FACE-TO-FACE INTERACTIONS PATTERNS IN A PRIMARY SCHOOL

Giulia Lumia, Marco D'Antoni

INTRODUCTION

This report is aimed to show the outcome of a network analysis carried out to identify and characterize social patterns among school children (6-12 y.o.).
This kind of outcome is useful not only from a social point of view, such as the study of friendships patterns, but also from an epidemiological one. Actually, the data were originally collected as a part of an epidemiological research [1], trying to identify situations at risk of infections transmission in the school environment.
They refer to interactions between couples of 228 children attending different school classes and are presented in the form of a weighted social network in which:

- each node is a student;
- each student is characterized by its gender, the class name and the class year;
- each edge is an interaction between two students;
- each interaction is weighted by its frequency, its duration and its average duration (in minutes).

The data raised two main research questions:

- Does exist a natural way to divide the network in groups according to the interaction duration?
- If yes, do the kids belonging to these groups have some discriminant features?

To answer these questions several steps were done:

1. MACROSCOPIC NETWORK DESCRIPTION;
   1.1 GRAPHICAL VISUALISATION;
   1.2 MACROSCOPIC FEATURES.
2. MESOSCOPIC NETWORK DESCRIPTION;
   2.1 COMMUNITIES DETECTION;
   2.2 COMPARISON BETWEEN COMMUNITIES DETECTION ALGORITHMS;
   2.3 COMMUNITIES CHARACTERIZATION.

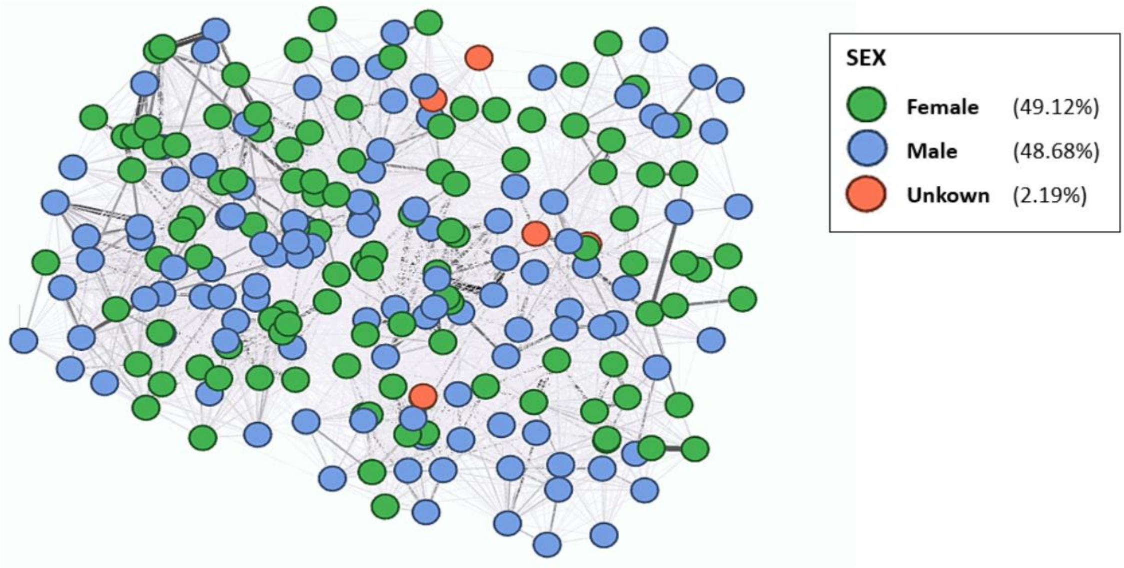
1. MACROSCOPIC NETWORK DESCRIPTION

A network can be described at different levels of detail: a macroscopic network description concerns all the features related to the network considered as a whole.
In this stage, a general description of the network was carried out in addition to a graphical representation of nodes' features.

