Network constraints on the mixing patterns of binary node metadata

Matteo Cinelli,1 Leto Peel,2 Antonio Iovanella,3 and Jean-Charles Delvenne2,4

1Applico Lab, CNR-ISC, 00185, Rome, Italy
2ICTEAM, Université catholique de Louvain, Louvain-la-Neuve, Belgium
3University of Rome Tor Vergata, Via del Politecnico 1, Rome, Italy
4CORE, Université catholique de Louvain, Louvain-la-Neuve, Belgium

We consider the network constraints on the bounds of the assortativity coefficient, which measures the tendency of nodes with the same attribute values to be interconnected. The assortativity coefficient is the Pearson’s correlation coefficient of node attribute values across network edges and ranges between -1 and 1. We focus here on the assortativity of binary node attributes and show that properties of the network, such as degree distribution and the number of nodes with each attribute value place constraints upon the attainable values of the assortativity coefficient. We explore the assortativity in three different spaces, that is, ensembles of graph configurations and node-attribute assignments that are valid for a given set of network constraints. We provide means for obtaining bounds on the extremal values of assortativity for each of these spaces. Finally, we demonstrate that under certain conditions the network constraints severely limit the maximum and minimum values of assortativity, which may present issues in how we interpret the assortativity coefficient.

I. INTRODUCTION

Assortative mixing is the tendency of nodes with similar attribute values (also referred to as node metadata) to be connected to each other in a network. For example, in a social network node metadata might include age or gender. Assortativity with respect to a particular attribute is often measured using Newman’s assortativity coefficient [1], which is the network analogue of Pearson’s correlation. Just as correlation plays an important role in identifying relationships between pairs of variables, assortativity plays a fundamental role in understanding how a network is organised with respect to a given attribute of the nodes.

Similar to Pearson’s correlation, assortativity is a normalised measure that lies in the range $r \in [-1, 1]$. However, the structure of the network can further constrain the range of assortativity and often makes these maximum and minimum values unattainable [2, 3]. This issue arises because of the dependence between “samples” of edges and the metadata of the nodes they are incident upon. Consider that when we calculate the correlation between variables $x$ and $y$, each sample pair $(x_t, y_t)$ is assumed to be sampled independently. For assortativity of node metadata, this assumption no longer holds. Each sample represents the pair of metadata values of nodes connected by an edge. Therefore if node $i$ is incident upon $k_i$ edges, then this node’s metadata value $c_i$ affects not just a single edge, but all $k_i$ edges it is involved in. Figure 1(left) illustrates the dependencies invoked on binary variables $x$ and $y$ once we consider the underlying network structure. These dependencies introduce issues when we try to interpret values of assortativity, as $r$ is no longer the proportion of the possible extremal values and therefore we face difficulties in determining whether or not the network is highly assortative (or disassortative) with respect to a given attribute.

Here we focus on the special case in which the graph is undirected and unweighted and the node metadata are restricted to binary values, e.g., gender of actors in social networks, which allow us to study important phenomena such as gender homophily [4] and related perception biases in social networks [5, 6]. We provide methods to calculate bounds on the permissible range of assortativity under the assumption that specific properties of the network are fixed. We focus on two types of properties: those relating to the network structure (specific degree sequence or specific graph topology) and those relating to the node metadata (proportion of nodes per category or specific assignment of nodes to categories). Considering all possible combinations of network and metadata properties to keep fixed provides us with three different spaces of configurations (omitting the fourth combination as it corresponds to just the single configuration that we observe):

1. the metadata-graph space (mgs) – the ensemble of configurations with a given degree sequence and proportion of metadata labels
2. the graph space (gs) – the ensemble of configurations with a given degree sequence and specific node metadata assignment
3. the metadata space (ms) – the ensemble of configurations with a specific topology and proportion of metadata labels.

In the metadata-graph space the range of assortativity can be explored by computing $r$ over the set of all possible graphs with the observed degree sequence (the graph space) combined with set of all possible permutations of the
metadata vector (the metadata space), i.e., a vector $c$ in which each entry $c_i$ represents the metadata value of node $i \in \{1, ..., n\}$ in the network. For this space, we present combinatorial bounds that represent the largest possible range of assortativity values since it contains both the graph space and the metadata space. Thus, bounding assortativity in metadata-graph space means bounding assortativity with respect to all the possible values it can assume within the other spaces. Similarly for the graph space, we present combinatorial bounds for the range of assortativity, $r$, for all the possible configurations of the observed degree sequence, but this time with the metadata vector fixed.

In what follows, we will see that the relationship between the metadata-graph space and the graph space is relatively straightforward. Also, both the metadata space and the graph space are subsets of the metadata-graph space. However, the relationship between the graph space and the metadata space in terms of assortativity is somewhat more nuanced. Depending on the topology and the node metadata vector, the assortativity range in the graph space can be either narrower or wider than the metadata space. The assortativity range of the metadata space is harder to define combinatorially, but can be explored via a complete enumeration (when computationally feasible) of all possible permutations of the metadata vector. However, for larger graphs we must resort to heuristic methods. Figure 1 illustrates qualitatively the relationship of the bounds of each of these spaces.

By exploring the bounds using both combinatorial and empirical methods, we demonstrate that these bounds can be substantially far from $-1$ and $1$ in each of these three spaces. We reinforce these results by demonstrating that for some real-world networks the full range of assortativity is not possible for any of these three spaces. Such evidence may provide some insights about the interpretation of common configurations as well as boundary ones that, without such a knowledge, would be misinterpreted by being considered less significant than they are.

II. ASSORTATIVE MIXING OF BINARY METADATA

Newman’s assortativity, $r$, for binary node metadata is equivalent to Pearson’s correlation for binary variables, also known as the $\phi$ coefficient [9]. Pearson’s correlation is a fundamental statistic used to identify linear relationships between variables. However, despite the widespread use of Pearson’s correlation, it is not without its limitations e.g., the ambiguity of interpreting specific coefficient values [10]. Given the relationship between Newman’s assortativity and Pearson’s correlation it is unsurprising that the assortativity coefficient also suffers from similar issues of interpretability [11]. Here we consider a further issue that affects the interpretation of assortativity – the extremal values and Pearson’s correlation it is unsurprising that the assortativity coefficient also suffers from similar issues of interpretability [10]. Given the relationship between Newman’s assortativity coefficient [9] is stated as:

$$\phi = \frac{\mathbb{E}[x,y] - \mathbb{E}[x]\mathbb{E}[y]}{\sigma_x\sigma_y},$$

where $\sigma_x$ is the variance of $x$. For binary variables, the sample $\phi$ coefficient is based on the $2 \times 2$ contingency table:

$$
\begin{array}{c|cc}
  & y = 0 & y = 1 \\
\hline
x = 0 & e_{00} & e_{01} \\
 x = 1 & e_{10} & e_{11} \\
\end{array}
$$

