Network homophily via tail inequalities

Nicola Apollonio [*]

Istituto per le Applicazioni del Calcolo, "Mauro Picone," Consiglio Nazionale delle Ricerche, Via dei Taurini 19, 00185 Rome, Italy

Paolo G. Franciosa [†]

Dipartimento di Scienze Statistiche, Università di Roma "La Sapienza," Piazzale Aldo Moro 5, 00185 Rome, Italy

Daniele Santoni [‡]

Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti," Consiglio Nazionale delle Ricerche, Via dei Taurini 19, 00185 Rome, Italy

Homophily is the principle whereby "similarity breeds connections." We give a quantitative formulation of this principle within networks. Given a network and a labeled partition of its vertices, the vector indexed by each class of the partition, whose entries are the number of edges of the subgraphs induced by the corresponding classes, is viewed as the observed outcome of the random vector described by picking labeled partitions at random among labeled partitions whose classes have the same cardinalities as the given one. This is the recently introduced *random coloring model* for network homophily. In this perspective, the value of any homophily score Θ, namely, a nondecreasing real-valued function in the sizes of subgraphs induced by the classes of the partition, evaluated at the observed outcome, can be thought of as the observed value of a random variable. Consequently, according to the score Θ, the input network is homophillic at the significance level α whenever the one-sided tail probability of observing a value of Θ at least as extreme as the observed one is smaller than α. Since, as we show, even approximating α is an NP-hard problem, we resort to classical tails inequality to bound α from above. These upper bounds, obtained by specializing Θ, yield a class of quantifiers of network homophily. Computing the upper bounds requires the knowledge of the covariance matrix of the random vector, which was not previously known within the random coloring model. In this paper we close this gap. Interestingly, the matrix depends on the input partition only through the cardinalities of its classes and depends on the network only through its degrees. Furthermore all the covariances have the same sign, and this sign is a graph invariant. Plugging this structure into the bounds yields a meaningful, easy to compute class of indices for measuring network homophily. As demonstrated in real-world network applications, these indices are effective and reliable, and may lead to discoveries that cannot be captured by the current state of the art.

I. INTRODUCTION

Network homophily is the phenomenon, first observed by social scientists, whereby people who share the same characteristics tend to be connected [1]. More abstractly, given a simple undirected graph G (modeling interactions) and a partition $\mathcal{P}$ of its vertices (modeling a classification based, for instance, on certain characteristics), G is said to be homophilic with respect to $\mathcal{P}$ if the subgraphs induced by the classes in the partition (i.e., people who presumably share the same characteristics) are significantly dense (thus showing a tendency to connect). What makes this definition vague is the inherent difficulty of explaining what "significantly dense" means. The main aim of this paper is to give a precise quantitative meaning to this term. A well-established descriptive approach to quantifying network homophily consists in defining a real-valued nondecreasing function Θ of the edge densities (or simply of the sizes) of the subgraphs induced by the classes of the given partition $\mathcal{P}$, and to locate the value of Θ evaluated at the input on a universal scale. The function Θ is called *homophily score* or *descriptive homophily index*. One of the simplest descriptive indices is the *homophily ratio* [2] (called *inside edge fraction* in [3]), which is simply defined as the fraction of the *homophillic edges*, namely, those edges that are induced by the classes of the partition. Also, by dividing the sum of the edge densities of the subgraphs induced by the classes by the number of classes yields another index in the scale [0,1]. Often the role of the universal scale is replaced by numerical experiments on benchmark instances while homophily indices are taken from a consolidated catalog. We refer the reader to Refs. [4,5] for (among other things) an annotated bibliography on both aspects.

In this paper we take an inferential statistics approach, namely, we think of the partitioned input network $(G, \mathcal{P})$ as the observed value of a random object in a probability space which induces a probability distribution on its measurable functions and, in particular, on homophily indices. The

---

[*]nicola.apollonio@cnr.it
[†]paolo.franciosa@uniroma1.it
[‡]daniele.santoni@iasi.cnr.it

latter ones are now random variables (statistics) of which we observe a particular realization. Depending on our ability to compute their distributions, a task which is very far from obvious (see the next section and Sec. V), we can perform hypothesis testing, locate confidence intervals, or, less ambitiously, simply compare the observed values of these statistics with their expected values under the null model represented by the chosen probability space for the random object $(G, \mathcal{P})$. We can think of the celebrated *network modularity index* by Newman [6] as probably the most notable instance of this approach. The probability space for $(G, \mathcal{P})$ consists of all multigraphs on the set of vertices of G with the same degrees as G sampled according to Bollobás' configuration model [7] (see next section). The statistic in this case is the random variable

$$\frac{1}{m} \sum_{i}^{s} [M_i - \mathbb{E}(M_i)],$$

where m is the size of G, M_i is the random number of edges induced by the ith class of the given partition $\mathcal{P}$ of $V(G)$ into s classes, and $\mathbb{E}(M_i)$ is its expected value under the configuration model. The modularity of G with respect to $\mathcal{P}$ is simply the value of this statistic in the observed data $(m_1, \ldots, m_s)$ where m_i is the number of observed homophilic edges induced by the vertices of class i, $i = 1, \ldots, s$. Since network modularity is a nondecreasing monotone function of the m_i's, besides its statistical interpretation, it directly quantifies homophily in the scale $[-1/2, 1]$. Actually, the range of the index could be a proper subinterval of $[-1/2, 1]$ whose extreme points correspond to the perfectly *antihomophilic* regime and to the perfectly homophilic regime, respectively. In this sense the scale $[-1/2, 1]$ is not universal. However, finding the maximum of modularity over the set of labeled partitions whose classes have prescribed cardinalities, and hence the width of the range of the index over this set, is an NP-complete problem, and the same is true for the homophily ratio (see theorem 5).

In this paper we take a complementary approach: we think of G as given while we think of $\mathcal{P}$ as a sample from the probability spaces consisting of all labeled partitions of $V(G)$ having the same structure of $\mathcal{P}$, equipped with the uniform distribution (see [8,9] for alternative probability measures on $\mathcal{P}$). This is the random coloring model introduced in [10]. The random coloring model is simple enough to derive exact closed formulas for the first two moments of the random variables M_i's, which are now defined on the probability space of the labeled partition of the same structure as $\mathcal{P}$ (see Sec. II). On these grounds, we propose a multidimensional extension of the well-known three-sigma rule of thumb by providing a Cantelli-type upper bound for the probability of the tails $[\Theta(M_1, \ldots, M_s) \geqslant \Theta(m_1, \ldots, m_s)]$ and $[\Theta(M_1, \ldots, M_s) \leqslant \Theta(m_1, \ldots, m_s)]$ where Θ is the homophily score. These bounds are coupled together and yield an index that ranges in $[-1, 1]$, which we take as a quantifier of network homophily. Moreover, its absolute value can be interpreted as an estimated significance of the deviation of the observed homophily from the expected one under the random coloring model. To illustrate the idea, consider the following extremely simplified problem: suppose we want to asses

whether a set C of c vertices of a graph G induces a subgraph H which is significantly denser among the subgraphs induced by the same number of vertices; we may think of the edge density y of H as a sample of the random variable Y described by the density of the subgraph induced by uniform sampling of sets of c vertices from G; by the results in [10], we known the expectation $\overline{Y}$ and the variance σ^2 of Y; hence we can compute the observed z score defined as $z = (y - \overline{Y})/\sigma$; the denser is H the higher is z. Well, but how much high z is? How can we assess that what we observe is statistically significant? If we knew the tail probability

$$\Pr\left\{\frac{Y - \mathbb{E}(Y)}{\sigma} \geqslant z\right\},$$

then we could immediately answer: if $y \leqslant \overline{Y}$, then H is not significantly dense at all, while the smaller is the tail probability above, the higher is the significance of the deviation from the expected value. However, already in this simplified problem, the probability above is even hard to approximate (see Sec. II A). Here comes to our aid a kind of three-sigma rule of thumb: let

$$Z = \frac{Y - \mathbb{E}(Y)}{\sigma}$$

be the Z score of Y, namely, the standardized version of Y; by Cantelli's inequalities ((7) and (8) in Sec. II A) the inequality

$$\Pr\{Z \geqslant z\} \leqslant \frac{1}{1 + z^2}$$

holds and tells us that, for $z \geqslant 0$, the function $z \mapsto \frac{z^2}{1+z^2} = 1 - \frac{1}{1+z^2}$ is not decreasing in the range $[0,1]$ and that the higher z, the higher its significance. By applying the same inequality to $-Z$ we obtain a symmetric bound for the hypo-dense behavior, namely, the significance of very large deviation of Y from $\overline{Y}$ in the opposite direction; similarly, for $z \leqslant 0$, the function $z \mapsto -\frac{z^2}{1+z^2}$ is not decreasing in the range $[-1, 0]$ and that the higher $|z|$ is, the higher is its significance. Coupling the bounds yields an index in the scale $[-1, 1]$ which ranges between the hypo-dense and hyper-dense case. The multidimensional extension of this idea is technically more difficult in that the bound crucially depends on the covariance matrix of the random vector M, which was not previously known within the random coloring model. In this paper we close this gap by computing the covariance matrix of subgraph sizes under the random coloring model. Interestingly, the matrix depends on the input partition only through its profile and on the network only through its degree sequence. Furthermore, all the covariances have the same sign, and this sign is a graph invariant.

