scientific reports

OPEN

A novel method for assessing and measuring homophily in networks through second-order statistics

Nicola Apollonio[1], Paolo G. Franciosa[2]✉ & Daniele Santoni[3]

We present a new method for assessing and measuring homophily in networks whose nodes have categorical attributes, namely when the nodes of networks come partitioned into classes (colors). We probe this method in two different classes of networks: (i) protein–protein interaction (PPI) networks, where nodes correspond to proteins, partitioned according to their functional role, and edges represent functional interactions between proteins (ii) Pokec on-line social network, where nodes correspond to users, partitioned according to their age, and edges respresent friendship between users.Similarly to other classical and well consolidated approaches, our method compares the relative edge density of the subgraphs induced by each class with the corresponding expected relative edge density under a null model. The novelty of our approach consists in prescribing an endogenous null model, namely, the sample space of the null model is built on the input network itself. This allows us to give exact explicit expression for the z-score of the relative edge density of each class as well as other related statistics. The z-scores directly quantify the statistical significance of the observed homophily via Čebyšëv inequality. The expression of each z-score is entered by the network structure through basic combinatorial invariant such as the number of subgraphs with two spanning edges. Each z-score is computed in $O(n + m)$ time for a network with n nodes and m edges. This leads to an overall efficient computational method for assesing homophily. We complement the analysis of homophily/heterophily by considering z-scores of the number of isolated nodes in the subgraphs induced by each class, that are computed in $O(nm)$ time. Theoretical results are then exploited to show that, as expected, both the analyzed network classes are significantly homophilic with respect to the considered node properties.

The *homophily principle* states that "similarity breeds connections"[1]. This principle—born in sociology—once declined into Network Theory, reads as *nodes in a network are more likely to be linked to nodes sharing similar attributes*. The effectiveness of homophily in social networks has been extensively demonstrated across various instances[2–7]: social networks exhibit homophily with respect to attributes such as gender, age, ethnicity, occupation, social class and many others. This simply means that people preferentially interact with people sharing the same cultural and sociological attributes. Putting it succinctely: "birds of a feather flock together"[1,8]. In contrast, heterophilic networks are those networks whose nodes preferentially interact with nodes having different attributes values. Homophily can also be seen as the categorical counterpart of "assortative mixing"—the correlation of attributes across link—and, as such, at least beyond a certain amount of assortativity, it binds the structure of networks[9], and influences the curvature of the cumulative degree distribution under the preferential attachment evolutionary mechanism[10]. In view of this discussion, homophily qualifies as a genuine network property, namely, a property that when possessed to some extent, impacts non trivially on the structure of the network. A quantitative understanding of homophily in networks is therefore useful both from a theoretical and a practical point of view. A step forward in this direction is taken once we realize that homophily in networks certainly fits in the frame of "community detection"[11,12]; observe that communities in complex networks identify high order homogeneous structures. Arguing as in[13], network community detection can be seen as a procedure consisting of two

[1]Istituto per le Applicazioni del Calcolo "Mauro Picone", Consiglio Nazionale delle Ricerche, Via dei Taurini 19, 00185 Rome, Italy. [2]Dipartimento di Scienze Statistiche, Università di Roma "La Sapienza", piazzale Aldo Moro 5, 00185 Rome, Italy. [3]Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", Consiglio Nazionale delle Ricerche, Via dei Taurini 19, 00185 Rome, Italy. ✉email: paolo.franciosa@uniroma1.it

stages: one stage consists of extracting communities by relying on the geometric structure of the networks, while the second stage consists in "abstracting" communities, namely, in identifying the common features and attributes of community members. Such common features and attributes are usually referred to as *functions* or *node characteristic*[14]. From this perspective, communities are first detected based on their geometry and then evaluated based on their functions. The other way round is also meaningful: given a functional description of a network, namely a partition of its nodes into sets of nodes with the same node characteristic, assessing whether or not the class of the partition have a certain amount of geometric structure, i.e. assessing whether or not such classes are communities, is tantamount to assessing whether or not the network is homophilic with respect to node characteristic. Innocent as it may seem, this observation already provides a way of quantifying homophily: if the functional description correlated with the geometry of the network or, equivalently, if the network were homophilic with respect to node characteristics, then the node-induced subgraphs of each class of the partition should be relatively denser than what we expect under some suitable null model—the relative edge density of a subgraph H of a given graph G is the density of H over the density of G. Newman's celebrated *modularity* index[15] formulates the null hypothesis as the relative expected density of a random graph with the same degree distribution as the input graph. Modularity is thus the relative edge density of monochromatic subgraphs minus the expected relative edge density of monochromatic subgraphs under the null hypothesis that edges are distributed at random among nodes. Since the index lies in interval $[-\frac{1}{2}, 1]$, its value directly quantifies network homophily: the larger the index the more homophilic the network is. Modularity thus provides a scale for comparing homophily of different networks. Notice that Newman's index resorts to an exogenous model (the *configuration model*) to test homophily. In this paper, by revising the idea in[14], we propose to measure network homophily by testing the observed structure (i.e. the relative edge density of functional classes) against the expected structure under an endogeneous random model: the input graph itself will be the sample space for the null hypothesis. With this aim in mind, in this paper we propose a new statistical model that builds on the approach in[14] (developed for networks with only two functional classes), extend it to an arbitrary number of classes, and strengthen it by exploiting second order statistics based on a uniformly random coloring of the input network with the same color distribution. This machinery yields an explicit exact formula for the $\mathbf{z}$-score of a suitable defined homophily index as well as of the number of isolated nodes of each functional class. The statistical significance of the observed homophily is then obtained through Čebyšëv inequality. As one may expect, the structure of the network enters second order statistics through the number of its subgraphs with two spanning edges, namely, the number of its P_3's (if the two edges are adjacent) and the number of its $2K_2$'s (if the two edges are not adjacent). This means that our analysis does not require exogenous models (random graphs, for instance) to make comparisons for assessing homophily. Throughout the rest of the paper P_3 is the graph on three nodes joined by two edges, namely the graph $\bullet - \bullet - \bullet$, while $2K_2$ is the graph on four nodes with two edges without common endpoints, namely, the graph $\begin{smallmatrix}\bullet - \bullet\\ \bullet - \bullet\end{smallmatrix}$.

In line with the work of[14], we probe our theoretical results on two different network classes: (i) Protein–Protein Interaction (PPI) networks, where nodes correspond to proteins, partitioned according to their functional role, and edges represent functional interactions between proteins (ii) on-line social network where nodes correspond to users, partitioned according to their age, and edges represent friendship between users. As expected, numerical results provide strong evidence of the homophilic nature of the considered networks with respect to the corresponding node properties: protein function for PPI and age class for social network.

Homophily in networks

As sketched in "Introduction" section, we look at homophily as a network parameter (actually as an array of parameters, see "Assessing and measuring homophily" section) measuring to what extent the attributes (node characteristics, functions) of the nodes of the networks correlate across the edges. To give a precise meaning to such a correlation, we follow the approach in[14] which we now discuss in more details. Of course, nothing bad is happening if we think of node characteristics as node colors and, consistently, of the functional description as a partition of the node set into color classes so that (potential) communities are sets of nodes with the same color. Consequently, we deal with a simple undirected graph G with n nodes and m edges whose nodes are partitioned into a number s of color classes. The simple original model in[14] refers to the case of two colors ($s = 2$) denoted by 0 and 1. Edges of G are then classified as $(0, 0)$-edges, $(0, 1)$-edges and $(1, 1)$-edges according to the color at their endpoints. Let c_0 (resp., c_1) be the number of nodes of G having color 0 (resp., 1), with $c_0 + c_1 = n$; furthermore, let $m_{i,j}$ be the number of (i, j)-edges, $i, j \in \{0, 1\}$, with $m_{0,0} + m_{0,1} + m_{1,1} = m$: if the functional definition of the communities correlated with the structure of G, then we should expect a statistical significant deviation between $m_{0,0}$, say, and what we would expect if characteristic 0 were randomly distributed among the nodes of the graph, namely, if any node had an equal chance of possessing it. In[14], it is proposed to measure this deviation by the three ratios:

$$\omega_0 = \frac{m_{0,0}}{\overline{m}_{0,0}}, \quad \eta_{0,1} = \frac{m_{0,1}}{\overline{m}_{0,1}}, \quad \omega_1 = \frac{m_{1,1}}{\overline{m}_{1,1}}$$

where, for $i, j \in \{0, 1\}$ and $i \neq j$

$$\overline{m}_{i,i} = m\frac{c_i(c_i - 1)}{n(n-1)} \quad \text{and} \quad \overline{m}_{i,j} = m\frac{2c_ic_j}{n(n-1)},$$

are the expected number of (i, i)-edges and (i, j)-edges, respectively, under the hypothesis that properties 0 and 1 are randomly distributed over the node set of G (see "Homophily, heterophily and isolated nodes: first and second order moments" section for proofs). Just by rewriting ω_i and $\eta_{i,j}$ as

$$\omega_i = \frac{2m_{i,i}}{c_i(c_i-1)} \Big/ \frac{2m}{n(n-1)} \quad \text{and} \quad \eta_{i,j} = \frac{m_{i,j}}{c_i c_j} \Big/ \frac{2m}{n(n-1)} \tag{1}$$