where $a_i = \sum_j e_{ij}$, $b_j = \sum_i e_{ij}$, and $e_{ij}$ is the proportion of pairs for which $x = i$ and $y = j$, and $\sum_{ij} e_{ij} = 1$. The $\phi$ coefficient [9] is stated as:

$$\phi = \frac{e_{11} - a_1 b_1}{\sqrt{a_1 a_0 b_1 a_0}}.\tag{3}$$

Note that the numerator of Eq. (3) is written only in terms of nodes of category $x_i, y_i = 1$ for the reason that $e_{11} - a_1 b_1 = e_{00} - a_0 b_0$. Since we are considering the correlation of node metadata in undirected networks, $e_{ij}$ represents half the proportion of edges in the network that connect nodes with type $i$ to nodes with type $j$ (or the proportion of edges if $i = j$) and $a_i = b_i$. Now we can simplify the denominator of Eq. (3),

$$\sqrt{a_1 a_0 b_1 a_0} = a_1 a_0 = a_1(1 - a_1) = a_1 - a_1^2.$$
FIG. 1: (a) The difference between correlation and assortativity. In the correlation (left) of two variables $x$ and $y$ each of the samples (rows) are considered independent. For assortativity (right) samples are no longer independent because a single node corresponds to as many samples as it has links. (b) Graphical representation of the assortativity range in the metadata-graph space, graph space and metadata space. The assortativity range in the graph space is represented differently as it may be either smaller or larger than the assortativity range in the metadata space. (c) Graphical representation of block diagrams related to two different types of partition. In the case of a bipartite-like configurations most of the links fall between different partitions (i.e., connect nodes with different metadata) while in the case of a bisected-like configurations most of the links fall within each of the partitions (i.e., connect nodes with the same metadata).

Then by making these substitutions and summing over categories (since we do not assume that $a_0 = a_1$) we recover Newman’s assortativity $r$:

$$r = \sum_i e_{ii} - a_1^2 \left/ \left(1 - \sum_i a_i^2\right)\right..$$

(4)

The fact that the minimum value of assortativity $r_{\text{min}}$ may be greater than $-1$ was previously considered in [1] where a relatively conservative bound was derived by inspecting Eq. (4) and considering that the minimum occurs when all edges connect nodes with different metadata values, such that $\forall i, e_{ii} = 0$. Thus, we see that the assortativity is always lower bounded by $r_{\text{min}}$,

$$r_{\text{min}} = \frac{-\sum_i a_i^2}{1 - \sum_i a_i^2} \geq -1 ,$$

(5)

and when the metadata are binary $e_{ii} = 0$ implies that $e_10 = 1$ and the metadata form a bipartite split such that the bound above is saturated. However, we will show that, depending on the space of configurations considered, it is not always possible to arrange the metadata to form a bipartite split. In these cases $r = -1$ will be unattainable.
In what follows, it will become clearer to describe assortativity in terms of edge counts, rather than proportion of edges. To do so we consider that the $m$ edges are divided into three subsets such that $m = m_{11} + m_{10} + m_{00}$. Then we make the simple substitution $m_{ij} = (2 - \delta_{ij})c_{ij}m$, where $\delta$ is the Kronecker delta. Consequently the assortativity of binary node metadata can be written as:

$$r = \frac{(m_{00} + m_{11})m - (m_{00} + \frac{m_{00}m}{2})^2 - (m_{11} + \frac{m_{11}m}{2})^2}{m^2 - (m_{00} + \frac{m_{00}m}{2})^2 - (m_{11} + \frac{m_{11}m}{2})^2}.$$ (6)

The maximum value $r = 1$ occurs when edges only occur between nodes with the same metadata, i.e., $m_{00} + m_{11} = m$ and implies that the network is made up of multiple connected components, each containing only nodes with the same metadata value.

III. BOUNDS ON THE EDGE COUNTS

There are instances in which the bounds for assortativity, $-1 \leq r \leq 1$, can be attained. However, this is often not the case without changing certain properties of the network. In particular, the degree sequence, a specific set of edges and the way that node metadata values are assigned to specific nodes all play a role in limiting the range of permissible values of assortativity.

Instrumental to exploring the effect of structural properties on the bounds of assortativity is the dependence of assortativity on the edge counts $m_{11}$, $m_{10}$ and $m_{00}$, as shown in Eq. (6). To demonstrate, we consider the largest of the aforementioned ensembles of graphs, the metadata-graph space (mgs), in which we preserve only the degree sequence of the observed graph and the relative proportions of observed metadata values. Within this space we can state bounds on the possible edge counts $m_{11}$, $m_{10}$ and $m_{00}$ [13]. We denote the upper bounds with a superscript $u$ (e.g., $m_{11}^u$) and the lower bounds with a superscript $l$ (e.g., $m_{11}^l$).

In a graph with $n$ nodes that have a binary metadata assignment, we have $n_0$ nodes with metadata value $c_i = 0$ and $n_1$ nodes with metadata value $c_i = 1$. We define bounds on the edge counts by partitioning the ordered degree sequence $D_G$ and using this partition of the degree sequence to consider the maximum and minimum edge counts that it imposes. For example, to determine the upper bound of the number of edges $m_{11}^u$ that connect pairs of nodes with metadata value 1 we should consider that the maximum value of $m_{11}$, given $n_0$ and $n_1$, occurs when $n_1$ nodes are arranged into a complete subgraph. If $D_G$ doesn’t allow such a configuration, then the maximum value of $m_{11}$ occurs when $n_1$ nodes with the highest degree only connect to each other and not to any nodes with metadata $c_i = 0$. Therefore we partition the degree sequence into a head $D_G^H(n_1)$, comprised of the $n_1$ highest degrees, and a tail $D_G^T(n_0)$, containing the $n_0$ lowest degrees, such that $D_G = D_G^H(n_1) \cup D_G^T(n_0)$.

Figure 2 shows a simple example of such a partition of the degree sequence with two different values of $n_1$. For example, the maximum possible $m_{11}$ in any network that has degree sequence $D_G$ and $n_1$ has to be necessarily less than or equal to the number of links contained in the subgraph with degree sequence $D_G^H(n_1)$ (when we consider $D_G^H(n_1)$ the degree-sum of the $n_1$ elements of the network is maximised) or to the number of links contained in a clique of size $n_1$. The same reasoning is applicable also when we consider $m_{00}$ whose upper bound can be computed partitioning $D_G$ in a way such that $D_G = D_G^H(n_0) \cup D_G^T(n_1)$ as shown in Figure 2 (right). Following a similar rational of partitioning the ordered degree sequence, it is possible to either maximise or minimise the degree-sum of the two groups thus obtaining specific upper and lower bounds to the edge counts. Full details on the derivation of the upper and lower bounds ($m_{11}^u$, $m_{10}^u$, $m_{00}^u$) and ($m_{11}^l$, $m_{10}^l$ and $m_{00}^l$) are given in Appendix A.1.

