- **BASIC VARIABLES**

  Answer: It encodes whether the person answered the survey.
  1 - The person answered the survey.
  0 - The person did not answer the survey.

  Course: It encodes the academic course the student is enrolled in.
  It ranges from 1 to 4.

  Section: It encodes the itinerary the student is enrolled in.
  1 - The student attends classes both in Spanish and English.
  0 - The student attends classes in Spanish.

  Group: Courses are divided in groups. Students belonging to the same group attend, in general, the same classes together.
  This variable is a 3-digit number. The first digit encodes the course displayed in the column *Course*, and the other two digits specify the group.

  Sex: It encodes the sex of the student.
  1 - Male
  0 - Female

  ID: An integer number that uniquely identifies the node.

- **DEGREE VARIABLES**

  in_degree_1p: Number of **+1 (good relationship)** links that are directed **towards the node**.

  in_degree_2p: Number of **+2 (very good relationship)** links that are directed **towards the node**.

  in_degree_1n: Number of **-1 (bad relationship)** links that are directed **towards the node**.

  in_degree_2n: Number of **-2 (very bad relationship)** links that are directed **towards the node**.

  out_degree_1p: Number of **+1 (good relationship)** links that are directed **from the node outwards**.

  out_degree_2p: Number of **+2 (very good relationship)** links that are directed **from the node outwards**.

out_degree_1n: Number of **-1 (bad relationship)** links that are directed **from the node outwards**.

out_degree_2n: Number of **-2 (very bad relationship)** links that are directed **from the node outwards**.

Nodes which did not answer the survey (Answer = 0) have all the *out_degree_XX* variables empty (to differentiate them from those which actually have 0 outgoing links).

- **RECIPROCITY VARIABLES**

  There are **35 RECIPROCITY VARIABLES**.

  All of them have the same structure: $\underline{\pm X \pm Y}$. Each variable represents, for the specific node, the number of links the node is involved in for which our node has a $\pm X$ outgoing link, and the other node responds with a $\pm Y$ link.

  For instance, if for node 137 we have a value of 5 in variable +2+1, it means that, from all the outgoing +2 links of node 137, 5 of them have been reciprocated with a +1 incoming link.

  We give integer numbers, instead of densities, to be able two define two types of densities: outgoing densities (from all the outgoing links of a certain type associated with a node, which proportion is reciprocated by a certain type of incoming link by the rest of nodes) and incoming densities (from all the incoming links of a certain type associated with a node, which proportion is reciprocated by the node with a certain type of outgoing link).

  Some of the variables contain a -3 value. It represents the situation in which the node did not answer the survey. For instance, variable -3+2 represents the number of +2 links received by a node that did not respond the survey. Of course, if the node answered the survey, this value is 0. Similarly, the variable +2-3 represents the number of +2 links directed from our node towards a node that did not answer the survey. This variables exist to be able to differentiate the situation in which a link is absent because a person did not respond the survey from the situation in which a link is absent because on purpose a certain person chose not to put a link there (i. e., to differentiate a +2-3 link from a +2+0 link).

- **OUTLIER VARIABLES**

  outlier_type: A node can be identified as an outlier if it displays a very large value in one of the 4 variables corresponding to the out degree. Specifically, we consider a node as an outlier if:
  It gives **more than 15 very bad outgoing relationships**.

It gives **more than 15 bad outgoing relationships**.
It gives **more than 50 good outgoing relationships**.
It gives **more than 30 very good outgoing relationships**.

The outlier_type variable is a 4 digit number that represents whether a certain node is an outlier according to one (or more) of its 4 types of out degree. Digits are associated to -2, -1, +1 and +2 out degrees, in this order (-2)(-1)(+1)(+2). Thus, the number 0000 means that a node is not an outlier. The number 0010 means that a node is an outlier due to its +1 out degree. The number 0101 means that a node is an outlier due to its -1 and +2 out degrees. And so on, and so forth.

We consider that nodes can only be outliers due to their out degrees. A very large out degree could be a proxy of a student not taking the survey seriously, answering at random or having very loose definitions of good and bad relationships. And network analyses are very sensitive to hubs, i. e. nodes with a very large number of connections. Thus, their presence can distort all analyses.