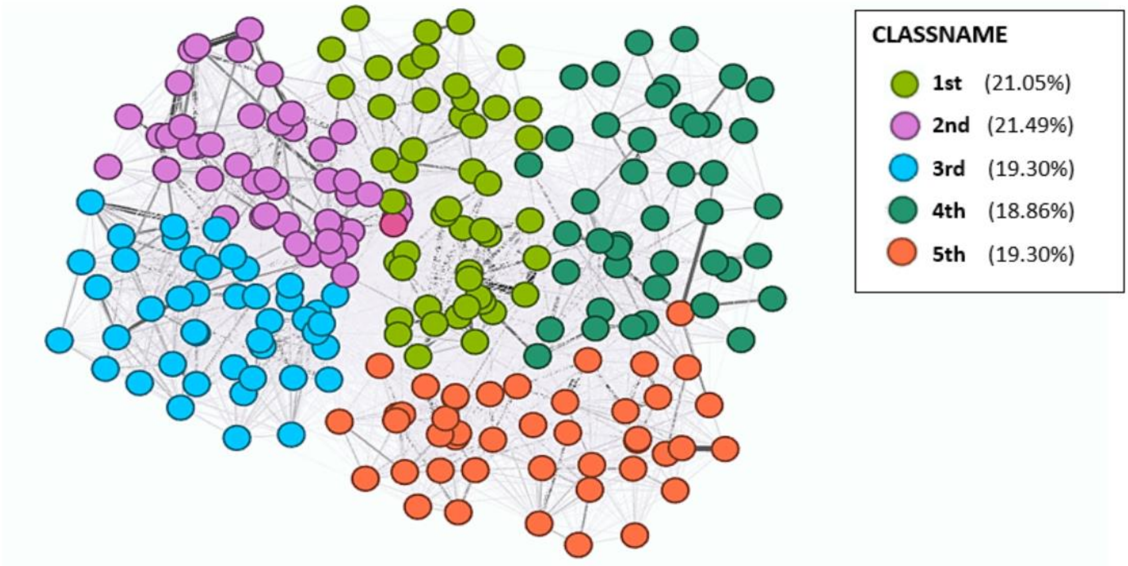
1.1 GRAPHICAL VISUALISATIONS

The following representations aim to offer a graphical visualisation of the network along with its features. They were made by the mean of a dedicated software [2], using an algorithm [3] by which linked nodes attract each other and non-linked nodes push apart, e.g. deskmates will be displayed closer than children who have never talked to each other. By looking at the GRAPH 1.1 it's clear that no evident concentration of nodes with the same sex exists. On the other hand, class name and class name and section lead to think that children having more or less the same age (the class name variable can be considered a proxy of the age) or attending the same class are most likely to interact. All the students' features are well balanced and represented in our dataset, with the exception of the unknown sex.

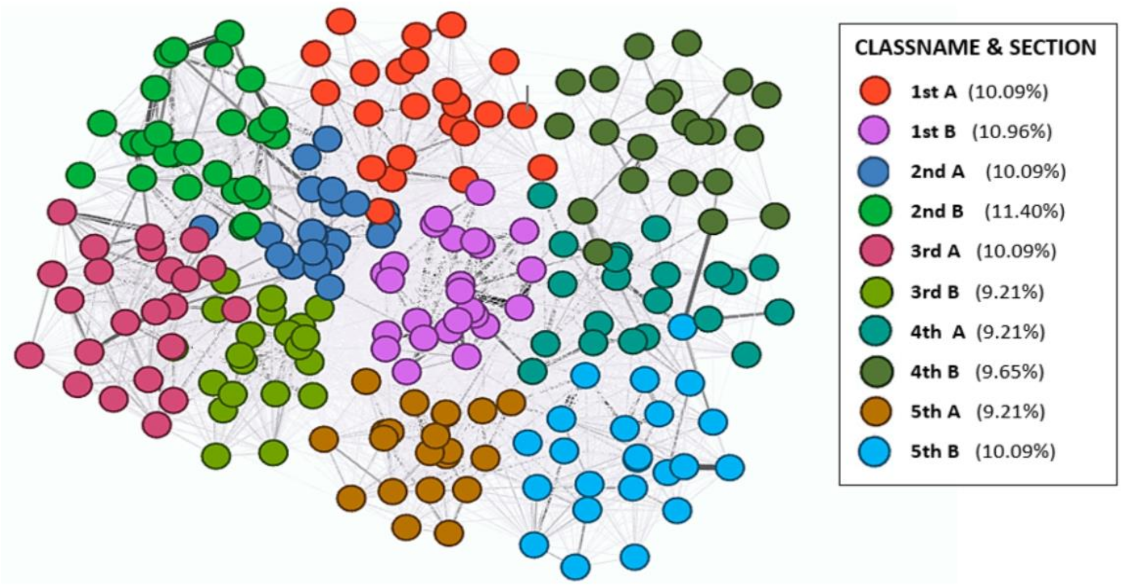
GRAPH 1.1: Network representation highlighting the students' sex.