The knowledge of the covariance structure of the random outcome M allow us to define a class of quantifiers of network homophily by specializing the function Θ. Although this leads to a variety of possibilities to measure network homophily (see Sec. IV), what we deem more interesting of our work is the methodological approach, which consists in viewing the input partitioned network, as a random object where randomness is inherited from random partitions rather than from random graphs. In principle, this is nothing more than a

reasonable modeling choice. However it has some important implications:

We provide an universal scale for our indices; in fact, their range is always (and tightly) the interval $[-1, 1]$, unlike the ranges of other indices, such as the modularity and homophily ratio indices, whose maxima are difficult to locate, the corresponding optimization problems being typically NP-complete (see Theorem 5).

Due to the simplicity of the random coloring model we gain the possibility of sharper analysis with respect to random graph models; for example, we have complete knowledge of the covariance structure of M, which can be exploited to estimate the significance of the observed homophily score via tail inequalities and hence to take this estimated significance as a homophily quantifier.

We discuss and substantiate these claims in Sec. V where we show, for instance, as expected, that the inherent homophilic nature of the *Homo sapiens* PPI network with respect to the location of corresponding genes on chromosomes is captured neither by the descriptive indices like the homophily ratio nor by Newman's modularity, while it is captured by ours.

The rest of the paper proceeds as follows. In Sec. II the general concept of homophily in random models is formally described introducing the ground of our approach to homophily: the random coloring model; in Sec. III we elicit the second-order moment structure of the random outcome M and study its properties including its computational complexity; these results are next exploited in Sec. IV where we propose a class of quantifiers of homophily and show that each of them can be efficiently computed; and in Sec. V we substantiate our approach through examples and real-world applications. Section VI closes the paper with concluding remarks.

II. HOMOPHILY IN RANDOM MODELS

Throughout the rest of the paper we use the following standard notation: for a positive integer k the symbol $[k]$ stands for the set $\{1, \ldots, k\}$ of the first k positive integers. The edge set of a graph G is denoted by $E(G)$. Symbols K_n, P_n, $K_{1,n}$, and $K_2^{(t)}$ denote the isomorphism classes of the complete graph on n vertices, the path on n vertices, the tree with $n + 1$ vertices and n leaves, and the graph consisting of t disjoint edges. The number of copies of P_3 in G is denoted by $\pi_3(G)$. Let G be a graph on $[n]$ with m edges. Let $\mathcal{P}$ be a partition of $[n]$ into s distinguishable classes. Let c_i be the number of vertices in class i, $i \in [s]$. The s-tuple $\mathbf{c} = (c_1, \ldots, c_s)$ is the *profile* of $\mathcal{P}$. Since labeled partitions $\mathcal{P}$ of $[n]$ of profile $\mathbf{c}$ correspond bijectively to surjective maps $f : V(G) \to [s]$ where $f^{-1}(i) = c_i$, for $i \in [s] := \{1, 2, \ldots, s\}$, we can equivalently think of the partitioned network $(G, \mathcal{P})$ as the colored network (G, f) where for each vertex of G, f tell us the color of that vertex. Accordingly, we rephrase the definition of homophily by saying *homophillic with respect to f* rather than *homophillic with respect to $\mathcal{P}$*, and we refer to $\mathbf{c}$ as *the profile of coloring f*. Also, we say that an edge of G is i-homophilic (or homophilic of color i) if both endpoints have color i. Let $(m_1, \ldots, m_s)'$—here and henceforth u' denotes the transpose of vector $u \in \mathbb{R}^d$, for some natural number d, usually thought

of as column vector—be the observed outcome for the input $(G, \mathcal{P})$, where m_i is the number of i-homophilic edges under f.

There are mainly three options to prescribe randomness for (G, f):

(a) We may think both G and f as random objects, namely, $\mathcal{G}_n$ is a set of graphs on $[n]$, $\Pi_{\mathbf{c}}$ is the set of colorings of $[n]$ of profile $\mathbf{c}$, and $\mathbb{P}_{\mathcal{G}_n, \mathbf{c}}$ is a probability measure on $\mathcal{G}_n \times \Pi_{\mathbf{c}}$.

(b) We may think of G as given while of the coloring f as a sample from $\Pi_{\mathbf{c}}$ according to a probability measure $\mathbb{P}_{\mathbf{c}}$ on $\Pi_{\mathbf{c}}$ ($\mathbb{P}_{\mathbf{c}}$ can therefore be viewed as conditional probability).

(c) We may think of f as given while the graph G as as sample from the random graph $\mathcal{G}_n$ according to a probability measure $\mathbb{P}_{\mathcal{G}_n}$ on $\mathcal{G}_n$ ($\mathbb{P}_{\mathcal{G}_n}$ can therefore be viewed as conditional probability).

Since models in (a) can be recovered by the knowledge of models in (b) and (c) via the factorizations

$$\mathbb{P}_{\mathcal{G}_n, \mathbf{c}}(G, f) = \mathbb{P}_{\mathcal{G}_n, \mathbf{c}}(G, f|G)\mathbb{P}_{\mathcal{G}_n}(G),$$
$$\mathbb{P}_{\mathcal{G}_n, \mathbf{c}}(G, f) = \mathbb{P}_{\mathcal{G}_n, \mathbf{c}}(G, f|f)\mathbb{P}_{\mathbf{c}}(f),$$

where we identify the marginals of $\mathbb{P}_{\mathcal{G}_n, \mathbf{c}}$ with $\mathbb{P}_{\mathcal{G}_n}$ and $\mathbb{P}_{\mathbf{c}}$, we may concentrate on the models in (b) and (c). Note that, when $\Pi_{\mathbf{c}}$ is the set of proper colorings of $[n]$, i.e., adjacent vertices have different colors, model in (a) is investigated in [11], and an optimization version is analyzed in [12].

Whichever model of (a), (b), and (c) applies,

$M = (M_1, \ldots, M_s)'$ is the *outcome* in the chosen probability space [of which we observe $(m_1, \ldots, m_s)$]; M_i is the random variable defined as the number of i-homophilic edges.

$\overline{M} = (\overline{M}_1, \ldots, \overline{M}_s)$ with $\overline{M}_i = \mathbb{E}(M_i)$, $i \in [s]$ and $\mathbb{E}(\cdot)$ being the *expectation* under the chosen model, is the the *expected* or *most typical* outcome even if, despite being typical, it might not belong to the support of M (for instance, because its coordinates might not be integer numbers).

$\sigma_i^2 = \mathbb{V}ar(M_i) = \mathbb{E}((M_i - \overline{M}_i)^2)$ is the variance of M_i, $\Sigma = \mathbb{E}((M - \overline{M})(M - \overline{M})')$ is the covariance matrix of M, and, provided that $\sigma_i > 0$ for all $i \in [s]$, $\Gamma = D^{-\frac{1}{2}}\Sigma D^{\frac{1}{2}}$ is the correlation matrix of M, where D is the diagonal matrix with diagonal entries $\sigma_1^2, \ldots, \sigma_s^2$.

$Z = (Z_1, \ldots, Z_s)$ and $z = (z_1, \ldots, z_s)$ are the z- and the observed z-score vectors, where, for $i \in [s]$, $Z_i = (M_i - \overline{M}_i)/\sigma_i$ and $z_i = (m_i - \overline{M}_i)/\sigma_i$; notice that the covariance matrix of Z is Γ.