The obtained bounds require only the degree sequence and the proportion of metadata to be set and so are suitable for the metadata-graph space. However, they can be trivially extended to the graph space by considering the case of a fixed partition of the degree sequence as explained in Appendix A.2.

IV. BOUNDS OF BINARY ASSORTATIVITY

The bounds of assortativity are directly related to the bounds on the edge counts. Specifically, we can consider two types of bipartition depending whether we wish to minimise or maximise the assortativity. Figure 4 describes this relationship. Maximising the assortativity corresponds to forming a minimum cut bisection of the network such that the majority of the edges connect nodes of the same type [Fig. 4]. Minimising the assortativity corresponds to finding a bipartite (or near bipartite) partition such that all (or most) of the edges connect a node $i$ that has metadata $c_i = 0$ to a node $j$ that has metadata $c_j = 1$ [Fig. 4]. The relevant bounds on the edge counts for minimising assortativity are $\{m_{00}, m_{11}, m_{10}\}$ with the minimum of $r = -1$ occurring if and only if $m_{00} + m_{11} = 0$. Similarly maximising the assortativity depends on the edge counts $\{m_{00}, m_{11}, m_{10}\}$ and the maximum value of $r = 1$ is only attainable if $m_{10} = 0$ and must be composed of separate disconnected components that contain only nodes of the same type.
In the following we discuss bounds of assortativity in relation to the limits imposed by the three spaces, the metadata-graph space (mgs), the graph space (gs) and the metadata space (ms). Throughout we will assume that $n_1 \notin \{0, n\}$ to ensure that both groups of the partition are nonempty.

### A. Bounds for the metadata-graph space

The metadata-graph space contains all configurations of graph structures and node metadata assignments that have a specified degree sequence $D_G$ and given number of nodes of each type $\{n_0, n_1\}$. The bounds on the edge counts described in Section II depend only upon the specific degree sequence and the number of nodes of each type. We can therefore use these directly to define the bounds upon the metadata-graph space.

#### 1. Upper bound

The maximum value of $r_{\text{mgs}}$, for connected networks, occurs when the minimum edges link nodes of different types. Therefore we define our upper bound $r_{\text{mgs}}^u$ by setting $m_{10} = m_{10}^l$. The maximum value of assortativity $r = 1$ can only be attained if the graph can be partitioned into disconnected components that contain only a single type of node. This constraint implies that when we find that the lower bound $m_{10}^l$ is greater than zero, then the maximum possible value of assortativity is less than 1. Setting $m_{10} = m_{10}^l$ implies that $m_{11} + m_{00} = m - m_{10}^l$ and we can write $r$ as:

$$r = \frac{(m - m_{10}^l)m - (m_{00} + \frac{m_{10}^l}{2})^2 - (m_{11} + \frac{m_{10}^l}{2})^2}{m^2 - (m_{00} + \frac{m_{10}^l}{2})^2 - (m_{11} + \frac{m_{10}^l}{2})^2}.$$  \hspace{1cm} (7)
Noting that $m_{00} = m - m_{10} - m_{11}$ we can substitute this quantity into Eq. [7] and write \( r \) with respect to \( m_{11} \), that is also the only variable of Eq. [8]

\[
\begin{align*}
  r &= \frac{(m - m_{10})m - (m - m_{11} - \frac{m_{10}}{2})^2 - (m_{11} + \frac{m_{10}}{2})^2}{m^2 - (m - m_{11} + \frac{m_{10}}{2})^2 - (m_{11} + \frac{m_{10}}{2})^2} \\
  &= \frac{(m - m_{10})m - (m - m_{11} - \frac{m_{10}}{2})^2 - (m_{11} + \frac{m_{10}}{2})^2}{m^2 - (m - m_{11} + \frac{m_{10}}{2})^2 - (m_{11} + \frac{m_{10}}{2})^2}
\end{align*}
\]  

(8)

In order to obtain the value of \( m_{11} \) that maximises \( r \) we can solve the following equation:

\[
\frac{\partial r}{\partial m_{11}} = 0.
\]

(9)

We obtain \( m_{11} = \frac{m - m_{10}}{2} \) that implies \( m_{00} = \frac{m - m_{10}}{2} \) and therefore:

\[
\begin{align*}
  r_{\text{mgs}} &= 1 - \frac{2m_{10}}{m}.
\end{align*}
\]

(10)

The maximisation of \( r \) is confirmed by the fact that assortativity, fixed \( m_{10} \), is a concave function (as also shown in Figure [3]). The concavity of \( r \) can be also proved by deriving Eq. [10] with respect to \( m_{10} \) obtaining \( -\frac{2}{m} \); such an outcome implies that \( r \) decreases when \( m_{10} \) increases and setting \( m_{10} = m_{10}^{\text{opt}} \) guarantees an upper bound to assortativity.

---

**FIG. 3:** Assortativity as a function of the difference in edge counts \( m_{11} - m_{00} \). The left panel displays the assortativity function \( r \) for different values of \( m_{10} \) as a proportion of \( m \) while the difference between \( m_{11} \) and \( m_{00} \), represented by \( \Delta = m_{11} - m_{00} \), varies. We notice that assortativity is maximised whenever \( m_{11} = m_{00} \). The right panel refers to the degree sequence displayed in Fig. [2] when \( n_1 = 5 \). In this case we set \( m_{10} = m_{10}^{\text{opt}} = 4 \), which implies \( m - m_{10} = m_{11} + m_{00} = 16 \). Thus, we obtain the maximum value of assortativity when \( m_{11} = m_{00} = \frac{m - m_{10}}{2} \)

i.e., when \( \frac{m - m_{10}}{2} = 8 \).