one sees that ω_i is nothing but the normalized intracommunity density; analogously, $\eta_{i,j}$ is the normalized intercommunity density[13]. In this perspective, homophily (and heterophily) provides a suggestive interpretation of basic structural graph properties (those that can be captured by first order moments of functions of random partitions into two classes with c_0 nodes labeled 0 and c_1 nodes labeled 1). In this simple model, graph G is declared i-homophilic (or homophilic with respect to property i), $i \in \{0, 1\}$, if $\frac{m_{i,i}}{\overline{\overline{m}}_{i,i}} > 1$; graph G is declared (i, j)-heterophilic if $\frac{m_{i,j}}{\overline{\overline{m}}_{i,j}} > 1$ (we ask the reader to bear the pedantic reference to the indices i, j in view of the generalization to more than 2 properties). Without any other clue about the likelihood or the variability of ω_i and $\eta_{i,j}$, it is clear that both the assertions have no statistical significance behind their descriptive power. Moreover, it follows from (1) that ω_i lies in the interval $[0, 1/\rho(G)]$, $\rho(G)$ being the edge density of G and such an interval might be really wide for sparse graphs. To overcome this limitation[14], they developed a computational model (feasible only for the case of two colors) aimed at evaluating the likelihood of an observed instance $(\omega_0, \eta_{0,1})$ in the form of a phase diagram in the $m_{0,0}m_{0,1}$-plane. Each point of such a diagram is the frequency of all partitions of the node set G into two parts C_0 and C_1 with c_0 and c_1 nodes, respectively, such that the subgraph of G induced by C_0 has $m_{0,0}$ edges, while the subgraph induced by C_1 has $m - (m_{0,0} + m_{0,1})$ edges, m being the size of G. The diagram is computed by exhaustive enumeration for small graphs, while for large graphs only the boundary of the diagram is heuristically computed. In either cases, the likelihood of the observed pair is determined by its position and its darkness (in a grayscale) in the phase diagram. Although this approach has been proven successfully for a wide range of real networks (with only two functional classes), including certain PPI networks[14], it still suffers of the following limitations:

(a) it is computationally expensive. In fact, an exact evaluation of the phase diagram requires time exponential in the number of nodes in the network, and can be applied to large instances only by exploiting heuristic algorithms on a sample. Also, after sampling a subgraph with $\tilde{m}$ nodes, the complexity is $O(n^2\tilde{m})$;
(b) it can be applied only to two functional classes;
(c) it is rather qualitative.

To overcome these limitations, we propose to compute the **z**-score of ω_i and $\eta_{i,j}$ under the null model described in the next section. Since, as we show, this can be done for any number s of colors in $O(s(n+m))$ time, and s is usually a small constant, we have that our algorithm is time optimal, hence (a) and (b) are settled. As for (c), if $Z(\omega_i)$, say, is the **z**-score of ω_i, then by Čebyšëv inequality the probability of the event $(Z(\omega_i) > \lambda)$ is at most λ^{-2} under the null model. Hence $Z(\omega_i)^{-2}$ directly measures the statistical significance of ω_i, at the same time making the method completely quantitative. Moreover, we propose to evaluate the **z**-score of the number of isolated nodes in the subgraph induced by each color, that are expected to be negative values in the case of homophilic network. This computation is computationally harder, requiring $O(nm)$ time, but experimental results show to be quite fast on networks with order of 10^5 edges, and is still applicable to sparse networks with about 10^6 nodes.

Design of the new model.

Throughout the rest of the paper, we think of a network as an undirected graph G with node-set $V(G)$ and edge-set $E(G)$. An *s-coloring* of G is a surjective map $g : V(G) \rightarrow [s]$, where $[s] := \{1, \ldots, s\}$ is the set of colors. As previously stipulated, we think of g as the functional description of the network, and of the set $g^{-1}(i)$, consisting of the nodes of G having color i, as the functional classes of the description. These classes are our (potential) communities. Hence, in the pair (G, g), G encodes the geometrical description of the network and g encodes its functional description. For instance, Protein–Protein Interaction networks (PPI for shortness) are graphs whose nodes are proteins and whose edges model functional interactions between proteins. Since proteins are classified by the biological function they are responsible for, each protein is uniquely associated with one of the 19 functional classes listed in Table 2 and which we identify by their labels. Therefore, given a PPI network G, the correspondence protein$\mapsto$function defines a surjective map g from the set of nodes of G into a set of 19 labels and, after thinking of the labels as colors, such a correspondence will be our 19-coloring g. For the Pokec social network graph, we partitioned the node set into five age classes. Therefore a correspondence user$\mapsto$age defines a 5-coloring. Notice that the classification of ages is not frequency based, so that node classes differ substantially in size.

Let c_i, $i \in [s]$, be the number of nodes of G of color i under g and call the integer vector $\mathbf{c} = (c_1, \cdots, c_s)$ the *profile* of g. Any other coloring $f : V(G) \rightarrow [s]$ with the same profile as g will be referred to as a **c**-*coloring of* $V(G)$ (or simply **c**-coloring when $V(G)$ is understood). Our next step is to introduce a probability space that allows us to formulate null hypotheses to test against alternative hypotheses about (G, g). To this end, let $\Phi(\mathbf{c})$ be the set of all **c**-colorings of $V(G)$. Since the *multinomial coefficient* with *parts* $c_1, c_2 \cdots c_s$, denoted by one of the two symbols below

$$\binom{n}{\mathbf{c}}, \quad \binom{n}{c_1 c_2 \cdots c_s},$$

counts the $\mathbf{c}$-colorings of $V(G)$ (see the Appendix for a definition of multinomial coefficient), it follows that $|\Phi(\mathbf{c})| = \binom{n}{\mathbf{c}}$. A *random* $\mathbf{c}$*-coloring* is the random variable F with values in $\Phi(\mathbf{c})$ and with probability mass function given by

$$\mathbb{P}_{n,\mathbf{c}}(F) = \Pr\{F = f\} = \binom{n}{\mathbf{c}}^{-1},$$

namely, all $\mathbf{c}$-colorings are equally likely (see the Appendix for a more formal definition not needed here). Having the probability space $(\Phi(\mathbf{c}), \mathbb{P}_{n,\mathbf{c}})$ we test functions of (G, g) versus the same functions under the null hypothesis (G, F), where F is a random $\mathbf{c}$-coloring of $V(G)$. We therefore define several random variables as functions of the random variable F, and such variables enable us to give first and second order moments of those statistics crucial for our purposes. We close this section by describing the former ones, deferring the description of the latter ones to the next section.

For a node $v \in V(G)$ and a color $i \in [s]$, let X_v^i be the Bernoulli random variable that equals to 1 if and only if node v has color i under the random $\mathbf{c}$-coloring F, i.e. X_v^i is the indicator of the event $F(v) = i$. Since X_v^i is a Bernoulli random variable, by (5) in the Appendix, one has

$$\mathbb{E}\left(X_v^i\right) = \Pr\left\{X_v^i = 1\right\} = \frac{c_i}{n}.$$

Analogously, for the product of two such variables for $u,\ v \in V$, $u \neq v$, and $i,\ j \in [s]$, after resorting to (5) and (6) in the Appendix, one has

$$\mathbb{E}\left(X_u^i X_v^j\right) = \Pr\left\{X_u^i X_v^j = 1\right\} = \Pr\left\{X_u^i = 1, X_v^j = 1\right\} = \begin{cases} \frac{c_i^{\underline{2}}}{n^{\underline{2}}} & \text{if } i = j \\ \frac{c_i c_j}{n^{\underline{2}}} & \text{if } i \neq j \end{cases} \tag{2}$$

where, after adhering to the notation in[16], for a positive integer a and a nonnegative integer r, we have denoted by the symbol $a^{\underline{r}}$ the *falling r-th power* of a (see also the Appendix for more details), namely $a^{\underline{r}} = a(a-1)\cdots(a-r+1)$, with $a^{\underline{0}} = 1$. Thus, the 2-nd falling power $a^{\underline{2}}$ of a equals $a(a-1)$. The above formula immediately shows that the random variables X_v^i as v runs in $V(G)$ and i runs in $[s]$ are not independent (neither are X_u^i and X_v^j). Without pretending to be rigorous, this is only due to the fact that a random $\mathbf{c}$-coloring can be thought of as the outcome of experiments where one draws from a bin "without replacement". However, variables in $\{X_v^j \mid v \in V, j \in [s]\}$ are *exchangeable*, in the sense that the joint distribution of any subset of them does not depend on the order of drawing (the distribution is symmetric with respect to permuting indices). Hence, as long as we consider statistics based only on linear combinations of X_v^i, there is no other dependency other than the one inherited by the sampling procedure. To let the graph come into the structure of the dependency among variables, we have to consider second order statistics.