GRAPH 1.2: Network representation highlighting the students' class name.



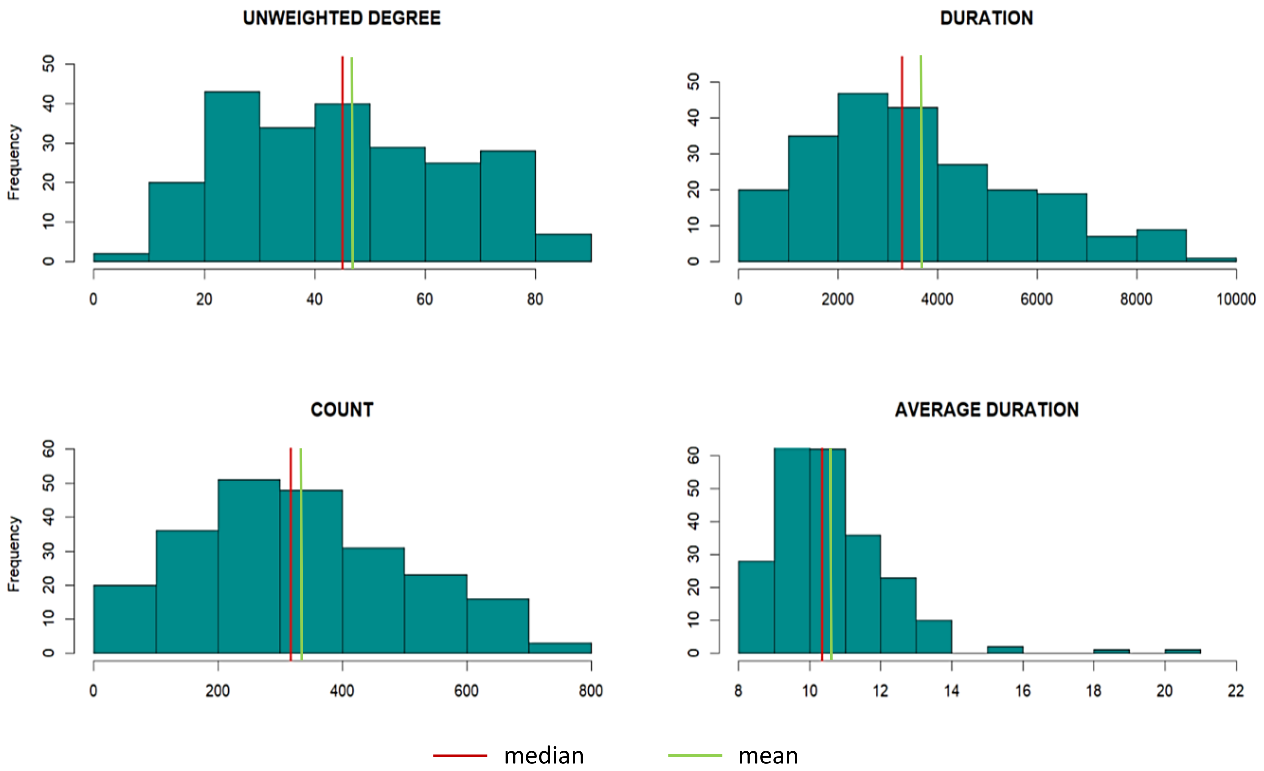
GRAPH 1.3: Network representation highlighting the students' class name and class section.



1.2 MACROSCOPIC FEATURES

According to the particular network type, the measures used to discover the main network features were generalised to their weighted networks form. In fact, most measures originally born to describe unweighted networks i.e. without taking into account the additional information related to the interactions among nodes.

GRAPH 1.4: Unweighted and weighted degree distributions



GRAPH 1.4 and TABLE 1.1 show the empirical distribution of some key descriptive measures, all slightly positively skewed. The unweighted degree of a node (a kid) is computed as the number of edges attached to it. Its distribution gives an intuition about the connectivity behaviour of the whole network: during the day half of the kids interacted at least with 45 different other kids.

However, the unweighted degree does not take into account the interaction's features, hence a new measure is needed.