Probably the most notable instance of the model in (c) is the celebrated *network modularity index* by Newman [6]; more generally, Newman studies from a quantitative viewpoint the broader phenomenon of *assortative mixing* of which network homophily can be seen as categorical manifestation [13]. Indeed, the modularity of G with respect to f in the sense of [6] is the index

$$\mathfrak{q}(f) = \sum_i^s \left[\frac{m_i}{m} - \left(\frac{D_i}{2m} \right)^2 \right], \tag{1}$$

where D_i is the sum of the degrees of the vertices in the ith color class. To see why the index above fits the model in (c) let M_i be the random number of i-homophilic edges in a (multi)graph G sampled according to Bollóbas' configuration model [7]—a random (multi)graph with the same degrees as G—and observe [6] that $(\frac{D_i}{2m})^2$ is the expectation of M_i/m

under this model. Therefore $\mathfrak{q}(f)$ takes the form

$$\frac{1}{m} \sum_i^s [m_i - \mathbb{E}(M_i)], \tag{2}$$

which expresses $\mathfrak{q}(f)$ as an appropriate multiple of the sum of the deviations of the observed homophillic edges in class i from their expected value under a particular null model (the configuration model in the present case). The reason why one resorts to probability spaces of multigraphs rather than simple graphs is because, unlike the Erdős-Rényi model $G(n, p)$, the edges of a simple random graph $\mathcal{G}(d)$ on $[n]$ with prescribed degree sequences d are correlated. In fact, even estimating the probability of a single edge can be challenging, and that makes analyzing $\mathcal{G}(d)$ difficult and more difficult than the case of multigraphs ([14]). However, in principle, the choice of any null model $(\mathcal{G}_n, \mathbb{P}_{\mathcal{G}_n})$ yields a modularity-like index having the same general structure (2).

A. The random coloring model

The models in (b) are in some sense complementary to those in (c). Choosing for $\mathbb{P}_\mathbf{c}$ the uniform measure we obtain the recently introduced *random coloring model* for network homophily [10] based on implicit ideas in [15]. The probability space consists thus of the set $\Phi_\mathbf{c}$ of all colorings of G of profile $\mathbf{c}$ equipped with the uniform measure

$$\mathbb{P}_\mathbf{c} = c_1! \cdots c_s!/n!. \tag{3}$$

This is the model we adopt in this paper. The random coloring model is simple enough to allow an exact closed formula for the first two moments of the random variables M_i's (see [10]). In fact, within this model there is a handy representation of the random variables M_i's by means of the Bernoulli random variables X_i^e, $e \in E(G)$, $i \in [s]$, defined as the indicators of the event: *edge e is i-homophilic*. Indeed,

$$M_i = \sum_{e \in E(G)} X_i^e,$$

so that we have the following expressions for the mean and the variance of M_i, $i = 1, \ldots, s$ ([10])

$$\overline{M_i} = m \frac{c_i^{\underline{2}}}{n^{\underline{2}}}, \tag{4}$$

$$\sigma_i^2 = m \frac{c_i^{\underline{2}}}{n^{\underline{2}}} \left(1 - m \frac{c_i^{\underline{2}}}{n^{\underline{2}}}\right) + 2\left[\left(\frac{c_i^{\underline{3}}}{n^{\underline{3}}} - \frac{c_i^{\underline{4}}}{n^{\underline{4}}}\right)\pi_3(G) + \frac{c_i^{\underline{4}}}{n^{\underline{4}}}\binom{m}{2}\right], \tag{5}$$

where, following the notation in [16], for a positive integer a and a nonnegative integer q, we denote by the symbol $a^{\underline{q}}$ the qth *falling factorial* of a, namely, $a^{\underline{q}} = a(a-1)\cdots(a-q+1)$, with $a^{\underline{0}} = 1$. Thus, $a^{\underline{2}} = a(a-1)$. After setting $\mathbf{c}^{\underline{2}} = (c_1^{\underline{2}}, \ldots, c_s^{\underline{2}})$, the most typical outcome in the random coloring model with parameters $\mathbf{c}$ and n reads as

$$\overline{M} = \frac{m}{n^{\underline{2}}} \mathbf{c}^{\underline{2}},$$

namely, the most typical outcome corresponds to a (possibly unattainable) coloring having the property that each color class induces a subgraph having the same density as input

graph. In [10] it is shown that both the marginal mean $\overline{M_i}$ and the marginal variance σ_i^2 can be computed in optimal $O(s + m)$ worst-case time. It is instructive to compute this vector for the complete graph on n vertices and a given profile $\mathbf{c}$. It is clear that, for any coloring $g \in \Phi_\mathbf{c}$, the number of edges in the subgraph induced by color i is exactly $c_i^{\underline{2}}/2$. Hence M is constant on $\Phi_\mathbf{c}$ and $\overline{M} = \mathbf{c}^{\underline{2}}/2$. On the other hand, the density of the complete graph is 1 while the coefficient of $\mathbf{c}^{\underline{2}}$ in the representation of $\overline{M}$ is half this density. Therefore, by applying the formula, we recover the same result. Using the corresponding formula, the reader may check that $\sigma_i^2 = 0$ for all $i \in [s]$. Notice that $\pi_3(K_n) = n^{\underline{3}}/2$.

Homophily in the random coloring model. The null hypothesis behind the random coloring model is the absence of homophily, namely, the coloring f does not correlate with G. This means that the effect of f on G, measured by some monotone function Θ of the number of homophilic edges, should be statistically the same as the effect of any other (equally likely) coloring of G with the same profile as f. We aim at devising an index, in the form of a nondecreasing function of the number of homophilic edges, such that

(i) On the one hand it locates the input (G, f) in the universal scale R (i.e., independent on the particular input graph) whose extreme values correspond to the antihomophilic and perfect homophilic behavior, respectively. Note that locating the range of the modularity value or homophily ratio index is a far from obvious problem (see Theorem 5), since it requires locating their maxima over the labeled partitions of the given profile.

(ii) On the other hand, values of the index are measures of the significance of the deviation of the number of homophilic edges from their expected value under the random coloring model.

Suppose for the moment that we know the distribution M induced by $\mathbb{P}_\mathbf{c}$. Consider the arithmetic mean of the z scores

$$A := A(Z_1, \ldots, Z_s) = \frac{1}{s} \sum_i^s Z_i.$$

This amounts to choose the statistic Θ as a function of $(M_1, \ldots, M_s)$ as

$$\Theta : \mathbb{R}^s \to \mathbb{R}, \quad x \mapsto \frac{1}{s} \sum_i^s \frac{x_i - \overline{M_i}}{\sigma_i},$$

which is trivially an affine nondecreasing function of its arguments. The index

$$\mathfrak{a}(G, f) = \begin{cases} 1 - \Pr\{A(Z) \geqslant A(z)\} & \text{if } A(z) \geqslant 0 \\ \Pr\{A(Z) \leqslant A(z)\} - 1 & \text{if } A(z) < 0 \end{cases} \tag{6}$$

ranges in $[-1, 1]$ and meets the requirements in (i) and (ii). Unfortunately, knowing the distribution of M is a very unlikely case. Indeed, a straightforward reduction to CLIQUE shows that given a certain positive integer κ even deciding whether the marginal tail probability $\Pr\{M_1 \geqslant \kappa\}$ is positive or not is an NP-complete problem: just choose for κ the number $c_1(c_1 - 1)/2$ where c_1 is the number of vertices of color 1: G has a clique on c_1 vertices if and only if $\Pr\{M_1 \geqslant \kappa\} > 0$. Therefore, since CLIQUE is not even in the APX class, it follows that the probabilities in (6) are even hard to approximate.