Let us come to edges now and, for an edge $uv \in E(G)$ and colors $i,\ j \in [s]$, let $Y_{uv}^{i,j}$ be the Bernoulli random variable which is equal to 1 if and only if one of the endpoints of uv has color i and the other one has color j. Hence , if $i = j$, then $Y_{uv}^{i,i} = X_u^i X_v^i$ while if $i \neq j$, then $Y_{uv}^{i,j} = X_u^i X_v^j + X_u^j X_v^i$. Therefore by (2)

$$\mathbb{E}\left(Y_{uv}^{i,j}\right) = \Pr\left\{Y_{uv}^{i,j} = 1\right\} = \begin{cases} \frac{c_i^{\underline{2}}}{n^{\underline{2}}} & \text{if } i = j \\ 2\frac{c_i c_j}{n^{\underline{2}}} & \text{if } i \neq j \end{cases}. \tag{3}$$

One more random variable is needed to compute the first two moments of the statistics we are interested in. Let T be a nonempty subset of $V(G)$ and let $i \in [s]$ be a color; define D_T^i as the number of elements of T having color i; by definition, D_T^i has the following expression:

$$D_T^i = \sum_{v \in T} X_v^i$$

Let A and B be disjoint subsets of $V(G)$. To determine the distribution of D_T^i we are interested in the probability of the event that all the elements of A have color i while all those of B have not. Let $\Omega_i(A, B)$ denote this event (for more on events of this type refer to the Appendix). Thus

$$\Omega_i(A, B) = (F(a) = i,\ \forall a \in A) \wedge (F(b) \neq i,\ \forall b \in B).$$

Hence

$$\left(D_T^i = h\right) = \bigvee_{\substack{R \subseteq T \\ |R| = h}} \Omega_i(R, T \setminus R)$$

and since the events on the right hand side of the identity above are mutually incompatible, after equation (3) in the Appendix and after setting $t = |T|$, one has

$$\Pr\left\{D_T^i = h\right\} = \binom{t}{h} \frac{c_i^{\underline{h}}(n-c_i)^{\underline{t-h}}}{n^{\underline{t}}} \tag{4}$$

and the close resemblance with the binomial distribution with parameters t and $\frac{c_i}{n}$ is clear: powers are replaced by falling powers. This is not an accident: D_T^i follows a hypergeometric distribution $\mathrm{Hyp}(n, c_i, t)$ giving the probability of success by drawing without replacement t balls from an urn containing n balls, c_i of which are successfull. By choosing T equal to the neighborhood of a node $v \in V(G)$, one immediately gets the distribution of the random number of neighbors of node v with color i, i.e. $D^i_{N_G(v)} \sim \mathrm{Hyp}(n, c_i, \deg_G(v))$.

Homophily, heterophily and isolated nodes: first and second order moments. We are now in position to describe statistics capable of assessing whether PPI networks are homophilic. Let (G, g) be a pair consisting of a PPI network G with n nodes and m edges and a $\mathbf{c}$-coloring g. We classify the m edges of G according to the colors of their endpoints. Consequently, we say that edge $uv \in E(G)$ is a (i, j)-edge of (G, g) if $\{g(u), g(v)\} = \{i, j\}$, $i, j \in [s]$—with a little abuse of notation we also admit $i = j$. Notice that (i, i)-edges, the *intra-community* edges, are the edges of G induced by the nodes in color class i (those responsible for the homophily of (G, g)) and, for $i \neq j$, (i, j)-edges, the *inter-community* edges, are the edges with one endpoint in color class i and the other one in color class j (those responsible for the heterophily of (G, g)). Let $m_{i,i}$ and $m_{i,j}$ be the number of (i, i)-edges and (i, j)-edges of (G, g), respectively. Therefore, for any two (possibly equal) colors $i, j \in [s]$, the random variable

$$M^{i,j} = \sum_{uv \in E(G)} Y_{uv}^{i,j}$$

counts the number of (i, j)-edges of (G, F) where F is a random $\mathbf{c}$-coloring. Let $\overline{m}_{i,j}$ be the expected value of $M^{i,j}$: by (3) and the linearity of expectation it follows straightforwardly that

$$\overline{m}_{i,j} = \begin{cases} m\frac{c_i^{\underline{2}}}{n^{\underline{2}}} & \text{if } i = j \\ 2m\frac{c_i c_j}{n^{\underline{2}}} & \text{if } i \neq j \end{cases},$$

which generalizes to an arbitrary number of colors the corresponding expressions given above for two colors. Analogously, we define the i-homophily of (G, g) and (i, j)-heterophily of (G, g), $i \neq j$, as the ratios

$$\omega_i = \frac{m_{i,i}}{\overline{m}_{i,i}}, \quad \eta_{i,j} = \frac{m_{i,j}}{\overline{m}_{i,j}},$$

namely, the relative intra- and inter-community density, respectively (recall the identities in (1)). If for all $i, j \in [s]$ (possibly $i = j$) we knew the variance $\sigma_{i,j}^2$ of $M^{i,j}$, then we could compute the $\mathbf{z}$-score of the observed ω_i e $\eta_{i,j}$ as the ratios

$$Z(\omega_i) = \frac{m_{i,i} - \overline{m}_{i,i}}{\sigma_{i,i}} = Z(m_{i,i}), \quad Z(\eta_{i,j}) = \frac{m_{i,j} - \overline{m}_{i,j}}{\sigma_{i,j}} = Z(m_{i,j}) \ . \tag{5}$$

By Čebyšëv inequality, if we assume, for instance, the null hypothesis that the observed value ω_i is a value assumed by the random variable $\frac{M^{i,i}}{\overline{m}_{i,i}}$ in the probability space $(\Phi(\mathbf{c}), \mathbb{P}_{\mathbf{c},n})$—which is tantamount to assume that (G, g) does not display i-homophily—then the confidence level for accepting the null hypothesis would be at most $Z^{-2}(\omega_i)$. Deferring for a while the computation of $\sigma_{i,j}^2$, let us examine another useful statistic for (G, g): the number l_i of isolated nodes in the subgraph induced by color i, i.e. the number of nodes in color class i having no neighbors in color class i. Call any such node *i-isolated* and observe that by definition the number of i-isolated nodes is

$$l_i = \left|\left\{v \in V(G) \mid g(v) = 1 \wedge g(w) \neq i, \ \forall w \in N_G(v)\right\}\right| \ .$$

Let L^i be the random variable defined as the number of i-isolated nodes of (G, F), where F is a random $\mathbf{c}$-coloring of $V(G)$. Although the random variables L^i's and $M^{i,i}$'s are clearly dependent (as confirmed by results plotted in Fig. 7) in the next section—at the extreme cases, for instance, $\Pr\left\{M^{i,i} = 0 \mid L^i \geq c_i - 1\right\} = 1$ and $\Pr\left\{M^{i,i} \geq \frac{c_i}{2} \mid L^i = 0\right\} = 1$—the joint knowledge of corresponding statistics l_i and ω_i is still quite informative. Indeed, consider two graphs G and $\tilde{G}$ on the same node set and let g be a $\mathbf{c}$-coloring of $V(G)$. The i-homophily of (G, g) and $(\tilde{G}, g)$ could be well the same, but the number of i-isolated nodes can be significantly different as in the following example.

Example 1 For a positive integer t denote by K_t the complete graph on t nodes and by $\overline{K}_t$ its complement, namely the graph with t nodes and no edges. Also denote by $K_{1,t}$ the complete bipartite graph with one node in a color class and t nodes in the other class. Finally, for graphs G and H denote by $G + H$ their disjoint union, namely the graph obtained by picking a copy of G a copy of H disjoint from G, and then forming the union of the two copies. Consider the subgraphs G_i and $\tilde{G}_i$ induced by color i in G and $\tilde{G}$, respectively. If, for some positive integer p, one has $G_i \cong K_p + \overline{K}_{2p}$ and $\tilde{G}_i \cong K_{p-1} + K_{1,p-1} + \overline{K}_p$, then G_i and $\tilde{G}_i$ have the same i-homophily but the number of i-isolated nodes in G_i is twice the number of i-isolated nodes in $\tilde{G}_i$.

Therefore, if we knew that $\omega^i \leq \tilde{\omega}^i$ and $l^i \geq \tilde{l}^i$, then this fact would support the claim that $(\tilde{G}, g)$ is more i-homophilic than (G, g) because the relative density of property i is less concentrated in $(\tilde{G}, g)$ than in (G, g). In conclusion, to assess i-homophily of (G, g) the use of the statistics $(Z(\omega^i), Z(l^i))$, where $Z(\omega^i)$ and $Z(l^i)$ are the **z**-scores of ω^i and l^i, respectively, could be useful. The next theorem, besides summarizing what we have said about the first order moments of the statistics considered so far, also gives the announced expression for $\sigma^2_{i,j}$ and the expression for the variance of L^i. We then exploit these results to compute **z**-scores as a tool for analyzing networks in the next section.