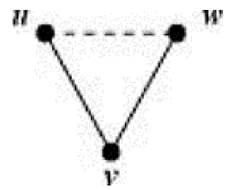
The weighted degree of a node regards not only the number of edges attached to it, but also its associated weights. It is computed as the sum of each edge attached to a node and weighted by a feature at a time. Therefore, three different weighted degree distributions were made:

- considering the DURATION of each iteration, it's clear that 3/4 of the kids interact daily for less than 4903 minutes, regardless of the number of interactions. In interpreting this version of the weighted degree, one should consider that weights refer to pairwise interactions so, for instance, if a kid interacts simultaneously with n other kids for a certain time t then this duration will be counted n times.
- The third plot refers to the weighted degree using the COUNT weight. It can be interpreted as the distribution of the number of different interactions, with respect to each child.
  Only for the 25% of the students the number of daily different iterations is higher than 443.
- The AVERAGE DURATION weighted degree distribution shows that only few kids interact on average for more than 14 minutes while most of them interact for an average time of 11.4 minutes.

TABLE 1.1: Unweighted and weighted degree distributions synthesis measures

	1st quartile	2nd quartile	3rd quartile	mean
UNWEIGHTED DEGREE	28	45	62	45.75
DURATION WEIGHTED DEGREE	2113.33	3280.00	4903.33	3603.04
COUNT WEIGHTED DEGREE	201.00	316.50	443.00	331.72
AVERAGE DURATION WEIGHTED DEGREE	9.50	10.35	11.40	10.62

Partial transitivity is a very useful property of real social networks: the fact that u knows v and v knows w doesn't guarantee that u knows w, but makes it much more likely. The friend of my friend is not necessarily my friend, but is far more likely to be my friend than some randomly chosen member of the population. In the weighted case, if both u-v and v-w are linked by a strong connection (i.e. the weight associated to their edge is high) then the chance of u-w being linked by a strong connection increases.



The transitivity can be quantified at two levels, a global and a local one for the single vertex, in both their unweighted and weighted forms. GRAPH 1.5 shows the distribution of the local clustering coefficients for the 228 nodes in the network while TABLE 1.2 highlight the average values of these distributions [4] in addition to the global clustering coefficients.

- The **unweighted global clustering coefficient** represents the number of closed triplets (or 3× triangles) over the total number of triplets (both open and closed). Since it is a proportion, it takes values between 0 and 1. In our network 47% of the triplets are closed;
- The **weighted global clustering coefficient** [5], takes into account the triplet value computed as the geometric mean of the weights involved in the triplet. It ranges between 0 and 1 and it's equal to 0.63.
- The **local clustering coefficient** of the generic node i represents the probability that a pair of i's friends are friends between them. The distribution of both its versions i.e. with or without considering ties' weights, are positively skewed (GRAPH 1.5). Putting all the ties on the same level, on average there's a probability of 0.56 that the neighbours of a randomly picked node are neighbours. This probability increases to 0.70 considering the duration of each interaction.

GRAPH 1.5: Unweighted and weighted local clustering coefficients distributions.