2. Lower bound

The minimum value of \( r_{\text{mgs}} \), for connected networks, occurs when the partition of the node metadata forms a bipartite split of the graph. When \( m_{11}^{\text{opt}} + m_{00}^{\text{opt}} > 0 \) for a given degree sequence \( D_G \) and group sizes \( n_0 \) and \( n_1 \) it means that a certain level of interaction between nodes that share a common property (intra-partition links) will be present and therefore a graph with a bipartite partition \( G_{n_1,n_0} \) is not realisable. As introduced in Sec. [I] in [1] Newman provides the following lower bound to assortativity:

\[
\begin{align*}
  r_{\text{min}} &= - \frac{\sum_i a_i^2}{1 - \sum_i a_i^2}
\end{align*}
\]

(11)
Such a lower bound, allowing the existence of a bipartite split of the nodes, assumes that the sum of the proportion of intra-partition links is zero, i.e. \( \sum_i e_{ii} = 0 \). Written in terms of edge counts for binary metadata we have \( m_{11} + m_{00} = 0 \) that implies \( m_{10} = m \), since \( m_{11} + m_{00} + m_{10} = m \), and by substituting such quantities in the previous equation we obtain:

\[
 r_{\text{min}} = - \frac{2(m_{\text{min}})^2}{1 - 2(m_{\text{min}}/2m)^2} = -1. \tag{12}
\]

Following the Newman’s reasoning but considering this time the bounds on the edge counts we have to take into account the following constraints: \( m_{11} \geq m_{11}' \geq 0 \), \( m_{00} \geq m_{00}' \geq 0 \) and \( m_{10} \leq m_{10}' \leq m \). Therefore, in order to obtain a lower bound to \( r \) we need to consider a realisation close to a bipartite split of the graph (i.e. with the highest \( m_{10} \)) that can be achieved considering two options. The first option consists in setting \( m_{11} = m_{11}' \) and \( m_{00} = m_{00}' \) that implies \( m - m_{11}' - m_{00}' = m_{10} \). Through a simple substitution of such quantities we obtain:

\[
 r_{\text{mgs}}^l = \frac{(m_{10}' + m_{11}')m - (m_{10} + m_{11} - \frac{m_{10}}{2})^2 - (m_{11} + m_{00} - \frac{m_{00}}{2})^2}{m^2 - (m - m_{11} - \frac{m_{10}}{2})^2 - (m + m_{00} - \frac{m_{00}}{2})^2}. \tag{13}
\]

The second option to obtain a lower bound to \( r \) is to exploit the upper bound to the edge count \( m_{10} \), instead of the lower bounds to the edge counts \( m_{11} \) and \( m_{00} \), setting \( m_{10} = m_{10}' \).

When \( m_{10} = m_{10}' \) then \( m_{11} + m_{00} = m - m_{10}' \); noting that, for a fixed value of \( m_{10} \), binary assortativity is a concave function and that \( m_{10}' \) can be different from \( m_{11} \) we have two further options to determine the lower bound. When \( m_{11} + m_{00} > 0 \) the minimum assortativity can be obtained when the absolute difference \( |\Delta| = |m_{11} - m_{00}| \) is maximised (as displayed in Figure 3). In the first case we set \( m_{10} = m_{10}' \), \( m_{11} = m_{11}' \) and if \( m_{10}' + m_{11}' \leq m \) then \( m_{00} = m - m_{10}' - m_{11}' \) which is the \( \Delta_{\text{min}} \) case. Thus:

\[
 r_{\text{mgs}}^l = \frac{(m - m_{10}')m - (m - m_{11}' - \frac{m_{10}}{2})^2 - (m_{11}' + m_{00}' - \frac{m_{00}}{2})^2}{m^2 - (m - m_{11}' - \frac{m_{10}}{2})^2 - (m + m_{00}' - \frac{m_{00}}{2})^2}. \tag{14}
\]

In the second case we set \( m_{10} = m_{10}' \), \( m_{00} = m_{00}' \) and if \( m_{10}' + m_{11}' \leq m \) then \( m_{11} = m - m_{10}' - m_{00}' \) which is the \( \Delta_{\text{max}} \) case. Thus:

\[
 r_{\text{mgs}}^l = \frac{(m - m_{10}')m - (m_{10}' + m_{11}' - \frac{m_{10}}{2})^2 - (m - m_{00}' - \frac{m_{00}}{2})^2}{m^2 - (m_{10}' + m_{11}' - \frac{m_{10}}{2})^2 - (m - m_{00}' - \frac{m_{00}}{2})^2}. \tag{15}
\]

In summary, we have three possible cases for determining the lower bound:

\[
 \{m_{00}, \, m_{11}, \, m_{10}\} = \begin{cases} 
 \{m_{00}', \, m_{11}', \, m - m_{11}' - m_{00}' \} \\
 \{m - m_{10}' - m_{11}', \, m_{11}', \, m_{00}' \} \\
 \{m_{00}', \, m - m_{10}' - m_{00}' \, m_{10}' \}
 \end{cases}, \tag{16}
\]

which we can substitute into Eq. 6 considering the minimum value of \( r \) as the bound \( r_{\text{mgs}}^l \).

**B. Bounds for the graph space**

Similar to the metadata-graph space, the graph space considers all configurations of graph structures that have the specified degree sequence. The difference is that in the graph space the assignment of metadata values to nodes is fixed and so the degree-metadata correlation is also fixed. We can bound the assortativity of the graph space using the same rationale as the metadata-graph space. In fact, we can use the same equations given in the previous subsection by replacing the bounds on the edge counts for the metadata-graph space with those of the graph space, which are given in Section A.2

The range of assortativity in the metadata-graph space is at least as large as the range in the graph space since the former has an extra degree of freedom by allowing all the possible arrangements of the metadata over the network nodes. As such, the following relations hold:

\[
 r_{\text{gs}}^u \leq r_{\text{mgs}}^u \leq r_{\text{max}}, \tag{17}
\]

\[
 r_{\text{min}} \geq r_{\text{mgs}}^l \geq r_{\text{gs}}^l. \tag{18}
\]
The difference in the assortativity bounds for the metadata-graph space and the graph-space vary according to \( n_1 \), as this controls how many distinct metadata assignments there can be per graph configuration. The ranges are equal when \( n_1 = \{0, n\} \), because there can be only a single metadata assignment for every graph configuration, and the difference in the ranges is maximised when \( n_1 = \frac{n}{2} \).

### C. Bounds for the metadata space

Unlike the metadata-graph space and the graph space, we cannot use any theoretical bound for assortativity in the metadata space because it relies upon the specific network topology. Therefore we must resort to a complete enumeration, when feasible, or a heuristic algorithm, based on a Monte Carlo exploration of the metadata space (one such algorithm is described in Section A.3). As in the previous section, we can clearly determine that the range of assortativity in the metadata-graph space is at least as large as the metadata space, because for each unique assignment of metadata values to node degrees, the metadata-graph space contains all possible graph configurations with the given degree sequence. Therefore,

\[
\begin{align*}
\frac{r_{\text{ns}}}{r_{\text{ms}}} &\leq \frac{r_{\text{ugs}}}{r_{\text{ms}}} \\
\frac{r_{\text{min}}}{r_{\text{ms}}} &\leq \frac{r_{\text{max}}}{r_{\text{ms}}}.
\end{align*}
\]

We cannot guarantee, however, any relationship between the metadata space and the graph space since they are constrained by different elements. The former is constrained by the topology and by the proportion of metadata labels while the latter by the degree sequence and by the assignment of metadata to specific nodes. Instead of a combinatorial bound in the metadata space, we demonstrate the bounds empirically on a synthetic network. Using the network in Figure 2 (top left network), we look at all the possible permutations of the metadata in order to compute the distribution of assortativity values.