While this circumstance is very frustrating, not being able to calculate the probability of a given event exactly is a rather common event in probability and statistics. The usual remedy is to resort to so-called tail inequalities. The typical couple of one-sided tail inequalities for the random variable $\Psi(Z)$, where $\Psi : \mathbb{R}^s \to \mathbb{R}$, read as

$$\Pr\{\Psi(Z) \geqslant \Psi(z)\} \leqslant B(\Psi(z)) \quad \text{if } \Psi(z) > 0$$
$$\Pr\{\Psi(Z) \leqslant \Psi(z)\} \leqslant B(\Psi(z)) \quad \text{if } \Psi(z) < 0\text{'}$$

where B is a non-non-negative real-valued function of its argument which depends in general on the mixed moments of the random vector $Z = (Z_1, \ldots, Z_s)'$. For our purposes, the most interesting concentration inequality is the celebrated Cantelli's inequality. This inequality requires the knowledge of covariance matrix of M, which was not previously known to our best knowledge and for which we give a closed form expression (see Theorem 2) in the next section which depends on G only through its degrees and on f only through its profile. Let us illustrate how we can mimic index $\mathfrak{a}$ via tail inequalities, assuming the knowledge of the covariance structure of M. Consider classical univariate Cantelli's inequalities for a random variable X with finite expectation and finite variance σ^2. For a positive real number t, these inequalities are

$$\Pr\{X - \mathbb{E}(X) \geqslant t\} \leqslant \frac{\sigma^2}{t^2 + \sigma^2}, \tag{7}$$

$$\Pr\{X - \mathbb{E}(X) \leqslant -t\} \leqslant \frac{\sigma^2}{t^2 + \sigma^2}. \tag{8}$$

Notice the symmetry in the bounds which implies that both the left and the right tail are approximated in the same way. The more symmetric the distribution of $X - \mathbb{E}(X)$ is, the more effective these bounds are. Since $A(Z)$ is a centered univariate random variable, with finite variance

$$\mathbb{V}ar(A(Z)) = \mathbb{E}\left(\left[\frac{\mathbf{1}'Z}{s}\right]^2\right) = \mathbb{E}\left(\frac{\mathbf{1}'ZZ'\mathbf{1}}{s^2}\right) = \frac{\mathbf{1}'\mathbb{E}(ZZ')\mathbf{1}}{s^2}$$
$$= \frac{\mathbf{1}'\Gamma\mathbf{1}}{s^2},$$

where Γ is the covariance matrix of Z and $\mathbf{1}$ is the s-dimensional all-ones vector, the bounds above apply. Therefore, the index

$$a(G, f) = \text{sgn}[A(z)]\frac{[sA(z)]^2}{[sA(z)^2] + \mathbf{1}'\Gamma\mathbf{1}}, \tag{9}$$

where z is the observed z score and sgn is the signum function, ranges in $[-1, 1]$, and meets the requirements in (i) and (approximately) the requirement in (ii). In view of the hardness result, this is the best we can hope for by pursuing this approach.

III. COVARIANCE STRUCTURE OF SUBGRAPH DENSITIES

We saw in the previous section that the covariance matrix of the outcome of the random coloring model crucially enters the upper bounds on the probability of deviating from the most typical outcome and finally in the estimates of $\mathfrak{a}$. This section is devoted to the study of the covariance structure of

M, and this completes the knowledge about the second-order moments of M started in [10].

Theorem 1. Let (G, f) be the input graph in the random coloring model. Let G have n vertices and m edges and f have profile **c**. Then, for all $i, j, i \neq j$

$$\text{cov}(M_i, M_j) = c_i^2 c_j^2 \left\{ \frac{2}{n^{\underline{4}}}\left[\binom{m}{2} - \pi_3(G)\right] - \left(\frac{m}{n^{\underline{2}}}\right)^2 \right\},$$

which can be written as

$$\text{cov}(M_i, M_j) = \gamma(G)c_i^2 c_j^2, \tag{10}$$

where

$$\gamma(G) = \frac{2}{n^{\underline{4}}}\left[\binom{m}{2} - \pi_3(G)\right] - \left(\frac{m}{n^{\underline{2}}}\right)^2 \tag{11}$$

depends only on G through its density and the distribution of its degrees.

Proof. Let $i \in [s]$ be any color. Edge $e \in E(G)$ is i-monochromatic under a coloring g, if both end vertices of e have color i. Let X_i^e be the Bernoulli random variable defined as the indicator of the event: *edge e is i-monochromatic*. Recall from Sec. II A that for every color i, M_i has the representation $M_i = \sum_{e \in E(G)} X_i^e$. Let now $i, j \in [s]$ be two distinct colors. By definition $\text{cov}(M_i, M_j) = \mathbb{E}(M_i M_j) - \mathbb{E}(M_i)\mathbb{E}(M_j)$ and by (4) we already know $\mathbb{E}(M_i)\mathbb{E}(M_j)$. The rest of the proof follows by expanding $\mathbb{E}(M_i M_j)$:

$$\mathbb{E}(M_i M_j) = \mathbb{E}\left(\left[\sum_{e \in E(G)} X_i^e\right]\left[\sum_{e' \in E(G)} X_j^{e'}\right]\right).$$

By expanding the product inside the expectation, we obtain the polynomial expression

$$\sum_{(e, e') \in E(G) \times E(G)} X_i^e X_j^{e'}.$$

Since no i-monochromatic edge e can share a vertex with a j-monochromatic edge e', we have the relations

$$X_i^e X_j^{e'} = 0 \quad \forall (e, e') \in E(G) \times E(G) \quad \text{such that} \quad e \sim e',$$

where, for edges e and e' (with possibly $e = e'$) we write $e \sim e'$ if e and e' have at least a vertex in common. Therefore, by linearity

$$\mathbb{E}(M_i M_j) = \sum_{e \nsim e'} \mathbb{E}(X_i^e X_j^{e'}).$$

Now $\mathbb{E}(X_i^e X_j^{e'}) = \Pr\{X_i^e = 1, X_j^{e'} = 1\}$ because the variables involved are Bernoulli and, if $e \nsim e'$, then

$$\Pr\{X_i^e = 1, X_j^{e'} = 1\} = \frac{(n-4)!}{(c_i - 2)!(c_j - 2)!}\left(\frac{n!}{c_i! c_j!}\right)^{-1}$$
$$= \frac{c_i^2 c_j^2}{n^{\underline{4}}}.$$

Since $\Pr\{X_i^e = 1, X_j^{e'} = 1\}$ does not depend on the pair (e, e'), it follows that $\mathbb{E}(M_i M_j)$ is $\frac{c_i^2 c_j^2}{n^{\underline{4}}}$ times the number of ordered pairs (e, e') with no vertex in common and this number is twice $(= 2!)$ the difference between the number $\binom{m}{2}$ of unordered pairs of edges and the number of pairs sharing exactly

one vertex. Since the latter number is precisely the number $\pi_3(G)$ of paths of G with two edges we obtain

$$\mathbb{E}(M_i M_j) = 2 \frac{c_i^2 c_j^2}{n^{\underline{4}}} \left[\binom{m}{2} - \pi_3(G) \right].$$

In view of (5), after subtracting $\mathbb{E}(M_i)\mathbb{E}(M_j)$ from the expression above we obtain the required expression. Now (10) and (11) follow by trivial manipulations. ∎

Notice that for $i = j$, the expression above does not specialize to the variance given in (5). However, what is remarkable is that the covariances have the same sign and this sign depends only on G and not on the probability space. Therefore the sign of $\gamma(G)$ is a graph invariant. Interestingly we have the following theorem.

Theorem 2. Let $\mathcal{G}$ be a class of graph whose members have the same number of edges and vertices and hence the same edge density. Then $\gamma(G)$ is maximized on $\mathcal{G}$ by the graphs having a minimum possible sum of squares of the degrees, and it is minimized on $\mathcal{G}$ by those graphs having a maximum possible sum of squares of the degrees.

Proof. By (11), $\gamma(G)$ is maximized (minimized) over $\mathcal{G}$, by the graphs G for which which $\pi_3(G)$ is minimum (maximum). Let $d_G = (d_1, \ldots, d_n)$ be the degree sequence of a graph G and observe that $2\pi_3(G) = \sum_i d_i^2 + 2m$. Therefore G^* maximizes (minimizes) $\gamma(G)$ over $\mathcal{G}$ if and only if d_{G^*} minimizes (maximizes) $\sum_i d_i^2$ over the same set. ∎

As an example, let us consider the case of trees. Since among trees $\sum_i d_i^2$ is maximized by stars, while trees that minimize $\sum_i d_i^2$ are paths, we have explicitly

$$\gamma(K_{1,n}) = -\frac{1}{(n+1)^2} \quad \text{and} \quad \gamma(P_n) = \frac{1}{n^2(n-1)}.$$

We can also express $\gamma(G)$ in terms of the first two moments of the degree distribution of G. Let δ_1 and δ_2 be the first two moments of the degree distribution of G—recall that degree distribution of G is the function which associates each integer $t \in \{0, 1, \ldots, n-1\}$ with the fraction of vertices having t neighbors in G. One has $\delta_1 = 2m/n$ and δ_2 is the sum of the squares of the degrees divided by n. After simple but tedious manipulations it follows that

$$\gamma(G) = \frac{n}{n^{\underline{4}}} \left(\frac{2n-3}{2n-2} \delta_1^2 + \frac{1}{2} \delta_1 - \delta_2 \right);$$

therefore,

$$\gamma(G) \leqslant 0 \iff \frac{\delta_2 - \delta_1^2}{\delta_1} \geqslant \frac{1}{2} \left(1 - \frac{\delta_1}{n-1} \right).$$

After setting $\upsilon(G) = \frac{\delta_2 - \delta_1^2}{\delta_1}$ and denoting by $\rho(G)$ the edge density of G, the relation above reads as

$$\gamma(G) \leqslant 0 \iff \upsilon(G) \geqslant \frac{1 - \rho(G)}{2}$$

and relates the sign of the correlation of the marginal of M (or of Z) to the so-called *index of dispersion* of the degree distribution of G. This index is defined as the ratio between the variance and the mean of a given distribution. The index of dispersion of the degree distribution of G is thus precisely $\upsilon(G)$. Since for any graph G it holds that $0 \leqslant \frac{1-\rho(G)}{2} < 1$, it follows that values of the index larger than one qualify

overdispersed distributions while values of the index smaller than one qualify underdispersed distributions. Hence networks with overdispersed degree distribution yield negatively correlated marginals, while regular graphs yield positively correlated marginals.