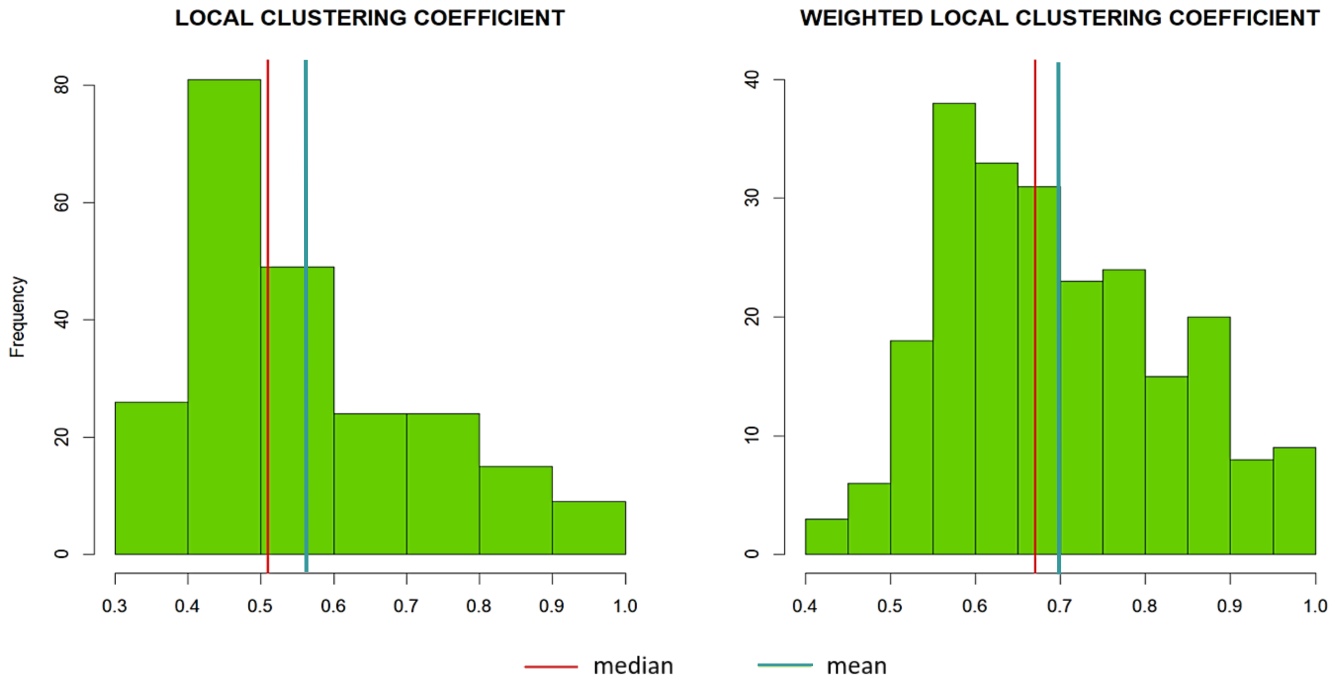


TABLE 1.2: Unweighted and weighted clustering coefficients

GLOBAL CLUSTERING COEFFICIENT	0.47
GLOBAL WEIGHTED CLUSTERING COEFFICIENT	0.63
TRIPLES	43731
TRIANGLES	278108
AVERAGE CLUSTERING COEFFICIENT	0.56
AVERAGE WEIGHTED CLUSTERING COEFFICIENT	0.70
MEDIAN CLUSTERING COEFFICIENT	0.51
MEDIAN WEIGHTED CLUSTERING COEFFICIENT	0.67

The weighted versions of these measures reduce to the unweighted one when the weights are all the same and can be seen as the answer to the question:
Is it more likely that a triangle will be formed by a triplet that shows strong links (in terms of duration)?
Since they are greater than their non-weighted versions, in our case, we can conclude that duration plays a role: the longer is the daily interaction between two children and another one in common, the greater are the chances that these two children will interact.

2. MESOSCOPIC NETWORK DESCRIPTION

The mesoscopic description of a network refers to the groups nodes tend to form within it.
As a result of the macroscopic network description (GRAPHS 1.1, 1.2), it is possible to notice that children inside the same classroom or attending the same school year tend to interact more.
A reasonable working question may be detecting the so called communities i.e. groups whose children interact more with the members of the same community rather than with those of the other ones.
The mesoscopic network description is made up of two main parts:

- community detection;
- community characterization.

2.1 COMMUNITIES DETECTION

Community detection is a tool often used to better understand the structure and the connectivity pattern of a network. It concerns the problem of finding the natural divisions of a network into groups of vertices such that there are many edges within groups and few edges between groups [6].
The most used measure to assess the quality of a partition into a certain number of communities is the **modularity**: a measure that quantifies how many edges lie within groups in our network relative to the number of such edges expected because of chance [7]. A good division of the given network into communities may be obtained by maximizing this quantity. It can be computed on both unweighted and weighted networks [8] so in our case the duration of each interaction was considered.
Since the number and the size of the communities is not fixed, the community detection problem turns out to be a NP problem, that can be solved suboptimally through an heuristic approach.
The algorithm used is the **simulated annealing algorithm** (SA) which evaluates a series of similar solutions (nearby) making use of a cooling function to decide whether to switch or not to a worse solution at a generic i^{th} iteration out of N, given a starting temperature T_0 and a final one T_f.