Figure 4 shows the distribution of assortativity values over a complete enumeration of metadata assignments for the network in Figure 2 (top left). The histogram on the left shows the distribution for \( n_1 = 5 \), while the one of the right shows the distribution for \( n_1 = 3 \). Here we can clearly observe the minimum and maximum values of the assortativity coefficient permissible in the metadata space.

### V. EXPERIMENTS ON REAL NETWORKS

In this section we investigate assortative mixing for binary metadata in real-world networks. Examples of binary metadata of network nodes can be found in a wide array of contexts, including: the functional categories of proteins in protein-protein interaction networks [8], the hydrophobic/hydrophilic nature of proteins in protein contact networks [14], and the use of a specific service in telecommunication networks [8]. Here we will focus on another natural case study on binary node metadata, which is gender assortativity in social networks of animals [15] and humans [16].

The investigation of gender assortativity is interesting for a number of practical reasons related to human behaviour and the adoption of specific habits [17–20]. Moreover, better understanding of the mixing patterns and preferences in social networks plays an important role in predicting missing metadata such as gender [21].

Here we investigate gender assortativity in two colleges, Smith and Wellesley, extracted from the Facebook 100 dataset [22] (see Appendix A.5 for dataset description). Smith displays a gender assortativity of \( r = 0.02 \) that is positive but close to 0 (i.e., close to a random distribution links relative to the node metadata), while in Wellesley we see a value of \( r = 0.24 \) that should indicate a distinctive pattern of assortativity by gender.

Figure 5 displays the assortativity bounds for the metadata-graph and graph spaces for the Smith (left) and Wellesley (right) social networks. To evaluate the metadata space, Figure 5 also shows histograms of assortativity values for \( 10^5 \) permutations of the binary metadata vector \( c \) for each network and the extremal values obtained from minimising and maximising the assortativity using an optimisation heuristic, described in Appendix A.3. We immediately observe that the disassortativity of both networks is bounded away from -1, in all three spaces, such that the network cannot be very disassortative and that there are relatively few configurations of the network and of it metadata that allow disassortative mixing. This effect is partially due to the huge gender imbalance as both colleges are female only and so there are relatively few males (staff members) in the network.

Comparing the ranges of assortativity in the metadata space and graph space, we observe that: \( r_{\text{ms}}^{\text{max}} \leq r_{\text{gs}}^{\text{u}} \) and \( r_{\text{ms}}^{\text{min}} \leq r_{\text{gs}}^{\text{u}} \) for both these networks. So for these networks the metadata space allows more disassortative mixing than the graph space does. It also appears as though the graph space allows for more assortative mixing than the metadata space. However, we should consider the latter with caution since \( r_{\text{ms}}^{\text{max}} \) is computed via a heuristic (thus it is a lower bound to the actual maximum) while \( r_{\text{gs}}^{\text{u}} \) is an upper bound. Therefore, given the fact that \( r_{\text{ms}}^{\text{max}} \) and
FIG. 4: Ranges of assortativity for the network shown in the top left of Fig. 2. The plot on the left has \( n_1 = 5 \) and the one on the right has \( n_1 = 3 \). The histograms show the distribution of assortativity values in the metadata space (complete enumeration of all permutations of metadata assignments). The dashed lines indicate the bounds \( r_{mgs}^l \) and \( r_{mgs}^u \) of the assortativity range in the metadata-graph space. The dotted lines indicate the bounds \( r_{gs}^l \) and \( r_{gs}^u \) in the graph space. When \( n_1 = 5 \) the values are \( r_{mgs}^l = -1 \), \( r_{gs}^l = -0.8 \), \( r_{ms}^{\text{min}} = -0.6 \), \( r_{ms}^{\text{max}} = 0.2 \), \( r_{gs}^u = 0.499 \), \( r_{mgs}^u = 0.6 \). When \( n_1 = 3 \) these values are \( r_{mgs}^l = -0.8 \), \( r_{gs}^l = -0.704 \), \( r_{ms}^{\text{min}} = -0.704 \), \( r_{ms}^{\text{max}} = -0.303 \), \( r_{gs}^u = 0.2 \), \( r_{mgs}^u = 0.827 \).

\( r_{gs}^u \) have very close values it indicates a very high similarity in terms of upper bound of assortativity in graph and metadata spaces.

With regard to the metadata space, we see that for both networks the observed assortativity value is higher than the assortativity of random permutations. We can interpret such a result as a test of statistical significance \[23\]. So even though the assortativity is relatively low (particularly for the Smith network) we can still conclude that the assortativity is significantly higher than a random partition of the network \( (p < 10^{-5}) \).

In order to complement the previous analysis, we also consider a smaller but much denser network, the Wolf Dominance network \[24\] (see Appendix A 5), over which we evaluate gender assortativity. In this smaller network it is possible to evaluate the metadata space via a complete enumeration of the possible metadata permutations and so we can compute the actual values of \( r_{ms}^{\text{min}} \) and \( r_{ms}^{\text{max}} \) to compare against the combinatorial bounds of the graph space and metadata-graph space.

Interestingly in this case, the upper bounds on assortativity in all three spaces are very close to zero and in the metadata space it is not possible to observe positive assortativity. Furthermore, the mean value of assortativity over the metadata space is not zero, as we can see from the histogram centred at \(-0.06\). This observation seems contrary to our expectation that assortativity of random partitions should be centred around zero. This result resembles that of Ref. \[7\] in which it was observed that under certain conditions the expected value of assortativity in the graph space is not equal to zero. Here we observe that this can also be true of the metadata space. Therefore, following a similar argument, we may conclude that in cases such as these an adjustment to the expected value may be necessary.

VI. DISCUSSION

The assortativity coefficient \( r \) is generally assumed to range between \(-1 \) and \( 1 \). Here we have shown how, in the case of binary node metadata, constraints of the network structure can further limit the range of attainable values. These constraints are represented by:

- the metadata-graph space: the range of assortativity values over the ensemble of configurations with a given
FIG. 5: Assortativity bounds for the colleges Smith (left) and Wellesley (right) of the largest connected component after eliminating nodes with missing gender metadata. Smith college has \( n = 2625, m = 77259, n_{\text{female}} = n_0 = 2596, n_{\text{male}} = n_1 = 29 \). Smith has a gender assortativity value \( r = 0.025 \) (solid line) with \( m_{11} = 25 \) and \( m_{10} = 1404 \). Wellesley college has \( n = 2689, m = 78853, n_{\text{female}} = n_0 = 2653, n_{\text{male}} = n_1 = 36 \). Wellesley has a gender assortativity value \( r = 0.246 \) (solid line) with \( m_{11} = 122 \) and \( m_{10} = 729 \).