Besides its interesting combinatorial meaning, invariant $\gamma(G)$ implies a strong linear algebraic structure for the covariance matrix of M, and this structure has striking algorithm consequences for computing Cantelli's bounds that our homophily quantifiers (to be defined in the next section) rely on. Recall that a *Z matrix* is a square matrix A over the real field whose off-diagonal elements are nonpositive and that an *M matrix* is a nonsingular Z matrix whose real eigenvalues are positive [17]. Clearly every symmetric positive definite Z matrix is an M matrix. M matrices are characterized in several different ways (see [17,18]). For our purposes the most interesting such characterization is the following.

Theorem 3. ([17]) A Z matrix is an M matrix if and only if it has a nonnegative inverse.

The following theorem essentially recognizes the rich structure of the covariance matrix of M, while the methodological implications are given in the next section.

Theorem 4. Let (G, f) be the input pair in the random coloring model with G and f as in Theorem 1.

(i) The covariance matrix Σ of the outcome M of the random coloring model $(\Phi_{\mathbf{c}}, \mathbb{P}_{\mathbf{c}})$ is of the form

$$\Sigma = Q + \gamma(G)\mathbf{c}^{\underline{2}}\mathbf{c}^{\underline{2}'},$$

where $Q = \text{diag}(q_1, \ldots, q_s)$ with

$$q_i = \sigma_i^2 - \gamma(G)\left(c_i^2\right)^2,$$

where σ_i^2 is given in (5) and $\gamma(G)$ in Theorem 1. Hence Σ is a rank 1 update of the diagonal matrix Q. If Q is invertible, so is Σ, and by the Sherman-Morrison formula, Σ^{-1} is of the form

$$\Sigma^{-1} = \text{diag}\left(q_1^{-1}, \ldots, q_s^{-1}\right) - \frac{\gamma(G)}{1 + \gamma(G)a'Qa}aa',$$

where $a = Q^{-1}\mathbf{c}^{\underline{2}}$.

(ii) If $\gamma(G) \leqslant 0$, then Σ_S is an M matrix for all principal minors Σ_S with $S \neq \emptyset$, while if $\gamma(G) \geqslant 0$, then $(\Sigma^{-1})_S$ is an M matrix for all principal minors $(\Sigma^{-1})_S$ with $S \neq \emptyset$.

The same conclusions in (i) and (ii) hold when Σ is replaced by the correlation matrix Γ.

Proof. Statement (i) follows straightforwardly by rewriting equations (5) and (10) in matrix form and then by applying the Sherman-Morrison formula [19]. As for the last statement, observe that if Σ is positive definite, then so is $D = \text{diag}(\sigma_1^2, \ldots, \sigma_s^2)$. Hence, replacing Σ by $D^{-\frac{1}{2}}\Sigma D^{-\frac{1}{2}}$ affects neither the sign of $\gamma(G)$ nor the property of being a rank 1 update of a diagonal matrix. It remains to prove Statement (ii). Both Σ and Σ^{-1} are positive definite matrices, and both are rank 1 updates of a diagonal matrix; namely, they are of the form $L + \lambda uu'$, where L is a diagonal nonsingular real matrix, λ is a real number, and $u \in \mathbb{R}_+^s$. Therefore, for every nonempty subset S of $[s]$ it holds that $(L + \lambda uu')_S = L_S + \lambda u_S u_S'$ and $u_S > 0$. Let $A_S = L_S + \lambda u_S u_S'$ It can be seen in various way (e.g., by the *Cauchy interlacing eigenvalues Theorem*, [18]) that A_S is a symmetric definite positive matrix

for every nonempty subset S of $[s]$. Therefore if $\lambda \leqslant 0$, then A_S is a positive definite Z matrix, and hence A_S is an M matrix for every nonempty subset S of $[s]$. ∎

In view of the theorem, finding Σ, Γ and their inverses can be accomplished very efficiently, in fact, in optimal $O(s^2 + m)$ worst-case time, where s is usually a small constant: computing $\gamma(G)$ requires $O(m)$, and due to the structure of the covariance matrix, $O(s^2)$ is needed to write Σ or Γ to perform inversion and vector multiplication.

Remark 1. Beside the sign, the absolute value of $\gamma(G)$ is also quite informative. Indeed, by (11), if G is sparse enough, like the graphs considered in Sec. V, $|\gamma(G)|$ is very small and the covariance is nearly zero. Consequently, M and Z have essentially uncorrelated marginals so that Σ is very close to the diagonal matrix D.

IV. MEASURING HOMOPHILY: A CLASS OF INDICES FOR NETWORK HOMOPHILY

We have already defined the homophily index $a(G, f)$ in Sec. II A. Since the index depends only on the observed z scores and the correlation matrix of M, after the results of the previous section, we see that $a(G, f)$ can be computed very efficiently. The rationale behind the definition of index a is that if G is homophilic with respect to f, then this behavior should be exhibited globally: either in most classes or in a few classes but in an extremely significant way. The same applies to antihomophily. We can therefore view the arithmetic mean $A(z)$ of the observed z scores as a *score* of the homophily of G with respect to f and $a(G, f)$ as a measure of the statistical significance of such a score. Moreover, as seen in Sec. II A, $A(z)$ is a real-valued nondecreasing affine function of the observed outcome $(m_1, \ldots, m_s)'$ so that, in the same vein, we can define and efficiently compute an entire class of indices simply by choosing a different score in the form of an affine nondecreasing real-valued function Θ. The recipe goes as follows. Let $w \in \mathbb{R}^s$ have nonnegative components and consider the linear map $\mathbb{R}^s \ni x \mapsto w'x \in \mathbb{R}$. This map clearly restricts to a measurable function of $\Phi_{\mathbf{c}}$ so that, by linearity, the score Θ defined by $\Theta(M) = w'(M - \overline{M})$, describes a centered random variable with covariance matrix

$$\Sigma_\Theta = w'\Sigma w, \qquad (12)$$

where Σ is the covariance matrix of M. Clearly Θ is a nondecreasing real-valued function in the coordinates of the observed outcome $(m_1, \ldots, m_s)'$. Therefore, by mimicking the construction of $a(G, f)$ and after setting $y = (m_1 - \overline{M_1}, \ldots, m_s - \overline{M_s})'$, we can associate with the score Θ the index $j_\Theta(G, f)$, varying in $[-1, 1]$, defined by

$$j_\Theta(G, f) = \text{sgn}[\Theta(y)] \frac{[\Theta(y)]^2}{[\Theta(y)]^2 + w'\Sigma w}. \qquad (13)$$

By construction, for any choice of the score Θ and hence of the weighting w, index $j_\Theta(G, f)$ meets the requirements in (i) and (approximately) the requirement in (ii) of Sec. II A. Index $j_\Theta(G, f)$ can be specialized in several reasonable ways. For instance, according to the the terminology of [3], typical choices for the score Θ could be based on one of the following:

The *edge-inside fraction*, also called *homophily ratio*, defined as m_i/m; this amounts to choosing $w = \mathbf{1}'/m$ so as to have a score in the range $[0,1]$.

The *average internal degree* defined as $2m_i/c_i$; this amounts to choosing w proportional to $2(\frac{1}{c_1}, \ldots, \frac{1}{c_s})'$ by a coefficient ν; possible choices for ν are the reciprocal of the maximum degree $\Delta(G)$ of G, the reciprocal of the number of color classes $\frac{1}{s}$, or even the reciprocal of the average degree $\overline{d(G)} = \frac{2m}{n}$.

The *internal density or dyadicity* defined as $2m_i/c_i^2$; this amounts to choosing w proportional to $2(\frac{1}{c_1^2}, \ldots, \frac{1}{c_s^2})'$ by the coefficient $\frac{1}{s}$ so to have a score in the range $[0,1]$.