To make a decision about which cooling function use a comparison was made:

- EXPONENTIAL

$$T_i = T_0 * exp(-\beta * i) \qquad \beta = -ln(T_f/T_0)/N$$

- LINEAR

$$T_i = T_0 - (\eta * i) \qquad \eta = (T_0 - T_f)/N$$

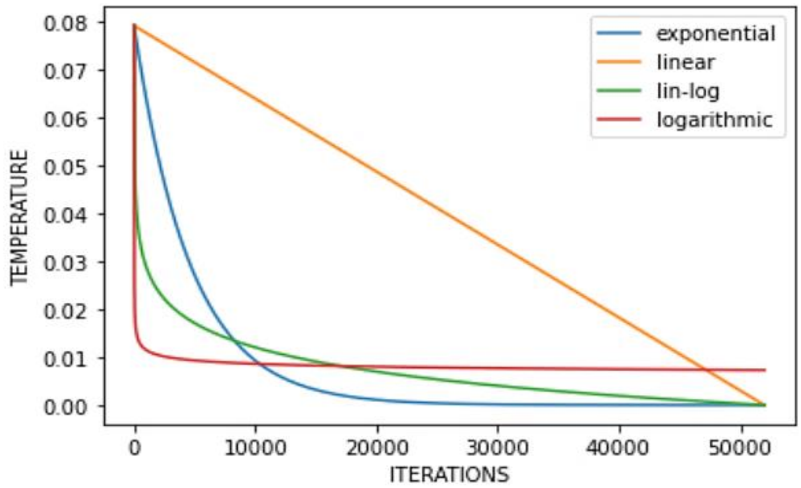
- LOG-LINEAR

$$T_i = T_0 - \eta * ln(i + 0.1) \qquad \eta = (T0 - Tf)/\ln(N + 0.1)$$

- LOGARITHMIC

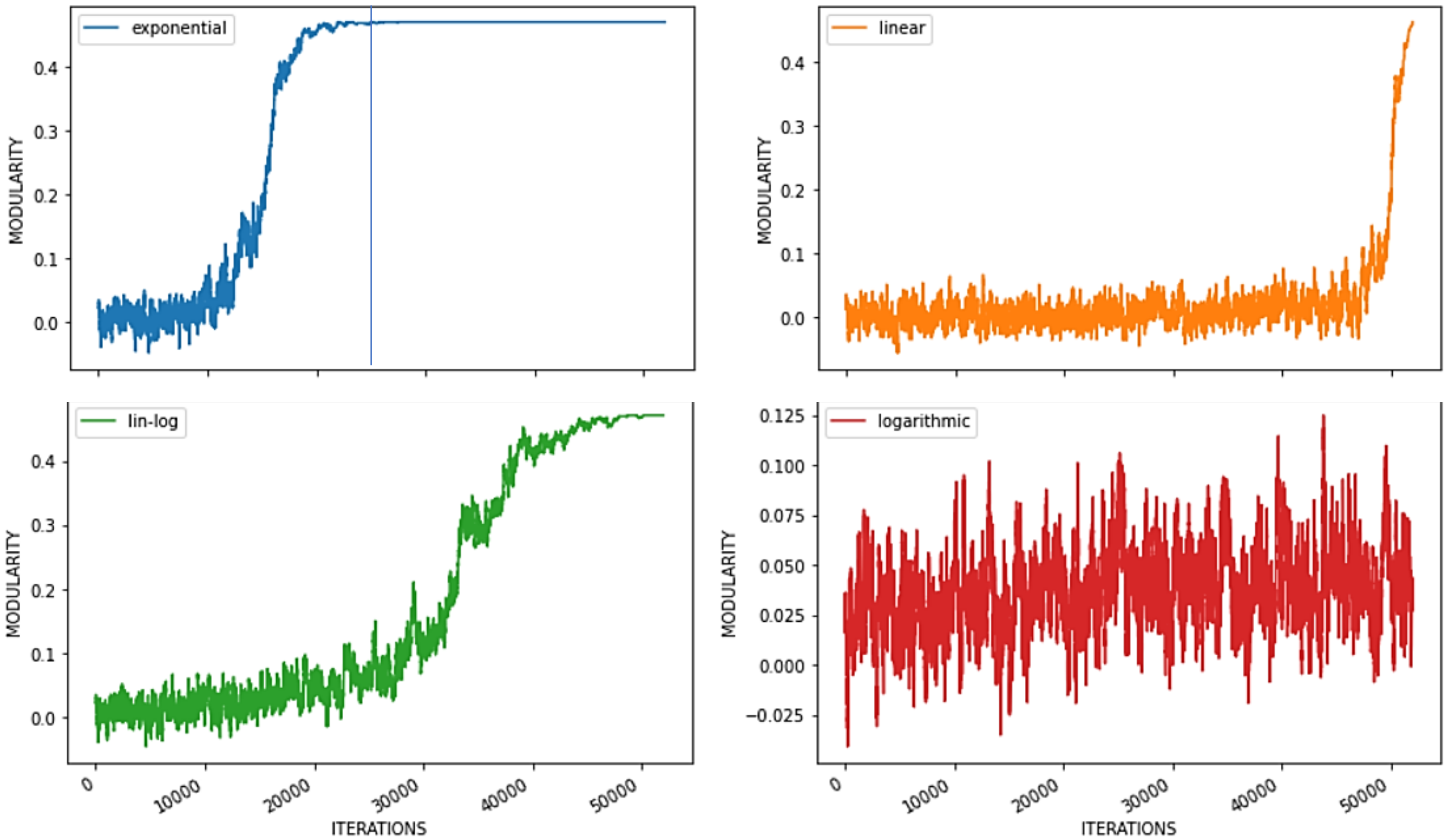
$$T_i = T_0 \,/\, \ln(i + e)$$

GRAPH 2.1: Cooling functions behaviour and speed.



The inspection of GRAPH 2.2 shows that with a constant number of iterations not all the cooling functions converge to a certain modularity value. The maximum modularity value (around 0.46) is reached by the first three methods (exponential, linear and log-linear) even if the number of iterations needed varies considerably.