Theorem 1 *Let G be a graph with n nodes and m edges and let $(\Phi(\mathbf{c}), \mathbb{P}_{n,\mathbf{c}})$ be the probability space of the random $\mathbf{c}$-colorings, where $\mathbf{c} = (c_1, \ldots, c_s)$. Assume $c_i > 0$, $\forall i \in [s]$. Moreover, let $\pi_3(G)$ denote the number of (not necessarily induced) copies of P_3 in G. For $i, j \in [s]$, consider the random variables $M^{i,j}$ and L^i defined on $(\Phi(\mathbf{c}), \mathbb{P}_{n,\mathbf{c}})$. Then*

(1) *for $i \in [s]$ the expected value and the variance of random variable $M^{i,i} = \sum_{uv \in E(G)} Y^{i,i}_{uv}$ where $Y^{i,i}_{uv} = X^i_u X^i_v$ for all $uv \in E(G)$, namely the random number of (i, i)-edges of (G, F) under a random coloring F, are respectively given by*

$$\overline{m}_{i,i} = m\frac{c_i^{\underline{2}}}{n^{\underline{2}}},$$

$$\sigma^2_{i,i} = m\frac{c_i^{\underline{2}}}{n^{\underline{2}}}\left(1 - m\frac{c_i^{\underline{2}}}{n^{\underline{2}}}\right) + 2\left\{\left(\frac{c_i^{\underline{3}}}{n^{\underline{3}}} - \frac{c_i^{\underline{4}}}{n^{\underline{4}}}\right)\pi_3(G) + \frac{c_i^{\underline{4}}}{n^{\underline{4}}}\binom{m}{2}\right\};$$

(2) *for $i, j \in [s]$, $i \neq j$, the expected value and the variance of random variable $M^{i,j} = \sum_{uv \in E(G)} Y^{i,j}_{uv}$ where $Y^{i,j}_{uv} = (X^i_u X^j_v + X^j_u X^i_v)$ for all $uv \in E(G)$, namely the random number of (i, j)-edges of (G, F) under a random coloring F, are respectively given by*

$$\overline{m}_{i,j} = 2m\frac{c_i c_j}{n^{\underline{2}}},$$

$$\sigma^2_{i,j} = 2m\frac{c_i c_j}{n^{\underline{2}}}\left(1 - 2m\frac{c_i c_j}{n^{\underline{2}}}\right) + 2\left\{\left(\frac{c_i c_j^{\underline{2}} + c_i^{\underline{2}} c_j}{n^{\underline{3}}} - 4\frac{c_i^{\underline{2}} c_j^{\underline{2}}}{n^{\underline{4}}}\right)\pi_3(G) + 4\frac{c_i^{\underline{2}} c_j^{\underline{2}}}{n^{\underline{4}}}\binom{m}{2}\right\};$$

(3) *for $i \in [s]$* let *L^i be the random number of i-isolated nodes of (G, F) under a random coloring F, namely the random variable $L^i = \sum_{v \in E(G)} W^i_v$, where W^i_v is the Bernoulli variable defined as the indicator of the event $(F(v) = i) \wedge (F(w) \neq i, \forall w \in N_G(v))$; then the expected value and the variance of L^i are respectively given by*

$$\mathbb{E}\left(L^i\right) = \frac{c_i}{n}\sum_{v \in V(G)} \frac{(n - c_i)^{\underline{\deg_G(v)}}}{(n - 1)^{\underline{\deg_G(v)}}},$$

$$\operatorname{var}(L^i) = \mathbb{E}\left(L^i\right)\left(1 - \mathbb{E}\left(L^i\right)\right) + \frac{c_i^{\underline{2}}}{n^{\underline{2}}} \sum_{\substack{(u, v) \in V(G) \\ u \neq v,\, uv \notin E(G)}} \frac{(n - c_i)^{\underline{b(u,v)}}}{(n - 2)^{\underline{b(u,v)}}},$$

where we have set $b(u, v) = |N_G(u) \cup N_G(v)| = \deg_G(u) + \deg_G(v) - |N_G(u) \cap N_G(v)|$. Clearly, $c_i - L^i$ is the random number of nodes of color i spanned by the (i, i)-edges.

A formal proof of Theorem 1 is given in the Appendix.

A couple of facts are notable before closing the section.

Statistics presented in points (1) and (2) in Theorem 1 can be easily computed in $O(n + m)$ time, where n is the number of nodes and m is the number of edges in the input graph, assuming we have a constant number of colors. Hence, computing the s^2 **z**-scores for the number of edges $M^{i,i}$ and $M^{i,j}$ is computationally efficient for any input instance. We observe that the method in[14] requires exponential time for an exact evaluation, or $O(s^2 n^3)$ time, where s is the number of functional classes, if optimisation heuristics are exploited. Computing statistics for the number of isolated nodes L^i presented in point 3) in Theorem 1 is more time consuming. As shown in the Appendix, it requires $O(smn)$ time, that can be improved to $O\left(s \cdot \sum_{v \in V} \deg(v)^2\right)$. This is still efficient for sparse large graphs, with up to millions of nodes and edges.

All of the second order statistics presented in the theorem have an expression that encodes part of the structure of the input graphs, e.g. its number of P_3's, $2K_2$'s as well as the cardinalities of the set of common neighbors of nonadjacent pair of nodes. This means that the *coefficient of variation* of ω_i, defined as $\sigma_{i,i}/\overline{m}_{i,i}$ is completely determined by G and c_i and that different **c**-colorings (inducing different functional description) have the same scale. In this respect the homophily of the pair (G, g) is an intrinsic measure of the same pair and the coefficient of variation of ω_i is an invariant of the pair $(G, \mathbf{c})$. We can thus answer the question "how homophilic the network is?" without resorting to comparisons with other networks.

Assessing and measuring homophily

In this section we reap the crops of the last theorem by devising a methodological recipe to assess and measure homophily in networks. The main tools in this respect are the **z**-scores computed in the previous section. Given a pair (G, g) consisting of a network and one of its functional description g—a partition of the node-set of the network into s classes of nodes having the same characteristic, e.g. age, marital status, biological function, kind of phone subscription, geographical localization etc.—we can define the $s \times s$ random matrix $\mathbf{D}$ whose i, j-th entry is the standardized random variable $(M^{i,j} - \overline{m}_{i,j})/\sigma_{i,j}$ and, analogously, the s-dimensional random vector $\mathbf{d}_0$ whose i-th entry is the random variable $(L^i - \mathbb{E}(L^i))/\sqrt{\mathrm{var}(L^i)}$—notice that $\mathbf{D}$ is symmetric because $M^{i,j}$ and $M^{j,i}$ are the same variable. From (G, g) we can compute the arrays $\mathbf{Z}$ and $\mathbf{z}_0$ consisting, respectively, of the **z**-scores of intra- and inter-community edges (with the former displayed on the main diagonal of the $s \times s$ matrix $\mathbf{Z}$) and of the s **z**-scores of the i-isolated nodes (nodes of color i none of whose neighbors has color i), for $i = 1, \ldots, s$. We refer to $\mathbf{Z}$ and $\mathbf{z}_0$ as the **z**-score *arrays of* (G, g). Hence we may think of $\mathbf{Z}$ and $\mathbf{z}_0$ as the observed values of $\mathbf{D}$ and $\mathbf{d}_0$, respectively—notice that $\mathbf{Z}$ is symmetric as well. For an array $\mathbf{A}$ (matrix or a vector) denote by $1/\mathbf{A}^2$ the array of the same dimensions as $\mathbf{A}$ whose generic entry b is a^{-2}, a being the corresponding entry of $\mathbf{A}$. Call the arrays $1/\mathbf{Z}^2$ and $1/\mathbf{z}_0^2$ *U-values arrays*. By Čebyšëv inequality, the U-values arrays give (entry-wise) an upper bound of the probability of observing a value at least as extreme as the one observed for the corresponding random variable. Hence U-values are upper bounds of the corresponding p-values—so called in the Theory of statistical hypotheses. Although U-values arrays:

- do not capture the statistical dependency structure of the corresponding random arrays—this subject deserves further research;
- do not ensure a tight approximation of the corresponding p-values: though using only second order moments Čebyšëv bounds are undoubtedly the best possible bounds, such bounds can be actually rather loose yielding (possibly) too conservative methods (especially in conjunction with the pervious point),

U-values arrays certainly exhibit the following merits:

- robustness: U-values do not require distributional assumptions and therefore have an endogenous nature;
- complexity: U-values can be efficiently computed (see "Implementation details" section);
- rigour: U-values are computed exactly and do not require sampling or estimates and have precise quantitative meaning for homophily.

Notice that the U-values arrays $(1/\mathbf{Z}^2, 1/\mathbf{z}_0^2)$ and the **z**-scores arrays $(\mathbf{Z}, \mathbf{z}_0)$ convey the same statistical information. Hence $(\mathbf{Z}, \mathbf{z}_0)$ is already a direct measure of the homophily of G with respect to g. We spend the remainder of the section to substantiate this claim.

Descriptive power of z-score arrays and comparisons of networks The generic entry of $\mathbf{Z} = \{z_{i,j}\}$ measures the distance from the expected value of the corresponding random variables on a scale whose unit is the mean square error. At the same time, such an entry bounds from above the likelihood of this distance through the U-values, namely, the map $z_{i,j} \mapsto z_{i,j}^{-2}$. Similar considerations hold for the array $\mathbf{z}_0$.

It follows that **z**-score arrays can be conveniently described as heat-maps that provide a visual representation of homophily. These kind of diagrams can be particularly useful when comparing different networks that use the same set of colors because all the arrays involved have the same dimensions and thus the corresponding heat-maps are comparable. This can be done for PPI networks, for instance, because they have the same functional description (see "Protein-protein interaction networks" and "Numerical results" sections). In this case one can also refine the analysis with the help of vector $\mathbf{z}_0$ to provide a measure of the concentration of homophily in each color class (however we did not pursue this idea numerically).

Multiple Testing The natural extension of Park and Barabasi's method[14] is the following procedure, which we present first in a scalar form to clarify the need for the Bonferroni correction and then in a more algebraic form to confirm the descriptive power of matrix $\mathbf{Z}$. Although in what follows, when dealing with hypothesis testing, it would be more appropriate to use one-sided Čebyšëv inequality (a.k.a. Cantelli's inequality)—this amounts to consider $(1 + z_{i,j}^2)^{-1}$ in place of $z_{i,j}^{-2}$—for simplicity we stick to the two-sided Čebyšëv inequality.