We can also consider weighted version of the score $A(z)$ [which leads to index $a(G, f)$] simply by defining $\Theta(z) = \sum_{i=1}^s w_i z_i$ with $\sum_{i=1}^s w_i = 1$. As an illustration consider the homophily-ratio-based score, which we denote by $R(G, f)$. Hence $R(G, f) = \frac{1}{m} \sum_{i=1}(m_i - \overline{M_i})$ and $\mathfrak{r}(G, f) = \frac{1}{m} \sum_{i=1} m_i$ is the descriptive homophily ratio [2], which, in our model, is viewed as the observed sample of the random variable $\mathfrak{R} = \frac{1}{m} \sum_{i=1} M_i$. In view of (13), the corresponding homophily index $r(G, f)$ is defined by

$$r(G, f) = \text{sgn}(R) \frac{\left[\sum_i(m_i - \overline{M_i})\right]^2}{\left[\sum_i(m_i - \overline{M_i})\right]^2 + \mathbf{1}'\Sigma\mathbf{1}}$$

and gives an upper on the statistical significance of the tails $[\mathfrak{R} \geqslant \mathfrak{r}(G, f)]$ and $[\mathfrak{R} < \mathfrak{r}(G, f)]$ according to, respectively, if $R(G, f) \geqslant 0$ or $R(G, f) < 0$.

Besides the quantifiers defined above, there are other possibilities for a global assessing of homophily in the form of estimated significance of an appropriate distance from the expected outcome $\overline{M} = (\overline{M_1}, \ldots, \overline{M_s})'$ of the observed outcome $(m_1, \ldots, m_s')$ sampled from the random outcome $M = (M_1, \ldots, M_s)'$. The most celebrated and widely used such distance is the *Mahalanobis distance*. For a random vector $X = (X_1, \ldots X_s)'$ with expectation $\overline{X} = (\overline{X_1}, \ldots, \overline{X_s})'$ and definite positive covariance matrix Λ, the Mahalanobis distance of X from a particular point x in the support of X is the distance $d_M(X, x) = \sqrt{(x - \overline{X})'\Lambda^{-1}(x - \overline{X})}$. This distance actually measures how far the point x is from being typical, where typical means lying in an ellipsoid centered at $\overline{X}$ whose axes are the principal components of X. Notice that if X is a centered random vector whose marginals have unit variance, then $d_M(X, x) = \|x\|_{\Lambda^{-1}}$, i.e, the Mahalanobis distance of a standardized random vector from a point x in the support of its distribution coincides with the vector norm induced by the correlation matrix Λ^{-1}. In particular, if, as usual, Z and z are the z score and the observed z score, respectively, corresponding to the input colored network (G, f), then $d_M(Z, z) = \|z\|_{\Gamma^{-1}}$, where Γ is the correlation matrix of M. Furthermore, $d_M(Z, z)$ is the Euclidean norm of z whenever M has uncorrelated marginals. For simplicity, we denote $d_M(Z, z)$ simply by $\|z\|$ and refer to it as the Mahalanobis norm of z. While it is clear that the higher $\|z\|$ is, the less typical is the observed result $(m_1, \ldots, m_s)$, $\|z\|$ is not a score in the previously defined sense, since $\|z\|$ is not generally monotone even in the absolute values of the coordinates; i.e.,

it is not true that

$$|z_i'| \leqslant |z_i|, \ \forall i \in [s] \Rightarrow \|z'\| \leqslant \|z\|.$$

However, if G is sparse enough so that M has nearly uncorrelated marginals and $\|z\|$ is close to the Euclidean norm of z (recall Remark 1), or G is overdispersed so that Γ has nonnegative entries, then the above implication holds (at least approximately in the first case). In these cases, $\|z\|$ can be taken as a homophily score, at least in a weaker sense. Similar to the homophily indices defined above, a quantifier of the homophily of G with respect to f based on the weak score $\|z\|$ is obtained by estimating from above the significance of $\|z\|$ under the random coloring model. By a multidimensional version of Chebyshev's inequality ([20]) one has

$$\Pr\{\|Z\| \geqslant \|z\|\} \leqslant \frac{s}{\|z\|^2} = \frac{s}{z'\Gamma z},$$

therefore

$$h(G, f) = \max\left\{0, \frac{\|z\|^2 - s}{\|z\|^2}\right\}$$

can be taken as a quantifier of homophily in the interval [0,1]: values of h close to 1, indicate that the observed outcome is very far from its expected value along the directions most responsible for the variance structure of M. However, this index alone does not distinguish between antihomophilic and homophilic behavior, and the analysis must be refined by inspecting the sign pattern of z. What is true, however, is that if the value of h is positive and small, then the observed outcome is typical.

Given the results of the previous section, it is clear that both the indices $j_\Theta(G, f)$ (for any choice of weighting vector w) and $h(G, f)$ can be computed in optimal $O(s^2 + m)$ worst-case time.

V. DISCUSSION, EXAMPLES, AND APPLICATIONS

The uniqueness of our approach is to consider the input colored (partitioned) network as a random object, where the randomness is inherited from random colorings rather than random graphs, and the way we exploit this modeling choice via tail inequalities. While this is nothing more than a reasonable modeling choice, it has some profound implications:

(A) We gain more information about the significance of the observed outcome when comparing our indices with other consolidated homophily scores. In fact, the significance of the latter cannot generally be assessed per se but rather by the comparisons with other benchmark (possibly randomly generated) instances. Moreover, even locating the values of these scores on a universal scale by comparing them to their respective maxima is a far from trivial problem, the corresponding optimization problems being typically NP-complete (see Theorem 5 below); rather, our indices designed to provide a measure of their significance by default always (and tightly) sit in the interval $[-1, 1]$; in this sense their scale is universal.

(B) Due to the simplicity of the random coloring model, we gain the possibility of sharper analysis with respect to random graph models. Recall that even estimating the edge probabilities in random graph models, such as $\mathcal{G}(d)$, can be challenging [14]. In contrast, the representation of M as a

sum of identically distributed Bernoulli random variables allows even the exact computation of the second-order mixed moments of M, which allows us to define a whole class of homophily quantifiers, each based on a particular homophily score defined as an affine function of the data or even quantifiers based on the distance from the most typical outcome of the observed one.

Let us elaborate on these points with an example and real-world applications. Before doing so, we give a formal proof that maximizing the modularity and homophily ratio index over the labeled partition of a given profile are difficult problems. Let **c**-MODULARITY and **c**-HOMOPHILY RATIO be the problems of maximizing the modularity and the homophily ratio index, respectively, over the labeled partition of profile **c**.

Theorem 5. The decision versions of **c**-MODULARITY and **c**-HOMOPHILY RATIO are NP-Complete problems.

Proof. Instances of both problems are of the form $(G, \mathbf{c})$ together with the value p of modularity (respectively, homophily ratio) where G is an undirected graph of order n and **c** is an s-tuple of positive integers that add up to n, and we are asked to decide whether a partition of vertices with profile **c** exists giving value at least p for the index. Consider the subset of instances where G is cubic, i.e., each vertex has degree 3, and $\mathbf{c} = (\frac{n}{2}, \frac{n}{2})$. Note that n is even because G is a regular graph of odd degree. The set of feasible partitions is thus restricted to the set of partitions into two classes with the same number of vertices. In this case, given (1), q and r differ by an additive constant. Therefore both indices are maximized when the total number of edges induced by the two classes is maximized and, equivalently, when the number of edges having exactly one end in each of the two classes is minimized. The latter problem is MIN BISECTION ON CUBIC GRAPHS, whose decision version is known to be NP-Complete [21]. This proves that **c**-MODULARITY and **c**-HOMOPHILY RATIO are hard by restriction, so they are hard in general. ∎