The chosen method is the exponential one since it requires only 25000 iterations to converge.

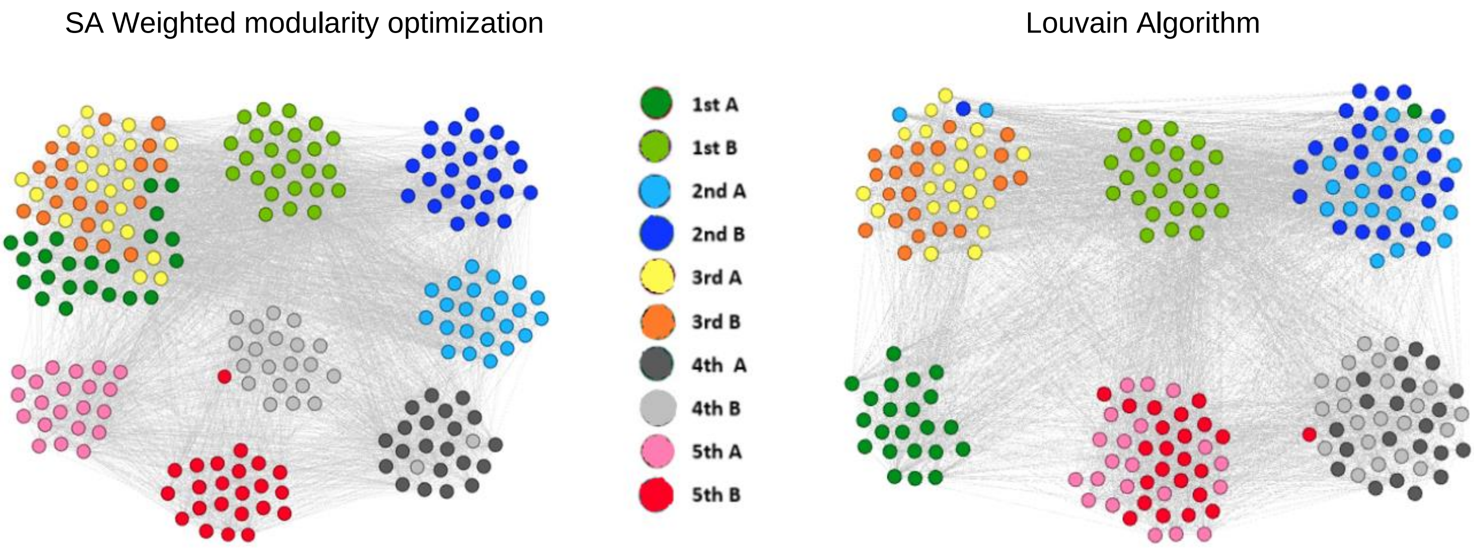
GRAPH 2.2: Modularity value of the solutions found using different cooling functions.



2.2. COMPARISON BETWEEN COMMUNITY DETECTION ALGORITHMS

Having specified the cooling schedule, both the number of communities and their composition were researched. The suboptimal solution found consists of 8 communities, reaching a modularity value equal to 0.6468. This result can be compared with that one obtained from an alternative algorithm [9]: the "Louvain Community Detection" algorithm, a greedy optimization algorithm also based on the maximization of the modularity score.
Its sub-optimal solution consists of 6 communities, with a modularity equal to 0.6581.