FIG. 6: Assortativity bounds for the Wolf Dominance network \((n = 16, m = 111, n_{\text{female}} = n_0 = 7, n_{\text{male}} = n_1 = 9)\). The network has a gender assortativity value \( r = -0.153 \) (solid line) that occurs in correspondence with \( m_{11} = 31 \) and \( m_{10} = 63 \). The dashed lines, obtained with the combinatorial bounds, occur in correspondence of \( r^{\text{min}}_{\text{mgs}} = -0.263 \) and \( r^{\text{max}}_{\text{mgs}} = 0.072 \). The dot-dashed lines occur in correspondence of the values \( r^{\text{min}}_{\text{ms}} = -0.16 \) and \( r^{\text{max}}_{\text{ms}} = 0.009 \) which are obtained via complete enumeration. The bounds to the graph space are represented by dotted lines at \( r^{\text{min}}_{\text{gs}} = -0.153 \) and \( r^{\text{max}}_{\text{gs}} = -0.007 \).
degree sequence $D_G$ and number of nodes of each type $n_0, n_1$. Here we provide a combinatorial lower bound $r^l_{mgs}$ and a combinatorial upper bound $r^u_{mgs}$.

- the **graph space**: the range of assortativity values over the ensemble of configurations with a given degree sequence $D_G$ and a specific assignment of metadata to nodes. We provide a combinatorial lower bound $r^l_{gs}$ and a combinatorial upper bound $r^u_{gs}$.

- the **metadata space**: the range of assortativity values over the ensemble of permutations of the metadata labels (preserving the counts $n_0, n_1$) on the specific topology of an observed graph. Here we propose the use of a heuristic to determine the upper bound $r^{\leq \max}_{ms}$ and lower bound $r^{\geq \min}_{ms}$.

The choice of ensemble should depend upon the specific problem at hand and relate to the specific assumptions we wish to make about the graph structure and metadata assignment. For instance, when investigating metadata such as gender in a social network we might consider the metadata and popularity of the nodes (i.e., their degrees) to be fixed and so the graph space might be most appropriate. Alternatively, in a road network in which the metadata indicates either presence or absence of road signals, occurrences of traffic jams or the locations of accident hotspots, we might want to consider the metadata space as the structure of the graph is fixed.

Although we have focused on binary metadata, the issue of attaining the extremal values $\{-1, 1\}$ of assortativity is still present for any categorical-valued metadata [25]. Taken altogether we can conclude that these constraints present questions about the interpretability of network assortativity, especially when comparing across networks [4, 22, 26, 27]. As a potential solution, we might consider normalising assortativity according to the bounds of the space most relevant to our given problem. For instance, if we consider the degree and metadata value of a node to be fixed, then an appropriate normalisation might be:

$$r_{gs} = \begin{cases} r_{gs} & \text{if } r \text{ is positive} \\ \frac{r_{gs}}{r^u_{gs}} & \text{otherwise} \end{cases}.$$  (21)

Such a normalisation has previously been suggested for related measures such as the $\phi$ coefficient and Cohen’s $\kappa$ [12, 25]. It also follows the rationale that assortativity is a normalised version of modularity $Q$, i.e., $r = Q/Q_{\text{max}}$. Alternatively we may consider comparing the observed assortativity with the distribution of assortativity values in the relevant ensemble, e.g., using assortativity as a test statistic in a one-sided hypothesis test to assess statistical significance [23, 28]. However, we leave the exploration of these ideas for future work.

Another avenue for future work would be to consider how a given ensemble constrains other network measures such as Freeman’s segregation [29], which is limited by the edge count $m_{10}$ (see note in Appendix A 4), and the clustering coefficient, which is closely related to assortativity to scalar features such as degree [30].

**ACKNOWLEDGEMENTS**

The authors thank Alexandre Bovet, Fariba Karimi and Leo Pen for insightful conversations. This work was supported in part by the F.R.S-FNRS Grant No. 1.B.336.18F [LP], the Concerted Research Action (ARC) programme (contract ARC 14/19-060) of the Federation Wallonia-Brussels [LP, JCD] and Flagship European Research Area Network (FLAG-ERA) Joint Transnational Call “FuturICT 2.0” [JCD]

[1] Mark EJ Newman, “Mixing patterns in networks,” Physical Review E 67, 026126 (2003).
[2] Rogier Noldus and Piet Van Mieghem, “Assortativity in complex networks,” Journal of Complex Networks 3, 507–542 (2015).
[3] Dan Yang, Liming Pan, and Tao Zhou, “Lower bound of assortativity coefficient in scale-free networks,” Chaos: An Interdisciplinary Journal of Nonlinear Science 27, 033113 (2017) [https://doi.org/10.1063/1.4976030]
[4] Mohsen Jadidi, Fariba Karimi, Haiko Lietz, and Claudia Wagner, “Gender disparities in science? dropout, productivity, collaborations and success of male and female computer scientists,” Advances in Complex Systems, 1750011 (2017).
[5] Kristina Lerman, Xiaoran Yan, and Xin-Zeng Wu, “The’ majority illusion’ in social networks,” PloS one 11, e0147617 (2016).
[6] Eun Lee, Fariba Karimi, Claudia Wagner, Hang-Hyun Jo, Markus Strohmaier, and Mirta Galesic, “Homophily and minority-group size explain perception biases in social networks,” Nature Human Behaviour (2019).
[7] Bailey K Fosdick, Daniel B Larremore, Joel Nishimura, and Johan Ugander, “Configuring random graph models with fixed degree sequences,” SIAM Review 60, 315–355 (2018).
1. Bounding the edge counts in the metadata-graph space

Here we summarize the bounds introduced in [13]. Given a degree sequence $D_G$, by using the quantities $n_1$ and $n_0$ which identify the amount of nodes with features 1 and 0 respectively, it is possible to define its head $D_G^H(n_1)$ or $D_G^H(n_0)$ and its tail $D_G^T(n_1)$ or $D_G^T(n_0)$ such that $D_G = D_G^H(n_1) \cup D_G^T(n_0)$ or $D_G = D_G^H(n_0) \cup D_G^T(n_1)$.