A. A toy example

This nice and meaningful example was suggested by an anonymous referee to highlight the difference between modularity and our indices based on tail inequalities. Consider a graph G of order n consisting of m disjoint edges where m is an even number. Hence $n = 2m$ and the degree sequence of G is the $2m$-dimensional all-ones vector **1**. Let us analyze the indices over all colorings f of profile (m, m) or, equivalently, over all partitions $\mathcal{P}$ of $[n]$ into m red vertices and m blue vertices. In this case there is no ambiguity in establishing whether an instance (G, f) is more homophilic than another instance (G', f'), where $G = G'$ in the random coloring model, and $f = f'$ in the configuration model. The only problem here is how to assess whether or not the given instance is homophilic *per se*. Let M_r and M_b be the random number of homophillic red and blue edges, respectively under the configuration model and the random coloring model. We use the same symbols for random variables on different probability spaces to facilitate comparisons. A random pairing of the vertices of G is called a (simple) configuration. Let $\mathcal{G}(\mathbf{1})$ be the set of graphs on $[n]$ with degree sequence **1**, and

let us equip this set with the uniform distribution (this yields the configuration model in this case). All graphs in $\mathcal{G}(\mathbf{1})$ are simple, and the probability of seeing a particular configuration by uniform sampling is $\frac{m!2^m}{2m!}$. For integers h and k such that $0 \leqslant h \leqslant \frac{m}{2}$ and $0 \leqslant k \leqslant \frac{m}{2}$, the event $(M_r = h, M_b = k)$ has probability 0 in both probability spaces whenever $h \neq k$. This because edges whose end vertices have different colors match the same number of vertices in both color classes. Therefore the joint distribution of (M_r, M_b) is supported by the diagonal $\{(k, k), 0 \leqslant k \leqslant m/2\}$ in both models. Hence

$$\Pr\{M_1 \geqslant k | \text{conf}\} = \Pr\{M_2 \geqslant k | \text{conf}\}$$
$$= \Pr\{M_1 \geqslant k, M_2 \geqslant k | \text{conf}\}$$

and

$$\Pr\{M_1 \geqslant k | \text{rand-col}\} = \Pr\{M_2 \geqslant k | \text{rand-col}\}$$
$$= \Pr\{M_1 \geqslant k, M_2 \geqslant k | \text{rand-col}\}.$$

Actually, it can be shown more, namely,

$$\Pr\{M_i \geqslant k | \text{conf}\} = \Pr\{M_i \geqslant k | \text{rand-col}\}$$
$$= \frac{2^m (m!)^2}{(2m)!} \sum_{t=k}^{m/2} \binom{m}{t, t} 2^{-2t}, \quad i \in \{b, r\},$$

$$(14)$$

where $\binom{m}{t, t}$ is the multinomial coefficient of m with parts $t, t, m - 2t$. That is, even though the random variables are defined on different probability spaces, their tails have the same measure so that we can omit the reference to the probability space and denote by $F(k)$ the probability $\Pr\{M_i < k\} = 1 - \Pr\{M_i \geqslant k\}$ where $\Pr\{M_i \geqslant k\}$ is given by (14). Either by the relation above or by direct computation, it follows that

$$\mathbb{E}_{\text{conf}}(M_r/m) = \mathbb{E}_{\text{conf}}(M_b/m) = \mathbb{E}_{\text{rand-col}}(M_r/m)$$
$$= \mathbb{E}_{\text{rand-col}}(M_b/m) = \frac{m(m-1)}{2m(2m-1)}$$
$$= \frac{m-1}{2(2m-1)}.$$

In fact, since each configuration has probability $\frac{m!2^m}{2m!}$, the probability of seeing an edge connecting two particular vertices is $\frac{1}{2m-1}$ rather than $\frac{1}{2m}$. However, the difference between $\frac{1}{2m-1}$ and $\frac{1}{2m}$ is negligible. In fact, $\frac{m-1}{2(2m-1)}$ is close to $\frac{1}{4}$ when m is large. We can thus stick to the definition of $\mathfrak{q}$ given in Sec. II. Therefore, by (2), the modularity of any observed partition $\mathcal{P}(k)$ with k red edges and k blue edges—recall that no partition with a different number of red and blue homophilic edges can be observed—is

$$\mathfrak{q}(\mathcal{P}) = 2\left(\frac{k}{m} - \frac{1}{4}\right).$$

In conclusion, Newman's modularity interpolates linearly between $-1/2$, corresponding to the perfectly antihomophilic case, and $1/2$, corresponding the perfectly homophilic case. Clearly the same result would have been obtained by taking expectations with respect to the random coloring model. In this particular case, however, the analysis provided by these indices is not sharper than what the simple homophily ratio
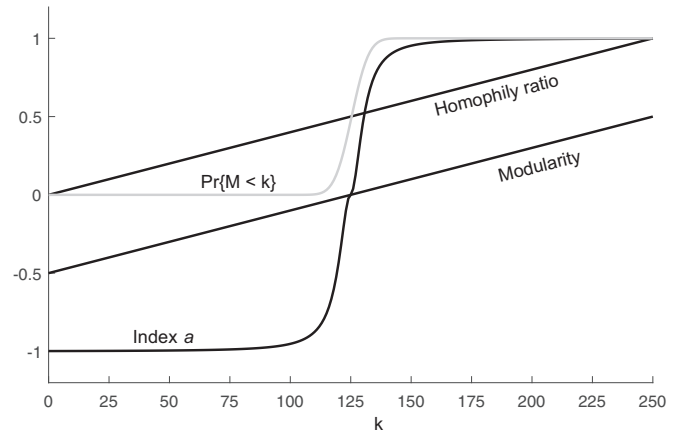

FIG. 1. Theoretical probability $\Pr\{M < k\}$ (in gray) for the matching example in Sec. V A compared against homophily ratio, modularity, and index a. Values are reported as a function of k. Experiments refer to an instance with 1000 vertices, 500 edges, and coloring profile (500, 500).

$\mathfrak{r}$ provides: $\mathfrak{r}(\mathcal{P}(k)) = \mathfrak{q}(\mathcal{P}(k)) + 1/2$. On the other hand, by looking at Fig. 1, one can appreciate the information gain provided by the indices based on concentration inequalities. The figure refers to a matching with 1000 vertices, 500 edges, and coloring profile (500,500). Since in every coloring of profile (500,500) the number of red edges is the same as the number of blue edges, all the indices $\mathfrak{q}$, $\mathfrak{r}$, and a are actually real-valued functions defined on the integers between 0 and 250. For $\mathfrak{q}$ and $\mathfrak{r}$ this is clear by their definition, while for a it suffices to observe that the arithmetic mean of the observed z scores corresponding to the observed outcome (k, k) coincides with the marginal z score of the red edges (or, equivalently, of the blue edges). In particular, the plot of $\mathfrak{q}$ and $\mathfrak{r}$ are (parallel) straight lines, while a is of the form $\frac{z^2(k)}{1+z^2(k)}$ with $z(k)$ being the marginal z score corresponding to k red edges. This z score is itself an affine function of k. In the figure we plot the corresponding values, as k runs between 0 and 250, against the plot of the (exact) probability $F(k)$.

By inspecting the plots of $F(k)$ it is clear that values of k [and hence the observed outcomes of the form (k, k)] outside the interval [110,140] are extremely unlikely, meaning antihomophily (left tail) or homophily (right tail). Neither $\mathfrak{q}$ nor $\mathfrak{r}$ can detect this phase transition, while index a clearly does. In conclusion, the situation is much the same as deciding whether the number of heads in a long series of coin tosses is consistent with the assumption that the coin is fair: $\mathfrak{q}$ and $\mathfrak{r}$ would count only the fraction of heads, while a would estimate the likelihood of the sequence.

B. Measuring homophily on real networks

If we want to decide whether a particular categorical property correlates with the structure of the network, we simply encode the property as a coloring and apply one of the proposed indices. Here we concentrate on $a(G, f)$ and $r(G, f)$. To illustrate this application we implemented a prototype function in Python 3 for computing these indices and run it on five instances of the form (G, f). In the first four

TABLE I. Indicators for the five considered instances.