GRAPH 2.3: Comparison between the two communities detection algorithms.



GRAPH 2.3 shows that the 8 communities detected by our algorithm in most cases are made up of whole classes; in particular, one of them contains 3 whole classes (2 of which are of the same grade). Communities from the Louvain algorithm are instead mainly made up of children attending the same school year, grade. The only exception refers to the first grade, whose students are separated into two communities according to their specific class (1^{st} A and 1^{st} B).

2.3 COMMUNITIES CHARACTERIZATION

GRAPH 2.3 gives the intuition that children attending the same class or having more or less the same age (the grade is a proxy) clearly tend to interact longer among them.
If communities were generated by pure chance, they would have a relative frequency for a certain characteristic equal to the relative frequency of the same characteristic in the whole network.
As instance, given 10 classes with a relative frequency of 10% in the whole network, we would expect that a community of 10 students will be made up of a student from each class.

Through community characterization it is possible to test whether the percentages of children with a certain characteristic within a community and in the whole network are statistically different.
TABLE 2.1 shows the results of a communities characterization analysis carried out over the variables:

- CLASSNAME & SECTION;
- CLASS YEAR;
- SEX.

TABLE 2.1: Conditioned frequencies of each feature with respect to the community and relative frequency p-value.

COMMUNITY	CLASSNAME & SECTION			CLASS YEAR			SEX		
	value	freq.	p-value	value	freq.	p-value	value	freq.	p-value
1	1B	1,00	0.000	1	1,00	0.000	F	0,52	0.303
							M	0,48	0.556
2	1A	0,34	0.000	1	0,34	0.001	F	0,46	0.341
	3A	0,34	0.000	3	0,66	0.000	M	0,51	0.292
	3B	0,31	0.000				Unknown	0,03	0.153
3	2B	1,00	0.000	2	1,00	0.000	F	0,58	0.128
							M	0,42	0.315
4	2A	1,00	0.000	2	1,00	0.000	F	0,61	0.079
							M	0,83	0.228
5	4A	0,91	0.000	4	1,00	0.000	F	0,35	0.109
							M	0,57	0.156
	4B	0,09	**0.613**				Unknown	0,09	**0.008**
6	4B	0,95	0.000	4	0,95	0.000	F	0,52	0.294
	5B	0,05	**0.347**	5	0,05	**0.058**	M	0,48	0.551
7	5B	1,00	0.000	5	1,00	0.000	F	0,45	0.446
							M	0,55	0.211
8	5A	1,00	0.000	5	1,00	0.000	F	0,48	0.534
							M	0,48	0.551
							Unknown	0,05	0.068

Classname&section and class year are over represented into the communities:
the p-values relative to the comparison between the relative frequency within each community and in the whole network are essentially zero. The only exception refers to classes with relative frequency of 9% (4B in community 5) and 5% (5B in community 6).

On the other hand, the gender variable doesn't show over/under representation.
That's because each community have relative frequencies similar to those in the entire network (GRAPH 1.1).

CONCLUSIONS

To sum up we can say that the longest interactions happen among children inside the same classroom or approximately having the same age, irrespective of the students' sex.
This information turns out to be particularly useful e.g. in a pandemic contest in which restrictive measures are need for the safety of the students inside a school.

REFERENCES

[1]
High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School
(J.Stehlé, N.Voirin, A.Barrat, C.Cattuto, L.Isella, J.Pinton, M.Quaggiotto, W.Van den Broeck, C.Régis, B.Lina, P.Vanhems)

[2]
Gephi is a free and open-source software for the visualization and the exploration of all kinds of graphs and networks.

[3]
The algorithm used is called "Force Atlas 2".

[4]
Collective dynamics of 'small-world' networks
D. J. Watts & Steven Strogatz (June 1998).

[5]
Clustering in weighted networks
(T.Opsahl, P.Panzarasa)

[6]
Networks, an introduction (chapter 11.2.1)
(M.E.J. Newman)

[7]
Networks, an introduction (chapter 11.6)
(M.E.J. Newman)

[8]
Analysis of weighted networks (2004)
(M.E.J. Newman)

[9]
The Louvain algorithm is contained in the Python library “Networkx”.