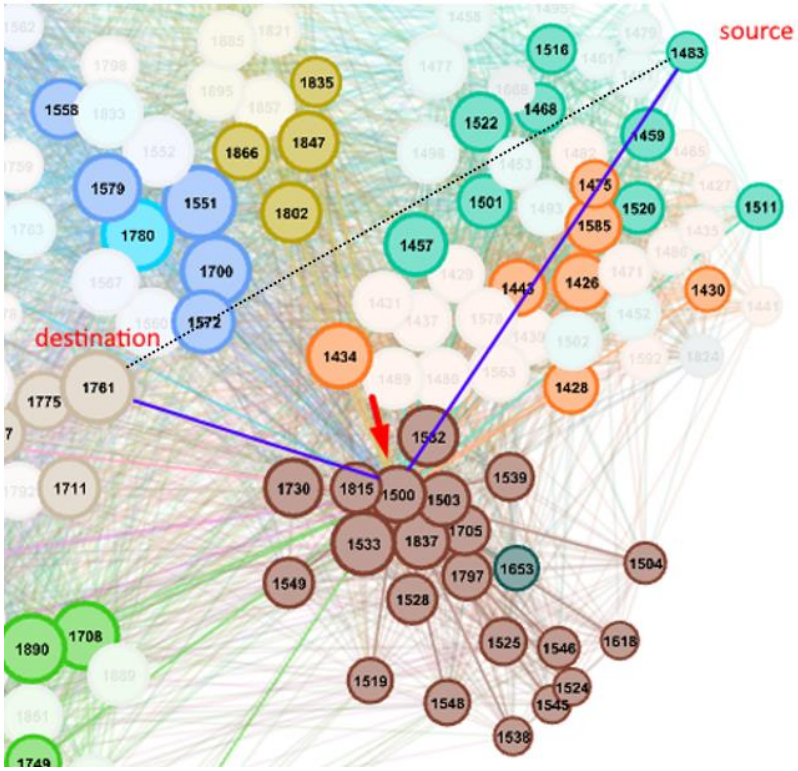

Figure 6 - Best walk from nodes Ids 1483 and 1761

The selected best path (Figure 6) is: ['1483', '1500', '1761'].
We ask at the student id 1500 to introduce the other two students' ids 1483 and 1761, so we add an undirected edge (as weight count=1 and duration=20) from node id 1483 and 1761.
In this exogenous inclusion operation, we wish to trigger an endogenous process to improve the inclusion of the isolated student.

Validation

Now we compare the original and modified network structures to verify if the changes respect the nature of the network and quantify the quality of the solution.
In the original network, the scale of degrees (Figure 7) goes from 18 to 98, student id 1483 has a degree of 18, and id 1761 has a degree of 97.

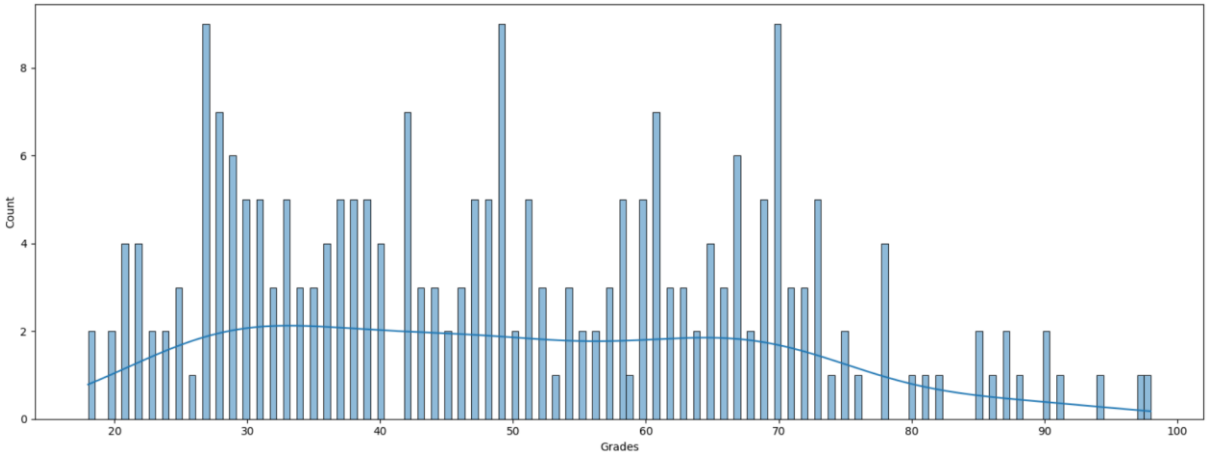

Figure 7 - Histogram of degrees with distribution

The assortativity coefficient r_1 of the original network is:

$$r_1 = 0.17292222281109404$$

The network is principal neutral (balanced) with a homophilic slightly shade for the degree, so connecting a low-degree student with a high-degree student is not perfectly keeping the nature of the network, but it is acceptable.
The assortativity coefficient r_2 of the modified network is:

$$r_2 = 0.1720638450712902$$

The difference is:

$$r_2 - r_1 = -0.0008583777398038506$$

After this change, the network goes in a heterophile direction of about 0.5%.
In order to reduce the impact by respecting the nature of the network, we could decide to connect the more isolated student with a popular student with minor degree centrality.
Now we inspect the k-core comparison, in Figure 8 the blue bars are related to the original network and the orange bars are related to the modified network.

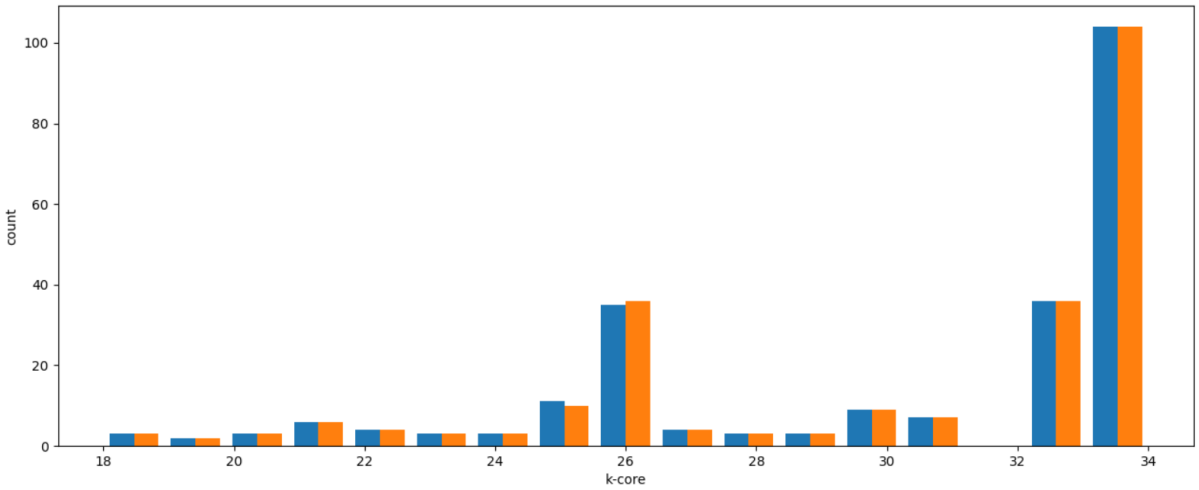

Figure 8 - k-core distribution on the original and the modified network

Changing the network, the student id 1483 goes up by one rank, from k-core of 26 to 27.
In order to detect potentially fragile students removing he from a maximal connected subgraph, we remove the high k-core and rank the remaining node by degree.
We remove the most reachable 104 students with k-core=34:
['1500', '1501', '1759', '1912', '1666', '1453', '1895', '1889', '1567', '1855', '1847', '1480', '1560', '1428', '1459', '1673', '1625', '1838', '1503', '1787', '1815', '1684', '1743', '1533', '1797', '1489', '1772', '1493', '1767', '1675', '1434', '1780', '1452', '1749', '1704', '1723', '1498', '1909', '1528', '1837', '1572', '1520', '1502', '1857', '1665', '1477', '1482', '1833', '1722', '1551', '1705', '1700', '1714', '1766', '1431', '1688', '1579', '1885', '1752', '1775', '1861', '1851', '1439', '1613', '1437', '1558', '1471', '1578', '1757', '1457', '1585', '1711', '1522', '1763', '1795', '1866', '1532', '1902', '1708', '1822', '1563', '1761', '1790', '1890', '1695', '1802', '1798', '1697', '1720', '1628', '1764', '1916', '1426', '1792', '1552', '1765', '1730', '1732', '1549', '1674', '1443', '1429', '1911', '1698']