Network	Vertices	Edges	$1 - r(G, f)$	$[1 - a(G, f)]10^6$	$\mathfrak{r}(G, f)$	$\mathfrak{q}(G, f)$
Treponema pallidum	894	8157	0.0014	160.8	0.545	0.3538
Escherichia coli	4020	29 748	0.0015	11.1	0.507	0.3332
Saccharomyces cerevisiae	6157	119 051	0.1233	18.0	0.498	0.2138
Homo sapiens	27 342	777 098	0.0025	1007.3	0.049	−0.0011
Pokec	1 212 349	8 320 600	1.6582e−06	3.3	0.466	0.1629

instances G is a Protein Protein Interaction (PPI) network: each vertex of the network represents a protein, while edges represent interactions between proteins in different organisms. More specifically, for the first three instances, G is one of *Treponema pallidum*, *Escherichia coli*, and *Saccharomyces cerevisiae*—retrieved from STRING database [22]—while in the fourth instance, G is the *Homo sapiens* PPI network. In the last instance, G is the network *Pokec*, which describes symmetric interactions among users of a popular Slovak social network. The coloring f in the first three instances associates each vertex with its functional class according to the standard taxonomy of the NCBI database [23]. As explained in detail in [10], the range of f consists of 19 colors rather than 24. In the fourth instance, f associates each protein with the chromosome the related gene belongs to. Finally, in the last instance, f associates each user with one out of the five classes of (declared) age. For biological reasons, the first four networks are inherently homophilic with respect to the chosen colorings, and there are reasons to expect that the same is true for the social network. The results of our experiments, summarized in Table I, confirm these expectations. The second and the third columns of Table I correspond to the number of vertices n and the number of edges m of G, respectively; the other columns of the table report the values of the indices used in the comparison: the random homophily ratio-based index $r(G, f)$, the arithmetic mean of the z-score-based index $a(G, f)$, the descriptive homophily ratio $\mathfrak{r}(G, f)$, and Newman's modularity $\mathfrak{q}(G, f)$. Actually, the values of $r(G, f)$ and $a(G, f)$ are complemented to 1 to give an immediate perception of the significance of the value of the corresponding statistic. For example, the last row of the fourth column of the table reads as follows: the probability of observing, by chance, a coloring with a higher homophily ratio than the observed one is less than 1 in half a million, that is, much less than the significance level typically set at 5%. Since locating the maximum of $\mathfrak{r}$ and $\mathfrak{q}$ is a difficult optimization problem [see Theorem 5 and recall (A)], the significance of these indices cannot be assessed by comparisons to their respective maxima but rather by a comparison with the results in Table II, where we reported the values of all of the indices averaged over five randomly generated instances $(G, f_1), \ldots (G, f_5)$ where G is kept fixed and the f_i's are picked independently at random among the colorings of G with the same profile as the input coloring f. Even without knowing the actual range of $\mathfrak{r}$ and $\mathfrak{q}$, it can be seen from Table II that all four considered indices behave as expected, i.e., none of them deviates significantly from its expected reference value. Also the comparison between the last column of the two tables shows that the index $\mathfrak{q}$ (as well as $\mathfrak{r}$) suggests homophily with respect to the corresponding coloring of all five networks except *Homo sapiens*. This network

is not considered to be significantly homophilic w.r.t. the given coloring by both the modularity index and the descriptive homophily ratio, while it is considered homophobic by both of our indices, consistent with the underlying biological reasons. Also note that although n, m span a wide range and the edge density differs by up to three orders of magnitude, our indices assume fairly comparable values, confirming the consistency of our method. The time needed to compute our indices is negligible with respect to the I/O time for reading the graphs, confirming that both the $a(G, f)$ and $r(G, f)$ indices can be computed in linear time without using any special data structure, provided that the number of colors is $O(\sqrt{m})$ (which is not restrictive in concrete applications).

VI. CONCLUSION

We proposed an inferential statistical approach, based on second-order tail inequalities, to quantify network homophily for an input $(G, \mathcal{P})$ consisting of a undirected graph G and a given partition of its vertex set. After adhering to the accepted paradigm that whatever network homophily means, it is measured by a nondecreasing function of the edge densities of the subgraphs induced by the classes of $\mathcal{P}$, we proposed to look at any such function, referred to as a *descriptive homophily index*, as a random variable defined on the probability space of the labeled partitions of the same profile as $\mathcal{P}$ endowed with the uniform measure. This is the random coloring model introduced in [10] and formally refined in the present paper. Within the random coloring model, values of descriptive indices are just observed samples from the corresponding random variables described by picking labeled partitions at random. In light of this view, we proposed to quantify network homophily as the (signed) statistical significance (p value) of the observed value under the null hypothesis of the absence of homophily. The p value of the observed value is the probability of the appropriate (right or left) tail of the corresponding random variable, namely, the probability of

TABLE II. Indicators for the randomized coloring of the considered instances, averaged each on five random colorings.

Network	$1 - r(G, f)$	$[1 - a(G, f)]10^6$	$\mathfrak{r}(G, f)$	$\mathfrak{q}(G, f)$
Treponema pallidum	0.6715	457 697	0.161	−0.00295
Escherichia coli	0.6487	499 182	0.467	0.00028
Saccharomyces cerevisiae	0.7734	539 713	0.282	−0.00013
Homo sapiens	0.5784	709 707	0.107	−0.00024
Pokec	0.5720	770 348	0.257	0.00003

observing a more extreme value than the observed one. We showed that the problem of computing the p value is hard to approximate, since the corresponding optimization problem is not even in the APX class. However, we gave upper bounds on this p value by mimicking the three-sigma rule of thumb, namely, by resorting to tail inequalities such as Chebyshev's and Cantelli's inequalities. While on the one hand we have no control on the quality of the approximation of the p value via these inequalities, on the other hand, in view of the inapproximability result, this is the best we can hope for unless $P = NP$. Estimating the p value via tail inequalities requires the knowledge of covariance matrix of the random vector M, which was not previously known within the random coloring model. Due the simplicity of the model, in this paper we closed this gap by giving a closed form expression for the covariance matrix of M. In particular, we discovered that the correlation among the marginals of the random vector M is ruled by a parameter $\gamma(G)$ whose sign is a graph property [hence $\gamma(G)$ is a graph invariant] and whose absolute value dictates the intensity of the correlations; for sparse enough graphs, the marginals are nearly uncorrelated. Interestingly, this parameter depends on the input partition only through its profile and depends on the network only through its degree sequence, implying that both the covariance matrix and its inverse can be computed very efficiently leading to an overall complexity of $O(s^2 + m)$ worst-case time for computing all the proposed indices. Our indices, which range in interval $[-1, 1]$ for any graph, supply some gain of information when compared to other well-established indices. The situation is much the same as deciding whether the number of heads in a long series of coin tosses is consistent with the assumption that the coin is fair. For this purpose, while classical indices count only the fraction of heads, our indices estimate the likelihood of the sequence. This is confirmed by the case of the *Homo sapiens* PPI network with respect to the location of corresponding genes on chromosomes. For biological reasons, the *Homo sapiens* PPI network is inherently homophilic with respect to the given partition. However, neither Newman's modularity nor the homophily ratio (nor other descriptive indices) is able to detect this property, whereas ours are.

[1] M. McPherson, L. Smith-Lovin, and J. M. Cook, Annu. Rev. Sociol. **27**, 415 (2001).

[2] J. Zhu, Y. Yan, L. Zhao, M. Heimann, L. Akoglu, and D. Koutra, NeurIPS. **33**, 7793 (2020).

[3] T. Chakraborty, A. Dalmia, A. Mukherjee, and N. Ganguly, ACM Comput. Surv. **50**, 54 (2017).

[4] J. Yang and J. Leskove, J. Knowl. Inf. Syst. **42**, 181 (2015).

[5] A. Lancichinetti, A. Kivelä, J. Saramäki, and S. Fortunato, PLoS ONE **5**, e11976 (2010).

[6] M. Newman, Proc. Natl. Acad. Sci. USA **103**, 8577 (2006).

[7] B. Bollobás, Eur. J. Comb. **1**, 311 (1980).

[8] A. J. Gates and Y. Ahn, J. Mach. Learn. Res. **18**, 1 (2017).

[9] M. Hoffman, P. Block, and T. A. B. Snijders, Sociol. Method. **53**, 1 (2023).

[10] N. Apollonio, P. G. Franciosa, and D. Santoni, Sci. Rep. **12**, 9757 (2022).

[11] M. Dyer and A. Frieze, Random Struct. Alg. **36**, 251 (2010).

[12] L. Zdeborová and F. Krząkała, Phys. Rev. E **76**, 031131 (2007).

[13] M. E. J. Newman, Phys. Rev. E **67**, 026126 (2003).

[14] P. Gao and Y. Ohapkin, Random Struct. Algorithms **62**, 911 (2023).

[15] J. Park and A. L. Barabasi, Proc. Natl. Acad. Sci. USA **104**, 17916 (2007).

[16] D. Knuth, *The Art of Computer Programming, Vol. 1: Fundamental Algorithms*, 3rd ed. (Addison-Wesley, Reading, MA, 2017).

[17] M. Fiedler, *Special Matrices and Their Applications in Numerical Mathematics* (Martinus Nijhoff, Dordrecht, 1986).

[18] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis* (Cambridge University Press, Cambridge, 1991).

[19] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. (Johns Hopkins University Press, Baltimore, 1996).

[20] X. Chen, arXiv:0707.0805 vl; second version: arXiv:0707.0805 v2.

[21] T. Bui, S. Chaudhuri, T. Leighton, and M. Sipser, Combinatorica **7**, 171 (1987).

[22] https://string-db.org.

[23] ftp://ftp.ncbi.nih.gov/pub/COG/COG/.