We consider that nodes cannot be outliers due to their in degrees, independently from how high the in degree is.

outlier: Apart from the outlier_type, we want to be able to know if a node is an outlier or not, independently from the specific out degree that tells us it is indeed an outlier. Thus, we define a second variable that identifies a node as an outlier dichotomically:
1 - The node is an outlier.
0 - The node is not an outlier

- **CENTRALITY VARIABLES**

  The network contains directed links with weights ranging from -2 to +2. Thus, the nature of the network conditions the way in which we are able to define centrality metrics. Most centrality metrics were developed for unweighted and undirected networks. Most of them have been extended to directed and weighted networks as well, but there are some caveats we need to take into account:

  First, these extensions are not prepared to handle together positive and negative links. Thus, we need to separate our network in two networks, and explore separately the central position of nodes in the positive and in the negative network. Conceptually, being a central node in the negative network is not the same as being central in the positive network, hence this separation makes theoretical sense as well.

  Besides, the metrics we will consider have extensions to weighted networks, but they interpret weights as distances. In our case, weights represent relationship intensities. It is not clear that we can draw a parallelism between both interpretations. We could interpret a more intense relationship as *emotionally closer* in terms of distance. But in this case, we need to change the weights of the network, because intensity and emotional distance are inversely related. Anyway, even if we consider this intepretation, it is not straigthforward to justify that it makes sense to include this information to compute centralities. For instance, let us consider the positive network, with +1 and +2 weights. We reinterpret weights as distances, such that a +2 weight becomes a distance 1, and a +1 weight becomes a distance 2. This means that, when computing centralities, walking through a very good relationship is equivalent as walking through two good relationships. For some centrality metrics this could make conceptual sense, but it is not clear.

So, taking all these caveats into account, for each centrality metric, we will compute three values: First, for the negative network, we will consider that both -1 and -2 links have the same weights, and interpret them as equivalent. We do this because previous analyses have shown that -1 and -2 links are quite similar (students are not very good at determining the intensity of bad relationships). Thus, we will compute the value of the centrality in the negative network in which all negative links are considered equivalent independently of the weights. Then, for the positive network we will compute two values of each centrality. The first one will be computed using the interpretation of emotional intensity as emotional proximity, thus considering both +1 and +2 weights. We refer to this one as type I. The second one will be computed using only +2 links. We refer to this one as type II. We will not compute centralities using +2 and +1 links as equivalent because the behavior for both types of links is quite different and it would only provide a noisy result.

Furthermore, due to the directed nature of the network, some centrality metrics can be defined either on incoming links, or outgoing links. For instance, consider the Closeness Centrality. In an undirected network, it measure how close a node is to the rest of the network. However, in a directed network you could measure how close a node is from the rest of the network (out-degree), or how close the rest of the network is to the node (in-degree). So, for such a metric we can define the *in* centrality and the *out* centrality.

For clarity of notation, let us consider an example, the Closeness Centrality. For the Closeness Centrality we will have 6 values for each node:

CC+I(in) - Closenesss Centrality computed for the positive network of type I (considering +1 and +2 links, interpreting them as distances) for the in-degree (how close the network is to the node).
CC+II(in) - Closenesss Centrality computed for the positive network of type II (considering only +2 links) for the in-degree (how close the network is to the node).
CC-(in) - Closenesss Centrality computed for the negative network (considering only -1 and -2 links, interpreting them equivalently) for the in-degree (how close the network is to the node).
CC+I(out) - Closenesss Centrality computed for the positive network of type I (considering +1 and +2 links, interpreting them as distances) for the out-degree (how close the node is to the network).

CC+II(out) - Closenesss Centrality computed for the positive network of type II (considering only +2 links) for the out-degree (how close the node is to the network).
CC-(out) - Closenesss Centrality computed for the negative network (considering only -1 and -2 links, interpreting them equivalently) for the out-degree (how close the node is to the network).

For a metric like the Betweenness Centrality, for which there is no difference between in and out interpretations, we will only have three values, one for each type of network considered.