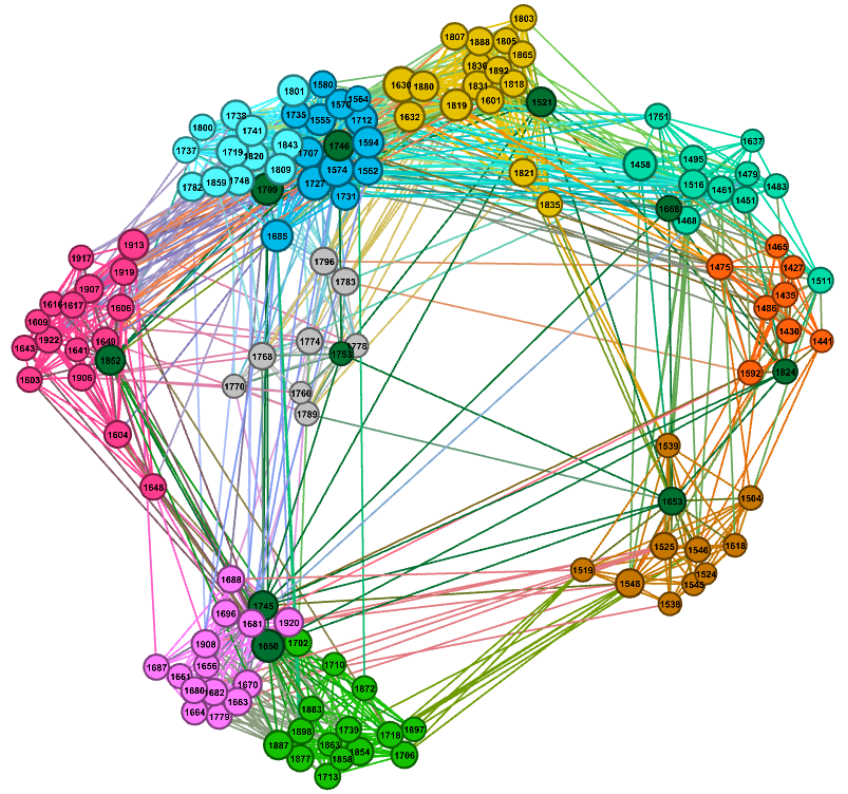

Figure 9 - Graph without nodes with k-core = 34

Ranking nodes by degree (Table 4) we found the nodes that can need an improvement in terms of inclusion in sub-network to which they are not already connected.

ID	Weighted degree centrality by count	Weighted degree centrality by duration
1465	14	320
1710	27	580
1483	32	740
1753	35	840
1427	47	1060

Table 4 - The first five nodes with weighted degree ordered by duration (ascending)

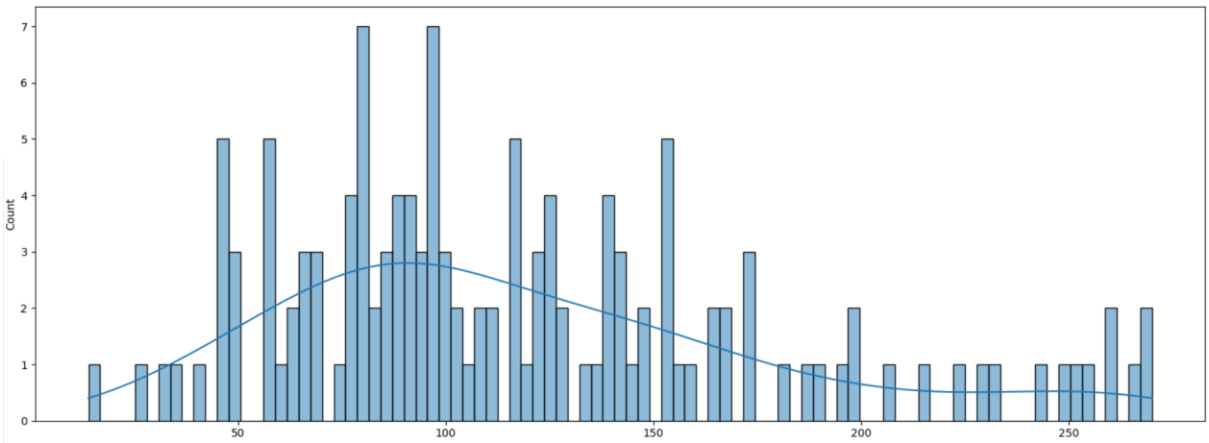

Figure 10 - Histogram of binning = √n weighted degrees by count with distribution without nodes k-core = 34

A wide application

After this first example, we apply the same methodology to a larger number of isolated students. We want to connect the first ten isolated students (Table 1) with the first ten popular students (Table 2).