Considering these partitions, the first upper bound $n_{11}$ is based on the fact that, especially in sparse networks, large cliques may be rare substructures. Therefore, using $D_G$ we check whether $G$ can actually contain a complete subgraph of size $n_1$ (i.e. if $D_G^H(n_1)$ satisfies the necessary condition for the realisation of a clique). If not, we take into account the densest hypothetical substructure that could be realised using the degree sequence of $G$. In equation A1 the first term is the number of links in the network, the second term is the number of links in a clique of size $n_1$, while the third term is the number of links in the sub-graph with $n_1$ nodes and maximum degree-sum (i.e. with
degree sequence $D^H_G(n_1))$.

$$m^{u}_{11} = \min \left( m, \frac{n_1}{2}, \left[ \sum_{i \in D^H_G(n_1)} \min(d_i, n_1 - 1) \right] \right),$$  \hspace{1cm} (A1)

In the second upper bound, $m^{u}_{11}$, we check if $G$ can contain a complete bipartite subgraph with partitions size $n_1$ and $n_0$. If not, we consider a set of stars made of the first $n_1$ elements of $D_G$ if $n_1 < n_0$ or made of the first $n_0$ elements of $D_G$ if $n_0 < n_1$. In equation [A2] the first term is the number of links in the network, the second term is the number of links in a bipartite graph with partitions of size $n_1$ and $n_0$, while the third term is the minimum between the number of $m_{10}$ deriving from the degree partition $D^H_G(n_1) \cup D^H_G(n_0)$ and the number of $m_{10}$ deriving from the degree partition $D^H_G(n_0) \cup D^H_G(n_1)$.

$$m^{l}_{10} = \min \left( m, n_1n_0, \min \left( \sum_{i \in D^H_G(n_1)} \min(d_i, n_0), \sum_{i \in D^H_G(n_0)} \min(d_i, n_1) \right) \right),$$  \hspace{1cm} (A2)

The first lower bound, $m^{l}_{11}$, considers the partition $D^H_G(n_1)$ and the minimum residual degree of its elements (when $n_1 > 0$), which is exploited in order to realise the minimum $m_{11}$. Since most of the real networks are sparse, this bound is effective mainly in the case of unbalanced partitions and of dense networks. The second term of equation [A3] counts the minimum number of links among the $n_1$ nodes in the graph deriving from the partition $D^H_G(n_0) \cup D^H_G(n_1)$, i.e. the amount of $m_{11}$ which is realizable from the residual degree of the partition $D^H_G(n_1)$.

$$m^{l}_{11} = \max \left( 0, \left[ \sum_{i \in D^H_G(n_1)} d_i - \sum_{i \in D^H_G(n_0)} d_i \right] \right),$$  \hspace{1cm} (A3)

The second lower bound, $m^{l}_{10}$, considers that the lower $m_{10}$ occurs in the case of a bisected network (i.e. a network with two separated components). Thus, if the degree sum in $D^H_G(n_1)$ overcomes the degree sum in a clique of size $n_1$ then we guarantee the presence of some $m_{10}$. Considering that any connected realization with $n_1 \neq \{0, n\}$ has at least one $m_{10}$, the second term of equation [A4] counts the minimum number of links between the $n_1$ and $n_0$ in the case the $n_1$ are arranged into a clique.

$$m^{l}_{10} = \begin{cases} 0, & \text{if } n_1 = 0, n \\ \max \left( 1, \sum_{i \in D^H_G(n_1)} d_i - n_1(n_1 - 1) \right), & \text{if } n_1 \in (0, n) \end{cases}$$  \hspace{1cm} (A4)

The bounds to $m_{00}$ can be obtained using the same rationale as that of $m_{11}$.

\hspace{1cm} a. **Improvements to lower bounds in the metadata-graph space**

The lower bound to the intra-partition links is $m^{l}_{11}$. It can be initially improved by correcting the term $\sum_{i \in D^H_G(n_0)} d_i$. This term keeps the bound low especially in the case of unbalanced partitions and in the case of heavy tailed and sparse, networks (i.e. when the degree sum of $D^H_G(n_0)$ has a high value because of the presence of hubs). Known the size of the two partitions, the second term in $m^{l}_{11}$ can be written as: $\sum_{i \in D^H_G(n_0)} \min(d_i, n_1)$.

Indeed, any node in $n_0$, despite its degree, can be connected at most to other $n_1$ ones in a different partition. Consequently the residual degree of the nodes in $D^H_G(n_1)$ can be exploited for the realisation of $m_{11}$.

Therefore:

$$m^{l}_{11} = \max \left( 0, \left[ \sum_{i \in D^H_G(n_1)} d_i - \sum_{i \in D^H_G(n_0)} \min(d_i, n_1) \right] \right),$$  \hspace{1cm} (A5)

The bound to the inter-partition links is $m^{l}_{10}$. The first extension consists in making the bound symmetrical by adding the term $\sum_{i \in D^H_G(n_0)} d_i - n_0(n_0 - 1)$ and in noticing that such term can be written in a more efficient way as $\sum_{i \in D^H_G(n_0)} \max(0, d_i - (n_0 - 1))$. As shown in [13], the current bound works better in the case of dense networks since, when $n_1$ becomes larger, the nodes in $D^H_G(n_1)$ may still have a residual degree which is higher than the degree of the nodes in a clique of size $n_1$ (i.e. certain elements in $D^H_G(n_1)$ have degree greater than $n_1 - 1$). Conversely, if
the considered network is relatively sparse we may not be able to provide a lower bound to \( m_{10} \) which is greater than zero even for very low values of \( n_1 \).

Therefore, given that \( n_1 + n_0 = n \), when \( n_1 \) increases we should also try to bound \( m_{10} \) by supposing a realisation in the tail of \( D_G \) that involves \( n_0 \) nodes. Thus, the symmetrical version of \( m_{10}^l \) comprises the term \( \sum_{i \in D_G(n_0)} d_i - n_0(n_0 - 1) \).

An additional improvement, possibly more appropriate in the case of heavy tailed and sparse degree sequences, derives from the following consideration: called \( D_G(n_1) \) and \( D_G(n_0) \) two arbitrary partitions of \( D_G \), any element in \( D_G(n_1) \) \((D_G(n_0))\) can be connected at most to other \( n_1 - 1 \) \((n_0 - 1)\) ones in the same partition. Thus, any element in \( D_G(n_1) \) \((D_G(n_0))\) can be involved in at least \( d_i - (n_1 - 1) \((d_i - (n_0 - 1)\) intra-partition links.

Given a certain arbitrary partition of \( D_G = D_G(n_1) \cup D_G(n_0) \), the minimum amount of \( m_{10} \) that can be realised is:

\[
m_{10} = \frac{1}{2} \left( \sum_{i \in D_G(n_1)} d_i - (n_1 - 1) + \sum_{j \in D_G(n_0)} d_j - (n_0 - 1) \right).
\]

In the case \( n_1 > n_0 \) the following relation holds:

\[
\frac{1}{2} \left( \sum_{i \in D_G(n_1)} d_i - (n_1 - 1) + \sum_{j \in D_G(n_0)} d_j - (n_0 - 1) \right) \geq
\frac{1}{2} \left( \sum_{i \in D_G(n_1)} d_i - (n_1 - 1) + \sum_{j \in D_G(n_0)} d_j - (n_1 - 1) \right).
\]

The second term of such a relation assumes, in order to provide a lower bound to \( m_{10} \), that any element of \( D_G \) has the lowest possible residual degree for the realisation of \( m_{10} \). Obviously, the quantity \( d_i - (n_1 - 1) \) has to be greater than 0 for each \( i \) and the second term of the previous inequality represents the lowest possible sum of residual degrees of any arbitrary partition, in the case \( n_1 > n_0 \). Thus, the previous relation can be written as:

\[
\frac{1}{4} \sum_{i=1}^n \max(0, d_i - (n_1 - 1)).
\]