<u>Procedure</u>. Given the pair (G, g) fix a significance level α. Compute the **z**-scores arrays $(\mathbf{Z}, \mathbf{z}_0)$. If $z_{i,i} \geq \dfrac{1}{\sqrt{\alpha}}$, then declare G i-homophilic at level α (recall that $z_{i,i} = Z(\omega_i)$). Analogously, if $z_{i,j} \geq \dfrac{1}{\sqrt{\alpha}}$, $i \neq j$, then declare G (i, j)-heterophilic at level α (recall that $z_{i,j} = Z(\eta_{i,j})$). Array $\mathbf{z}_0$ can be dealt with in the same way and can be used to refine the analysis. (6)

While the procedure above correctly assesses homophily (heterophily) of the marginal entries of $\mathbf{D}$, it is not true that the same significance level is valid for the joint distribution of $\mathbf{D}$. For assessing joint homophily (heterophily) we have to look at Procedure (6) as a *multiple testing* procedure which therefore requires multiple testing corrections. One of such correction, the most conservative one, is Bonferroni's correction which, in its simplest

Organism			PPI network		
Species	**Kingdom**	**Phylum/class**	**Nodes**	**Edges**	**Density**
Brucella melitensis (**Bm**)	Bacteria	Alphaproteobacteria	2675	15,450	0.43%
Escherichia coli (**Ec**)	Bacteria	Gammaproteobacteria	4020	29,748	0.37%
Haemophilus influenzae (**Hi**)	Bacteria	Gammaproteobacteria	1609	9202	0.71%
Helicobacter pylory J99 (**Hp**)	Bacteria	Epsilonproteobacteria	1264	7678	0.96%
Mycobacterium tuberculosis H37Rv (**Mt**)	Bacteria	Actinobacteria	3779	24,889	0.35%
Streptococcus pneumoniae TIGR4 (**Sp**)	Bacteria	Firmicutesi/Bacilli	1811	8813	0.54%
Treponema pallidum (**Tp**)	Bacteria	Spirochaetes	894	8157	2.04%
Vibrio cholerae (**Vc**)	Bacteria	Gammaproteobacteria	3153	20,844	0.42%
Pyrococcus abyssi (**Pa**)	Euryarchaeota	Thermococci	1564	9090	0.74%
Saccharomyces cerevisiae (**Sc**)	Fungi	Ascomycota/Saccharomycetes	6157	119,051	0.63%

Table 1. Complete list of considered organisms, together with their network size (nodes and edges). Density is expressed as the ratio between the actual number of edges and the number of edges in the complete graph with the same number of nodes.

form, scales level α—the level below which the null hypothesis is rejected—by the reciprocal of the number h of testing performed. For instance, suppose we want to assess whether a pair (G, g) is jointly homophillic at level α. Then we need to simultaneously test the s diagonal elements of $\mathbf{D}$. In this case, Procedure (6) specializes by declaring that (G, g) is i-homophilic when $z_{i,i} > \frac{s}{\sqrt{\alpha}}$. Clearly, as the number of testing increases, the procedure becomes too conservative especially in conjunction with Čebyšëv bounds. This limitation is unavoidable without further information about the statistical dependence structure among the marginals of $\mathbf{D}$. Nonetheless, by using a slightly refined form of Bonferroni correction, we can still devise a method to measure homophily in a given network and to compare homophily between different networks that use the same set of colors. For $(i,j) \in [s] \times [s]$, with $i \neq j$, consider the alternative hypothesis $\mathcal{H}^1_{i,j} : D_{i,j} > 0$ versus the null hypothesis $\mathcal{H}^0_{i,j} : D_{i,j} \leq 0$ at the significance level $\alpha_{i,j}$. Pair (i, j) is said to *positive at the significance level* $\alpha_{i,j}$ whenever Procedure (6) accepts $\mathcal{H}^1_{i,j}$. More generally, for $Q \subseteq \{(i,j) \in [s] \times [s] \mid i \neq j\}$, the joint confidence level of the family of tests $\mathcal{H}^0_Q = \{\mathcal{H}^0_{i,j} \mid (i,j) \in S\}$—a.k.a *family-wise error rate of the family of tests* $\mathcal{H}^0_Q$—is $\alpha = \min\{1, \sum_Q \alpha_{i,j}\}$ and set S is called *positive at the joint significance level* α whenever $\mathcal{H}^1_{i,j}$ is accepted by Procedure (6) for all $(i,j) \in Q$. The main observation is as follows. If we prescribe the individual significance level $\alpha_{i,j} = z^{-2}_{i,j}$ for $(i,j) \in Q$, then S will be positive at the joint significance level $\min\{1, \sum_Q z^{-2}_{i,j}\}$. In particular, if $Q = \{(i,i) \mid i \in [s]\}$, then the set of diagonal positions of $\mathbf{Z}$, namely the positions of the $\mathbf{z}$-scores of the intra-community densities, is positive at joint confidence level given by the trace of the U-value array $1/\mathbf{Z}^2$. This observation suggests that we can relate the number of positive elements in a set Q at a significance level α with the sum of entries of $1/\mathbf{Z}^2$ indexed by Q. Indeed, let $Q\ (\alpha)$ be the largest subset of Q such that $\sum_{(i,j)\in Q(\alpha)} z^{-2}_{i,j} \leq \alpha$ and let $q(\alpha)$ be the cardinality of $Q(\alpha)$. Notice that $q(\alpha)$ can be 0. Hence Q contains exactly $q(\alpha)$ positive elements at the joint significance level α. Parameter, $q(\alpha)$ depends only on Q, $\mathbf{Z}$ and α and therefore can be used to compare different networks that use the same set of colors. On the other hand, by definition, $q(\alpha)$ is related to $\mathbf{Z}$ by the following fact: for a real number λ, let $J(\lambda) = \{(i,j) \in [s] \in [s] \mid i \leq j \wedge z_{i,j} > \lambda\}$. It is clear that for each α there exists a λ (not in general unique) such that $Q(\alpha) = J(\lambda) \cap Q$. Therefore, family $\{J(\lambda) \mid \lambda \in \mathbb{R}\}$ globally conveys the same information as family $\{q(\alpha) \mid \alpha \in [0, 1)\}$ and we can get rid of the significance level α when comparing networks that use the same set of colors. Notice however that $\{J(\lambda) \mid \lambda \in \mathbb{R}\}$ conveys globally the same information as the heat-map of the $\mathbf{z}$-score matrix $\mathbf{Z}$ with the temperature acting as an inverse transform of the significance level.

Synthetic measure via Multidimensional Čebyšëv-type inequalities Multidimensional Čebyšëv inequalities[17] provide a somewhat dual method to the multiple testing procedure above. Recall that if $\mathbf{X}$ is a d-dimensional real random vector whose marginals have zero mean and unitary variance, $\|\mathbf{X}\|$ is the Euclidean norm of $\mathbf{X}$, and t is a positive real number, then the following multidimensional Čebyšëv-type inequality holds

$$\Pr\{\|\mathbf{X}\| \geq t\} \leq \frac{d}{t^2}$$

by a straightforward application of Markov inequality to the random variable $\|\mathbf{X}\|^2$. The same inequality holds for matrices but replacing the Euclidean norm by the Frobenius norm and adjusting for dimensions. More generally, it holds by vectorializing any subset of entries of a given matrix (after adjusting for dimensions). For instance, direct application of inequality above yields:

$$\Pr\left\{\|\mathrm{diag}(\mathbf{D})\| \geq \|\mathrm{diag}(\mathbf{Z})\|\right\} \leq \frac{s}{\|\mathrm{diag}(\mathbf{Z})\|^2} \ ,$$

with $\mathrm{diag}(\mathbf{A})$ denoting the vector formed by the diagonal entries of the square matrix $\mathbf{A}$. Hence, the sum of the squares of the diagonal entries of $\mathbf{Z}$ gives a global synthetic measure of homophily: the higher such sum is the more globally homophillic the network is. Therefore,

Information storage and processing	
J	Translation, ribosomal structure and biogenesis
K	Transcription
L	Replication, recombination and repair
Cellular processes and signaling	
D	Cell cycle control, cell division, chromosome partitioning
V	Defense mechanisms
T	Signal transduction mechanisms
M	Cell wall/membrane/envelope biogenesis
N	Cell motility
U	Intracellular trafficking, secretion, and vesicular transport
O	Posttranslational modification, protein turnover, chaperones
Metabolism	
C	Energy production and conversion
G	Carbohydrate transport and metabolism
E	Amino acid transport and metabolism
F	Nucleotide transport and metabolism
H	Coenzyme transport and metabolism
I	Lipid transport and metabolism
P	Inorganic ion transport and metabolism
Q	Secondary metabolites biosynthesis, transport and catabolism
Poorly characterized	
X	Function unknown or general function prediction only

Table 2. Protein functional classes, partitioned into higher categories.

$$\max\left\{0, 1 - \frac{s}{\|\mathrm{diag}(\mathbf{Z})\|^2}\right\}$$

is a global index of homophily lying in [0, 1], like Newman's modularity index[15].

Numerical tests on real networks

We now probe our theoretical results on two different network classes: (i) Protein–Protein Interaction (PPI) networks, where nodes correspond to proteins, partitioned according to their functional role, and edges represent functional interactions between proteins (ii) on-line social networks, where nodes correspond to users, partitioned according to their age, and edges represent friendship between users. As shown in the previous section, the major character of our methodology is the **z**-score matrix **Z**. Let us discuss data and the running time of the method in some details before going to the numerical tests.

Protein–protein interaction networks.