To respect the original nature of the network, having the most isolated and most popular student list, we connect the most isolated with the lower popular student not already connected, the second one most isolated with the second one lower popular, and so on. Table 5 shows how we connect nodes selecting the shortest path passing from the bridge node with higher betweenness centrality (Table 3).

Isolated student	Popular student	Shortest paths	Bridge student
1483	1579	['1483', '1468', '1579'], ['1483', '1493', '1579'], ['1483', '1520', '1579'], ['1483', '1500', '1579'], ['1483', '1501', '1579'], ['1483', '1443', '1579'], ['1483', '1562', '1579'], ['1483', '1751', '1579'], **['1483', '1477', '1579']**, ['1483', '1459', '1579'].	1477
1753	1673	['1753', '1783', '1673'], ['1753', '1787', '1673'], ['1753', '1772', '1673'], ['1753', '1792', '1673'], ['1753', '1796', '1673'], ['1753', '1711', '1673'], ['1753', '1434', '1673'], ['1753', '1767', '1673'], ['1753', '1911', '1673'], ['1753', '1745', '1673'], ['1753', '1752', '1673'], **['1753', '1761', '1673']**, ['1753', '1764', '1673'].	1761
1441	1833	['1441', '1482', '1833'], ['1441', '1480', '1833'], ['1441', '1489', '1833'], ['1441', '1522', '1833'], ['1441', '1437', '1833'], ['1441', '1434', '1833'], ['1441', '1426', '1833'], ['1441', '1563', '1833'], **['1441', '1477', '1833']**.	1477
1521	1560	**['1521', '1552', '1560']**, ['1521', '1821', '1560'], ['1521', '1709', '1560'], ['1521', '1857', '1560'], ['1521', '1847', '1560'], ['1521', '1630', '1560'], ['1521', '1815', '1560'], ['1521', '1835', '1560'], ['1521', '1429', '1560'], ['1521', '1818', '1560'], ['1521', '1819', '1560'], ['1521', '1866', '1560'], ['1521', '1885', '1560'], ['1521', '1802', '1560'], ['1521', '1601', '1560'], ['1521', '1746', '1560'].	1552
1913	1552	['1913', '1555', '1552'], ['1913', '1558', '1552'], ['1913', '1709', '1552'], ['1913', '1727', '1552'], ['1913', '1580', '1552'], ['1913', '1752', '1552'], ['1913', '1885', '1552'], **['1913', '1761', '1552']**, ['1913', '1625', '1552'], ['1913', '1809', '1552'], ['1913', '1880', '1552'], ['1913', '1763', '1552'], ['1913', '1574', '1552'].	1761
1465	1700	['1465', '1759', '1700'], ['1465', '1482', '1700'], ['1465', '1480', '1700'], ['1465', '1486', '1700'], ['1465', '1712', '1700'], ['1465', '1437', '1700'], ['1465', '1434', '1700'], ['1465', '1439', '1700'],	1477

		['1465', '1458', '1700'], ['1465', '1429', '1700'], ['1465', '1866', '1700'], ['1465', '1426', '1700'], ['1465', '1578', '1700'], ['1465', '1443', '1700'], ['1465', '1563', '1700'], ['1465', '1457', '1700'], ['1465', '1471', '1700'], **['1465', '1477', '1700']**, ['1465', '1459', '1700'].	
1805	1822	['1805', '1775', '1822'], ['1805', '1630', '1822'], ['1805', '1843', '1822'], ['1805', '1802', '1822'], ['1805', '1847', '1822'], ['1805', '1741', '1822'], ['1805', '1763', '1822'], **['1805', '1761', '1822']**.	1761
1609	1780	['1609', '1774', '1780'], ['1609', '1604', '1780'], **['1609', '1911', '1780']**, ['1609', '1917', '1780'].	1911
1710	1551	['1710', '1780', '1551'], ['1710', '1674', '1551'], ['1710', '1673', '1551'], ['1710', '1704', '1551'], ['1710', '1822', '1551'], **['1710', '1890', '1551']**, ['1710', '1714', '1551'], ['1710', '1855', '1551'], ['1710', '1625', '1551'], ['1710', '1749', '1551'], ['1710', '1743', '1551'].	1890
1917	1761	['1917', '1772', '1761'], ['1917', '1759', '1761'], ['1917', '1630', '1761'], ['1917', '1911', '1761'], ['1917', '1913', '1761'], ['1917', '1859', '1761'], **['1917', '1916', '1761']**.	1916

Table 5 – massive nodes connections

Now we compare the original network with the modified network (adding edges between more isolated and most popular students).
The assortativity coefficient r_2 of the modified network is:

$$r_2 = 0.1663106890619654$$

The difference is:

$$r_2 - r_1 = -0.006611533749128651$$

After this change, the network goes in the heterophile direction of about 3.8% mitigated by pairing isolated/popular criteria.
Finally, we inspect the k-core comparison, in Figure 11 the blue bars are related to the original network and the orange bars are related to the modified network.