Centrality metrics considered:

Betweenness Centrality (BC) - It is the sum of all shortest paths between every pair of nodes that pass through the node divided by the total number of all shortest paths between every pair of nodes. It has no in/out interpretation.

Closeness Centrality (CC) - It is the inverse of the average shortest path distance to the node over all n-1 reachable nodes. It has in/out interpretation.

Eigenvector Centrality (EC) - It is a centrality metric that takes into account the centrality of predecessor nodes (incoming link) to compute the centrality of a certain node. In simple terms, it measures how central a node is according to how central its neighbors are. It has in/out interpretation.

PageRank Centrality (PC) - It is a metric with a similar interpretation to the Eigenvector Centrality. It has in/out interpretation.

- **STRUCTURAL HOLES**

  We have introduced some metrics to measure concepts related to Structural Holes at a node level. Structural Holes convey the information about how an individual in a network can access information from other nodes, how many of these sources are not redundant and how this node occupies a broker role in the network. In simple terms, a Structural Hole is defined as a gap between two individuals who have complementary sources to information. The guiding idea is that individuals occupying these gaps hold certain positional advantages/disadvantages from how they are embedded in the network.

  These concepts have not been explored extensively in weighted and directed networks, and variables associated to Structural Holes may be ill behaved depending on how the network is structured and defined (and what it represents). Empty values in the DataFrame provided refer to these cases. Anyway, as in the case of centrality metrics, whenever is possible we define up to 6 values for each metric (positive type I, positive type II and negative, with in and out interpretations for each of them).

Structural Holes' variables considered:

Constraint (C) - It is a measure of the extent to which the node under consideration is invested in those nodes that are themselves invested in the neighbors of the node under consideration. It has in/out interpretation.

Effective Size (ES) - A person's ego network has redundancy to the extent that her contacts are connected to each other as well. The non-redundant part of a person's relationships is the effective size of her ego network. It has no in/out interpretation.

Efficiency (E) - It is the Effective Size of a node normalized with its degree. For the directed network we consider the degree as the number of unique nodes to which the node under consideration is connected, either with in or out links (this interpretation seemed suitable for metrics that measure the extent to which a certain node has access to the information flows of the network).

- **TRIANGLES - TRANSITIVITY AND BALANCE**

  Finally, we define a series of metrics that intend to measure the concepts of transitivity and balance at the level of nodes. Transitivity is the tendency of a network to create triangular structures A->B B->C A->C where arrows represent positive links. Balance is the tendency of a network to create structures with an even number of negative links.

  Nonetheless, there are important caveats to take into consideration. First, there is no single metric that integrates both concepts. They have been explored separately in the literature, but I believe that creating separate metrics for both is a mistake because they are interrelated, follow similar rationales and affect simultaneously the formation of triangles. Furthermore, these concepts have not been explored properly in weighted directed networks.

  As a result, the metrics used are defined by myself and are still in an experimental phase to see whether they represent what we intend them to represent.

  Without going into much detail, I define two metrics for each node: the level of *pleasure* a node experiences due to transitive and balanced structures ($TB+$), and the level of *tension* a node experiences due to intransitive and imbalanced structures ($TB-$). Besides, I count how many triads produce pleasure ($r+$) and tension ($r-$). With these two metrics I define the average level of pleasure and tension ($tb+$ and $tb-$) per triad for each node. Finally, I count how many triads are *vacuously transitive* for each node $r0$, defined as triads that produce no pleasure nor tension.

  These levels of tension and pleasure are defined according to whether triads a node belongs to conform to the traditional definitions of balance and transitivity. Pleasure is obtained when the triad conforms to these definitions. Tension is increased when they violate these definitions, and the amount of tension is computed according to how much balance or transitivity would exist in the triad if it was transitive or balanced.

As before, we compute each metric in two different networks: type I metrics (for instance, $TB+I$) are computed using the version of the network in which we have +2. +1 and -1 links. Similarly, type II metrics are computed using the version of the network in which we have only +2 and -1 links.

As we mentioned, these metrics are still experimental, and it may happen they do not represent what we intend to or with the strength we want to, so one must take this into account in the interpretation of the results. This study may even serve to confirm they represent what they should.