We consider ten PPI networks retrieved from STRING database (https://string-db.org/)[18,19], setting a high confidence score cut-off (0.70). The selected networks, listed in Table 1, are mainly related to Bacteria (8 out of 10, belonging to diffent Phyla or classes), we also included in the study *Saccharomices cerevisiae* (Fungi - Ascomycota) and *Pyrococcus abyssi* (Euryarcheota - Thermococci) for comparison. The 8 bacterial organisms were chosen as representatives of Bacteria Kingdom, including different Phyla (Alpha, Gamma, Epsilon proteobacteria, Actinobacteria, Firmicutes/Bacilli, Spirochaetes). Organisms were also chosen on the basis of their network sizes (number of nodes and edges), in order to build an etherogeneous dataset. Species, Kingdom, Phylum/Class as well as number of nodes, number of edges, and density of the relative network are reported in Table 1 for each organism.

Functional classes of proteins of the considered ten organisms were obtained from NCBI database (ftp://ftp.ncbi.nih.gov/pub/COG/COG/). Proteins were partitioned into 25 different functional classes, but only 19 were taken into account in this work, since:

- 5 classes (A—RNA processing and modification, B—Chromatin structure and dynamics, Y—Nuclear structure, Z—Cytoskeleton, W—Extracellular structures) had no representatives (or only a few) for most of bacterial organisms;
- classes R—general function prediction, and S—Function unknown, were merged into the X class.

The 19 considered classes are reported in Table 2. The number of proteins for each functional class in each organism is reported in the Appendix.

Class	Age	Nodes	Edges
C	[12,18)	152,659	348,617
D	[18, 25)	332,826	2,038,089
E	[25, 40)	270,299	521,228
F	[40, 60)	46,295	23,156
X	Otherwise	410,270	949,026
Whole network		1,212,349	8,320,600

Table 3. Classes of Pokec social network. For each class the number of nodes is reported, with the number of edges joining nodes in the same class.

	Size			Computing time (s)	
Network	Nodes	Edges	Squared degrees sum	Edgez-score	Singleton z-score
Bm	2675	15,450	942,470	0.042	15.338
Ec	4020	29,748	1,947,532	0.077	63.174
Hi	1609	9202	607,128	0.023	10.477
Hp	1264	7678	535,246	0.020	9.973
Mt	3779	24,889	1,574,806	0.068	43.241
Sp	1811	8813	555,570	0.023	9.010
Tp	894	8157	818,544	0.021	14.284
Vc	3153	20,844	1,505,448	0.054	39.030
Pa	1564	9090	713,514	0.022	12.510
Sc	6157	119,051	30,075,870	0.257	1062.981
Pokec	1,212,349	8,320,600	752,382,968	24.270	24,086.467

Table 4. Computing times for edge **z**-scores and singleton **z**-scores, on organisms and Pokec networks. For each network we report the number of nodes, the number of edges and the sum of squared degrees. The complexity of singleton **z**-scores computation strongly depends on the sum of squared degrees.

Each organism's network is an undirected graph, in which each node represents a protein associated to a color denoting one of the functional classes listed in Table 2, and each edge represents the interaction between two proteins, weighted according to the likelihood of the given interaction. A PPI graph is thus represented by two text files, the first lists node labels and the associated colors, the second lists edges as pairs of nodes and the associated weight in range [0, 999]. Edges have been cut-off at a 700 minimum weight, usually considered as a high confidence threshold. Isolated nodes in the resulting graph have been deleted. Some networks present a very limited number of nodes (some units) labeled by similar values (e.g. `jhp0681_1` and `jhp0681_2` in the node file for organism *Helicobacter pylori*) representing different isoforms of the same protein, but these nodes were simply denoted by a unique label (e.g. `jhp0681`) in the edge listing file. We merged such nodes in a single node; in the few cases in which they were associated to different functional classes, we merged them associating the functional class X to that node.

Pokec social network. Pokec is the most popular Slovak on-line social network. Datasets, obtained during May 25–27 2012, are anonymized and contain relationships and user profile data of the whole network[20]. Friendships in the Pokec network are originally oriented. We decided to consider only symmetric pairs, so that we derived an undirected graph where nodes x, y are adjacent if and only if both x is a friend of y and y is a friend of x, so that it can be assessed that the two considered members had an actual interaction; also in this case, isolated nodes have been discarded. The network obtained contains more than one million nodes and 8 millions edges. Nodes are partitioned in classes according to the age declared by members, where about 34% of them either did not declare age, or declared a patently untrue value—in some cases even less than 10 or over 100. So, we decided to put into a "fake" age class denoted by X all members whose age is not a numeric value in [12, 60). The size of each subgraph induced by the 5 age classes, possibly containing isolated nodes, is shown in Table 3, together with the size of the entire network.

Implementation details. We developed a Python 3 prototype implementing our model, source code is available at http://www.statistica.uniroma1.it/users/pfrancio/homophily/.

Experiments have been performed on an Intel Core i5 PC with 4 cores, 2.3 GHz clock, 16 GB RAM, 256 KB L2 cache and 6 MB L3 cache, equipped with MAC OS 10.14.6. For the huge Pokec network, a 250 GB RAM machine running 18.04.5 LTS has been used.

Computing times, using a single core, are reported in Table 4, excluding time elapsed in file I/O. As it clearly appears from the table, the ratio between the number of edges in the graph and the time needed to compute edge **z**-scores is close to be constant (varying from 340k to 460k edges per second), confirming the asymptotic complexity $O(n+m)$—assuming the number of colors is constant.

An efficient computation of singleton **z**-scores requires some more care. Expression for var(L^i) in point 3) in Theorem 1 requires $O(n^3)$ time to be computed. Actually, it can be manipulated (details are discussed in the Appendix), so that the complexity of computing var(L^i) for each color i is lowered to $O(nm)$. More precisely, its complexity is strictly related to the number of pairs of nodes at distance 2, which in turns is bounded by π_3, i.e. the number of P_3's in the graph. It is immediate to see that

$$\pi_3 = \frac{1}{2}\sum_{v \in G}(\deg_G(v))^{\underline{2}} \leq \frac{1}{2}\sum_{v \in G}(\deg_G(v))^2$$

The sum of squared degrees for all experimented networks is reported in Table 4, where it is confirmed to be proportional to computing times for singleton **z**-scores (with a ratio varying from 28k to 61k P_3's per second).

Numerical results. In order to have a pictorial quantitative perception of homophily and heterophily in the considered networks, we present matrix **Z** of the **z**-scores of the intra- and inter-community edges (see "Assessing and measuring homophily" section) in the form of heat-maps. Color scale is logarithmic on **z**-scores, traslated in order to avoid negative values. Each entry of **Z** corresponds to a square in the diagram. Green squares corresponding to entry i, j represent positive **z**-scores, while pink squares represent negative **z**-scores. Results related to PPI networks are shown in Fig. 1. Homophily of PPI's with respect to their functional description is clearly readable from all the heat-maps by the green squares in all diagonals—showing the relative intra-community density—except for the poorly characterized X function class. A majority of off-diagonal **z**-scores are negative (more than 79.6%), while diagonal **z**-scores tend to show very high values. As a global result, neglecting all i, j pairs where either $i = \text{X}$ or $j = \text{X}$, we recap that:

- the average value of the diagonal entries **Z** is 36.26, with standard deviation 49,85, ranging from a − 0.3183 minimum to a 326.6 maximum;
- more than 91% of diagonal entries **Z** are greater than 5;
- the average value of off-diagonal entries **Z** is − 0.836, with standard deviation 3.707, ranging from a − 5.983 minimum to a 55.67 maximum;
- more than 65% of off-diagonal entries **Z** are less than − 1.

Concerning the off-diagonal entries of **Z** (namely, those corresponding to inter-community edges) it is worth noting that some classes show significant values, highlighting a unexpected heterophily although in most cases the associated classes belong to close functional classes such as class J, K and L, that can be grouped in the higher category *Information, storage and processing*.

In particular significant heterophilic **z**-scores are reported, in most of the organism networks, for classes J-L and class J-U representing *Translation, ribosomal structure and biogenesis* (class J), *Replication, recombination and repair* (class L) and *Intracellular trafficking, secretion, and vesicular transport* (class U). These heterophilic relationships can be considered reasonable from a biological point of view, since nodes associated to protein synthesis in the ribosome (class J) are related to nodes involved in DNA replication (class L) and also to intracellular transport (class U) according to the mechanics of protein biosynthesis (when DNA is transcribed, the resulting RNA copy is transported to the ribosome and after translation the protein can be transported away from the ribosome and onto the relevant part of the cell). These results provide consistency to our work as a real-world validation of our method.

To have a global and comparative glimpse of the whole scenario concerning PPI, we isolated the diagonal entries of **Z** and plotted them in Fig. 2 on a different scale.