Finally, \( m_{10}^l \) can be expressed as:

- if \( n_1 = 0, n \)

\[
m_{10}^l = 0.
\]

- if \( n_1 > n_0 \)

\[
m_{10}^l = \max \left( 1; \sum_{i \in D_G(n_1)} \max(0, d_i - (n_1 - 1)); \sum_{i \in D_G(n_0)} \max(0, d_i - (n_0 - 1)); \left[ \frac{1}{2} \sum_{i=1}^n \max(0, d_i - (n_1 - 1)) \right] \right)
\]

- if \( n_1 \leq n_0 \)

\[
m_{10}^l = \max \left( 1; \sum_{i \in D_G(n_1)} \max(0, d_i - (n_1 - 1)); \sum_{i \in D_G(n_0)} \max(0, d_i - (n_0 - 1)); \left[ \frac{1}{2} \sum_{i=1}^n \max(0, d_i - (n_0 - 1)) \right] \right)
\]

Or in a more compact way when \( n_1 \neq 0, n \):

\[
m_{10}^l = \max \left( 1; \sum_{i \in D_G(n_1)} \max(0, d_i - (n_1 - 1)); \sum_{i \in D_G(n_0)} \max(0, d_i - (n_0 - 1)); \left[ \frac{1}{2} \sum_{i=1}^n \max(0, d_i - (\max(n_1, n_0) - 1)) \right] \right)
\]

2. Bounding the edge counts in the graph space

We consider the graph space into which the degree sequence \( D_G \) and the vector of binary node metadata are both fixed. In such a case, we say that \( D_G = D_G(n_1) \cup D_G(n_0) \) which represents the current partition of the considered degree sequence, given the node metadata assignment. Therefore, we can exploit the combinatorial bounds of the metadata-graph space in order to bound the different edge counts in the graph space. The rationale behind the bounds remain the same as well as the formulas (presented in Sections \ref{sec:14} and \ref{sec:14a}) which can be, however, contracted as we can’t leverage, within the graph space, the different ways of partitioning \( D_G \). Therefore, the bounds in the graph space can be written as:

\[
m_{11}^n = \min \left( m_{11} \left( \frac{n_1}{2} \right); \left\lceil \sum_{i \in D_G(n_1)} \frac{\min(d_i, n_1 - 1)}{2} \right\rceil \right)
\]
\begin{align*}
    m^u_{10} &= \min\left(m, n_1n_0, \min\left(\sum_{i \in D_G(n_1)} \min\left(d_i, n_0\right), \sum_{i \in D_G(n_0)} \min\left(d_i, n_1\right)\right)\right) \quad (A11) \\
    m^l_{11} &= \max\left(0, \frac{\sum_{i \in D_G(n_1)} d_i - \sum_{i \in D_G(n_0)} \min\left(d_i, n_1\right)}{2}\right) \quad (A12) \\
    m^l_{10} &= \max\left(1; \sum_{i \in D_G(n_1)} \max(0, d_i - (n_1 - 1)); \sum_{i \in D_G(n_0)} \max(0, d_i - (n_0 - 1))\right) \quad (A13)
\end{align*}

3. Swap of node metadata

In order to approximate the maximum and minimum values of binary assortativity in the metadata space we use the following heuristic procedure which provides admissible solutions to the graph bisection problem also in the case of unbalanced partitions.

1. Take into account the network, the metadata vector \(c\) and compute \(r^{\text{current}}\)

2. Take into account two randomly chosen entries of \(c\), called \(c_i\) and \(c_j\), such that \(c_i = 1\) and \(c_j = 0\) (or viceversa)

3. Swap the values of \(c_i\) and \(c_j\) and compute \(r^{\text{swap}}\)

   - In the case of assortativity maximization:
     - if \(r^{\text{swap}} > r^{\text{current}}\) then the switch is accepted and \(r^{\text{current}} = r^{\text{swap}}\)
     - if \(r^{\text{swap}} \leq r^{\text{current}}\) then with probability \(p = 0.001\) the swap is accepted, and \(r^{\text{current}} = r^{\text{swap}}\)

   - In the case of assortativity minimization:
     - if \(r^{\text{swap}} < r^{\text{current}}\) then the swap is accepted and \(r^{\text{current}} = r^{\text{swap}}\)
     - if \(r^{\text{swap}} \geq r^{\text{current}}\) then with probability \(p\) the swap is accepted, and \(r^{\text{current}} = r^{\text{swap}}\)

Steps 2 and 3 of procedure are iterated several times and different repetitions are performed.

4. Freeman’s Segregation

By using the notation of [29], segregation \(S\) can be expressed starting from the relation:

\[
    s = \begin{cases} 
        \mathbb{E}(e^*) - e^* & \text{if and only if } \mathbb{E}(e^*) \geq e^* \\
        0 & \text{otherwise}
    \end{cases} \quad (A14)
\]

in such a formula \(e^*\) is the number of cross-class edges (i.e. \(m_{10}\)) and \(\mathbb{E}(e^*)\) is the first moment of \(e^*\). Therefore, \(S\) is expressed as \(S = \frac{s}{\mathbb{E}(e^*)} \in [0, 1]\). Any value of \(S\) may be interpreted simply as the ratio of the number of missing cross-class links to the expected number of such links. By using the notation that we adopted throughout the paper, we can express the number of cross-class links as have \(e^* = m_{10}\). Thus:

\[
    S = \frac{\mathbb{E}(m_{10}) - m_{10}}{\mathbb{E}(m_{10})} \quad (A15)
\]

A value of \(S = 1\) indicates that there are no cross-class links and that segregation is complete. Whenever \(m^l_{10} \neq 0\) we can guarantee the absence of complete segregation for any realisation of the considered \(D_G\).
5. Dataset Description

a. Facebook100

The Facebook100 dataset [22] contains an anonymized snapshot of the friendship connections among 1208316 users affiliated with the first 100 colleges admitted to Facebook. The dataset contains a total of 93969074 friendship links between users of the same college. Each node has a set of discrete-valued social attributes: status \{undergraduate, graduate student, summer student, faculty, staff, alumni\}, dorm, major, gender \{male, female\} and graduation year.

b. Wolf Dominance

The network represents a set of dominance relationships among a captive family of wolves [24]. Common signs of dominance among wolves are two low postures (namely low and low-on-back) and two behaviors (namely body tail wag and lick mouth [31]). In such a network a node corresponds to a wolf and a link exists if a wolf exhibited a low posture to another one. The network with \( n = 16 \) nodes and \( m = 148 \) links is provided with metadata such as age and sex.