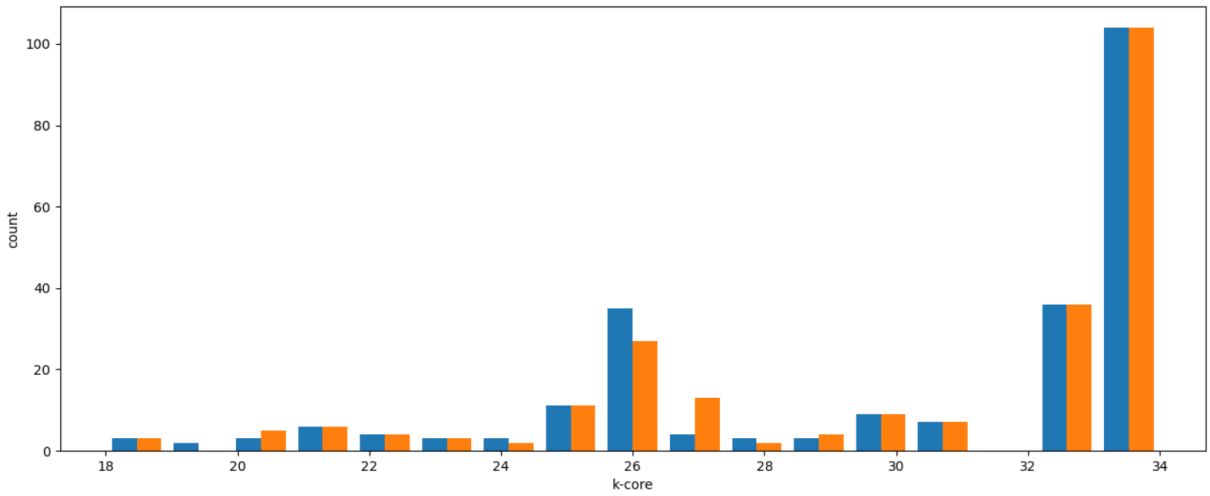

Figure 11 - k-core distribution on the original and the modified network

We conclude by looking at how these changes benefit the network (Table 6), raising 14 students by one point in terms of reachability.

Student ID	from k-core	to k-core
1917	19	20
1603	19	20
1805	24	25
1483	25	26
1461	26	27
1778	26	27
1753	26	27
1521	26	27
1745	26	27
1760	26	27
1479	26	27
1770	26	27
1836	26	27
1465	28	29

Table 6 - k-core changes

Critique

This article covers one improvable but practical approach using network science metrics to manage the lack of inclusion in students.
In data collection, RFID badges were a good low-budget solution to gather the data but need many recording days to avoid validity impact.
Having more budget, we can consider using a camera which removes the 20-second lower limit interaction, avoids lost and fictitious interactions, and using facial recognition, we can add the interaction quality in the edges.
The edges can be changed in directed edges giving information about which student is acting or reacting during an interaction.
For instance, one interaction between students A and B can have one good interaction with A acting for a cumulative time of 40 sec and B reacting for 20 sec friendly, and one bad interaction with B acting for 20 sec and B reacting for 5 sec annoyed.
Obviously, the use of cameras would generate privacy issues that should be managed.
In different contexts you can consider using weighted eigenvector centrality or weighted betweenness centrality, we didn't use the weighted version because we want to improve the inclusion of the isolated student with different new possibilities.
Betweenness centrality does not consider that bridges' connections are new for the isolated student, we could also compare the isolated students' structural equivalence with bridge students and perform an adjustment to the betweenness centrality.
Regarding which coupling to choose between the isolated student and a popular student, we can improve the association criteria to mitigate the impact of network changes that do not follow the nature of the network.
We can pre-determinate, using the assortativity coefficient formula, how much the network's nature will change adding an edge and by quantifying the k-core gain we can choose the best tradeoff for your network.

Before applying these techniques, it would be useful to map the students' personalities with a test or with the help of a sociologist or a psychologist, considering the personalities affinity and not only the degree.