A large majority of **z**-scores (diagonal) shows very high values corresponding to extremely significant deviation from expected ones. As expected, the exception regards last column related to X class (*Function unknown or General function prediction only*) showing **z**-score values typically negative including very small values (− 14 for *Saccharomyces cerevisiae*, − 7 for *Pyrococcus abyssi* and -6 for *Escherichia coli*) with only two organisms showing positive values (0.44 for *Mycobacterium tubercolosis* and 1.9 for *Vibrio cholerae*) (Fig. 2). This typical scenario is consistent with what we could expect from a biological point of view, since it is reasonable that proteins, envolved in a common task, could on average preferentially interact or be close to each other in the PPI. Proteins belonging to X class do not share a common task since in most of cases they are not associated to any given functional class, so it is reasonable that they are not likely to interact with each other. Some functional classes seem to show extremely high values, shared among almost all the organisms. It is evident for class J (*Translation, ribosomal structure and biogenesis*) showing the highest values, reaching huge **z**-scores (335 for *Escherichia coli*, 280 for *Mycobacterium tubercolosis*) always higher than 124. Also class N (Cell motility) shows extremely high **z**-score values reaching 229 for *Brucella mellitensis* and 206 for *Escherichia coli*, with the only exception of *Mycobacterium tubercolosis*—2.47—that is anyway more than two standard deviations greater than the expected one. Genes coding for proteins in bacteria are known to typically occur phisically close on chromosome, according to the operon paradigm, and it was shown, consistently with our findings (see[21]), that especially genes coding for proteins envolved in translation and cell motility task are very close to each other, favoring their syncronous transcription and the interaction of their protein products.

As for the Pokec social network, results are presented in a completely analogous manner: see Fig. 3 for the heat-maps, while in (4) we isolated the diagonal elements.

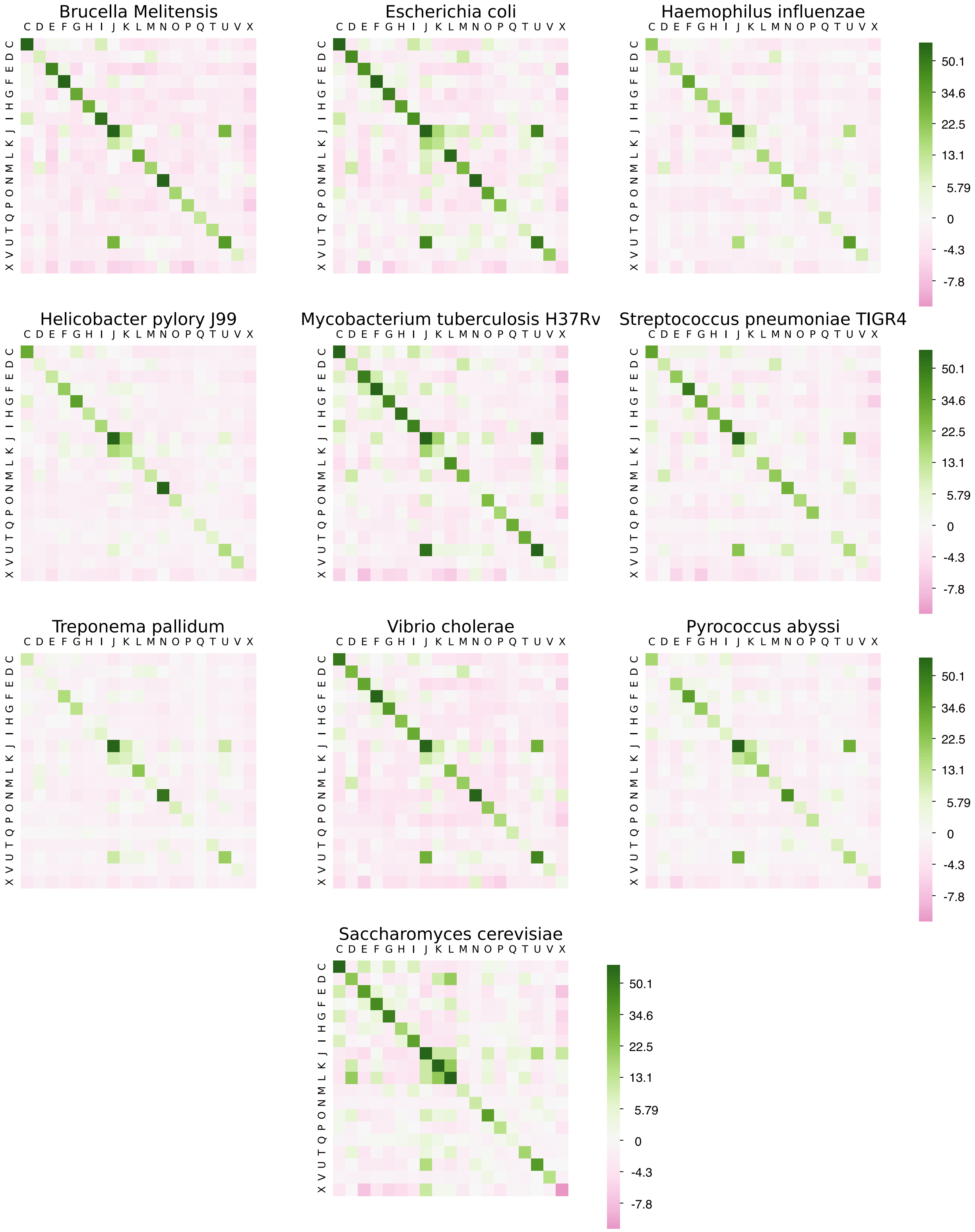

Figure 1. Heat-maps corresponding to **Z** matrices of the ten organism PPI networks. Diagonal entries correspond to intra-community edges **z**-scores, while off-diagonal entries correspond to inter-community edges **z**-scores. Values in the color scale have been cut to interval $[-10, 60]$.

As expected Pokec shows a significant homophilic beahavior with respect to the considered node attribute, age class, as reported in Table 3.

All diagonal **z**-scores, excepting class X (no age or non reliable value), reported in Figs. 3 and 4 show highly significant positive values, ranging from an astonishing value around 500 for class C ([12–18) years old) and

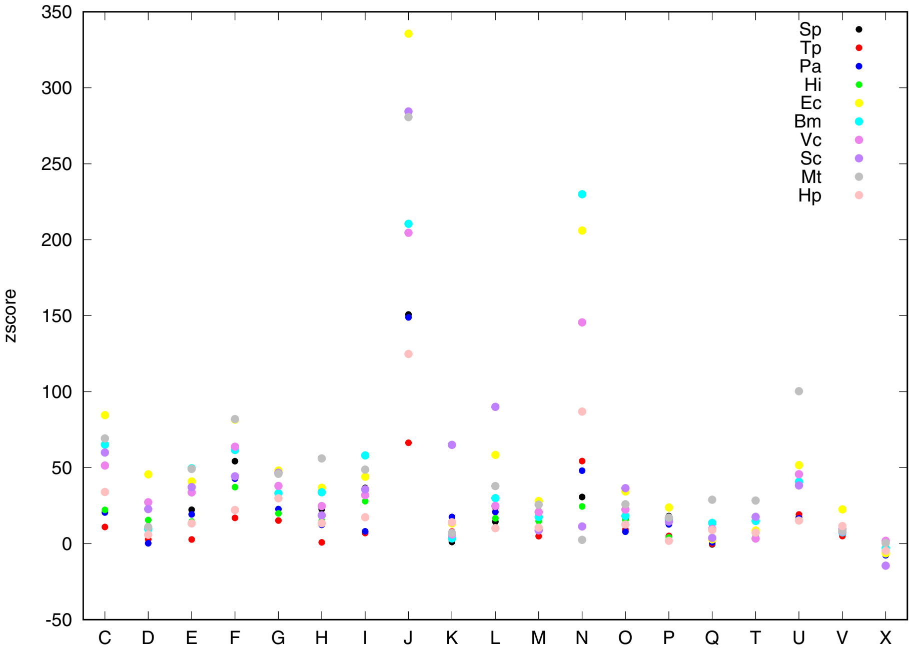

Figure 2. **z**-score intra-community density values (diagonal entries of **Z**) of each functional class (*x*-axis) are reported in different colors (each color representing a different organism as indicated in the top right legend of the plot).

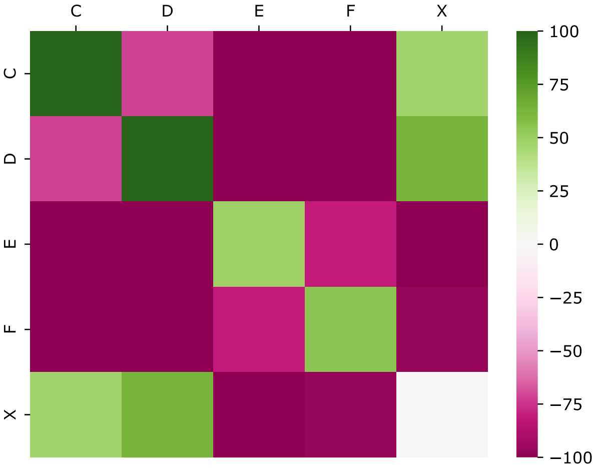

Figure 3. Heat-map corresponding to **Z** matrix of Pokec social network. Values in the color scale have been cut to interval [−100, 100].

around 200 for class D ([18–25) years old) till around 50 for classes E ([25–40) years old) and F ([40–60) years old). Diagonal **z**-score associated to class X is very close to 0, meaning that users that do not report their age (or report a non reliable age) do not interact with each other. They prefer to have relationships with other users reporting an age belonging to class C and D (showing positive values in the heat-map Fig. 3), while they do not interact with users belonging to class D and E. It can be hypothesized, if we trust in the homophilic nature of social network with respect to age, that most of those users (not reporting their age) have an age belonging to classes C and D.

To complement the analysis, we also computed vector $\mathbf{z}_0$. Recall that the *i*-th entry of such vector is the **z** -score of the number of isolated nodes in the subgraph induced by color *i* (functional class for the PPI and age class for Pokec). As explained in "Homophily, heterophily and isolated nodes: first and second order moments" section, although correlated with the intra-community densities (as confirmed for PPIs in Fig. 7: the higher the density, the lower the likelihood to find isolated nodes), the entries of $\mathbf{z}_0$ provides a measure of the concentration of the intra-community edges within color classes and, as expected, they are typically negative, consistently with what they represent. A negative entry means that subgraph induced by the corresponding functional classes for

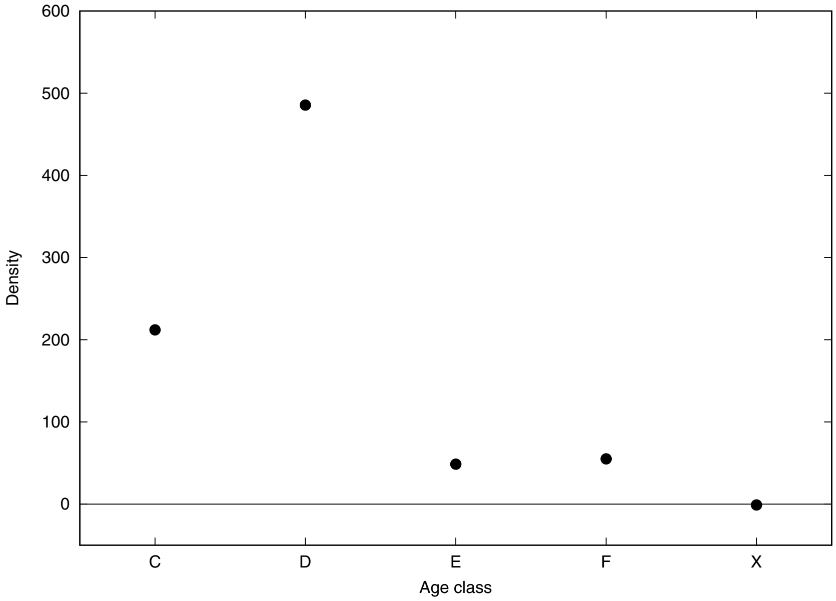

Figure 4. Diagonal **z**-score values related to age class.

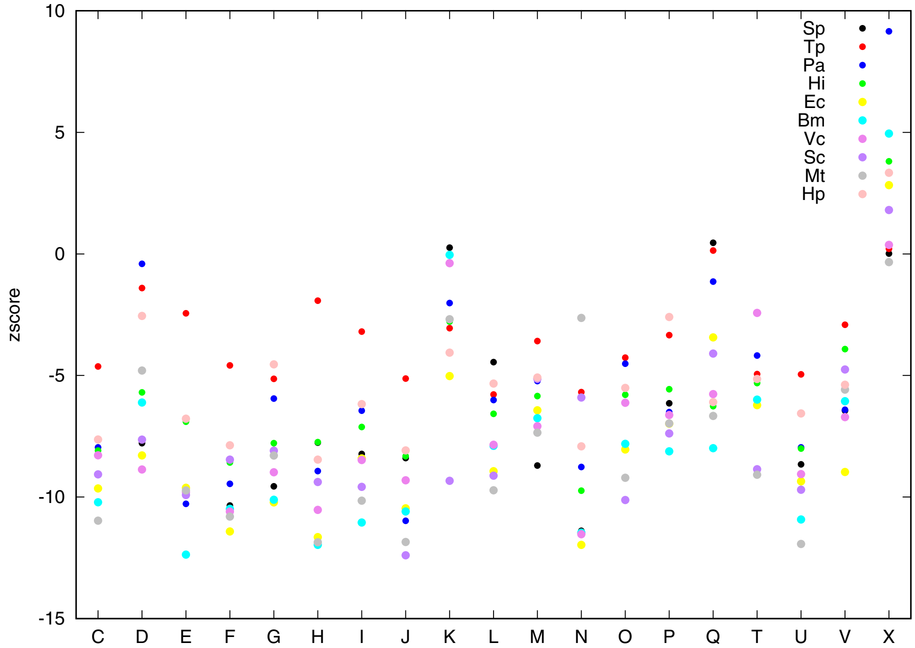

Figure 5. $\mathbf{z}_0$ values (y-axis) of each functional class (x-axis) are reported in different colors (each color representing a different organism as indicated in the top right legend of the plot).

the PPI and age class for Pokec contains less isolated nodes that expected. As can be observed in both Figs. 5 and 6, except for the X class which shows a **z**-score value close to zero for Pokec and few values close to zero **z** -scores associated to all other classes assume very low (negative) values (around 75% of values are smaller than −5, around −140, for class C and D and around −80, and −60, for class E and F in Pokec), that can be considered extremely significant from a statistical point of view.

Finally, as we said in "Homophily, heterophily and isolated nodes: first and second order moments" and "Assessing and measuring homophily" sections, the entries of the p-values arrays $1/\mathbf{Z}^2$ and $1/\mathbf{z}_0^2$ (obtained simply by squaring the reciprocal of the entries of the **z**-score arrays) can be rather loose estimates of the corresponding true quantiles. In this respect our method is rather conservative. Nonetheless, as shown in Fig. 8, a large majority of p-values entries are under the threshold of 0.05, which is usually considered as reliable (for individual testing) with the exceptions already discussed above.

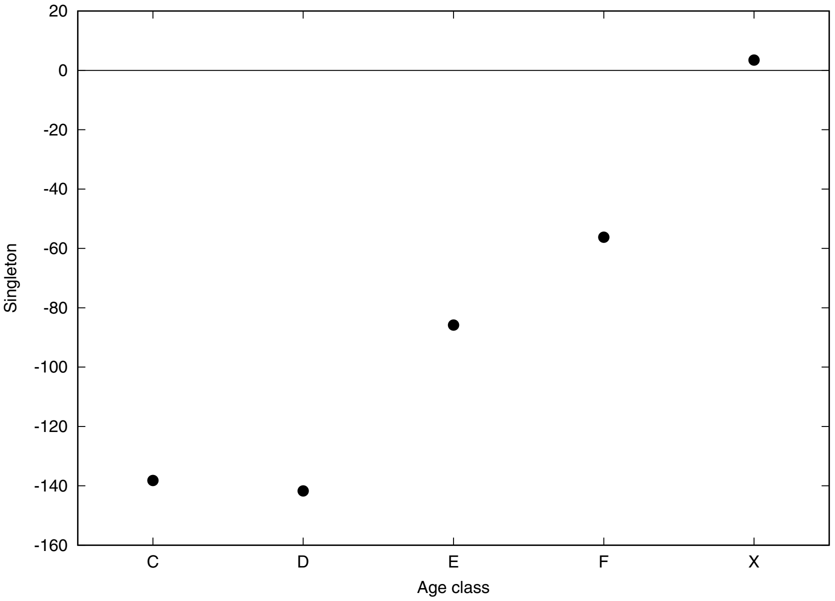

Figure 6. $\mathbf{z}_0$ values (y-axis) of each age class (x-axis) are reported.

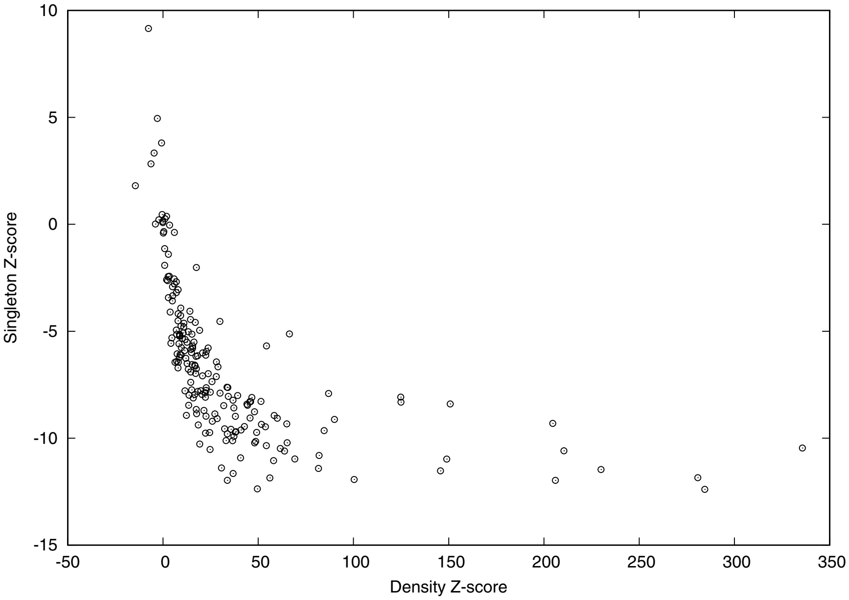

Figure 7. Correlation between diagonal **Z** values (x-axis) and $\mathbf{z}_0$ values (y-axis). Each point represents a given functional class of a given organism.

Conclusions and discussion

In this paper we presented a new approach to assess and measure homophily in networks. The model, described in "Assessing and measuring homophily" section, relies on computing

- the **z**-scores of $m^{i,j}$, the number of edges with one endpoint in functional class i and the other endpoint in functional class j (with possibly $i = j$),
- the **z**-scores of l^i, the number of nodes in functional class i with no neighbours in class i,

under the hypothesis that these numbers are samples from the corresponding random variables $M^{i,j}$ and L^i under the random coloring model $(\Phi(\mathbf{c}), \mathbb{P}_{n,\mathbf{c}})$ (the null model). These **z**-scores are either directly interpreted as a refined measure of network homophily (through heat-maps) or serve as the basis either for more synthetic measure via

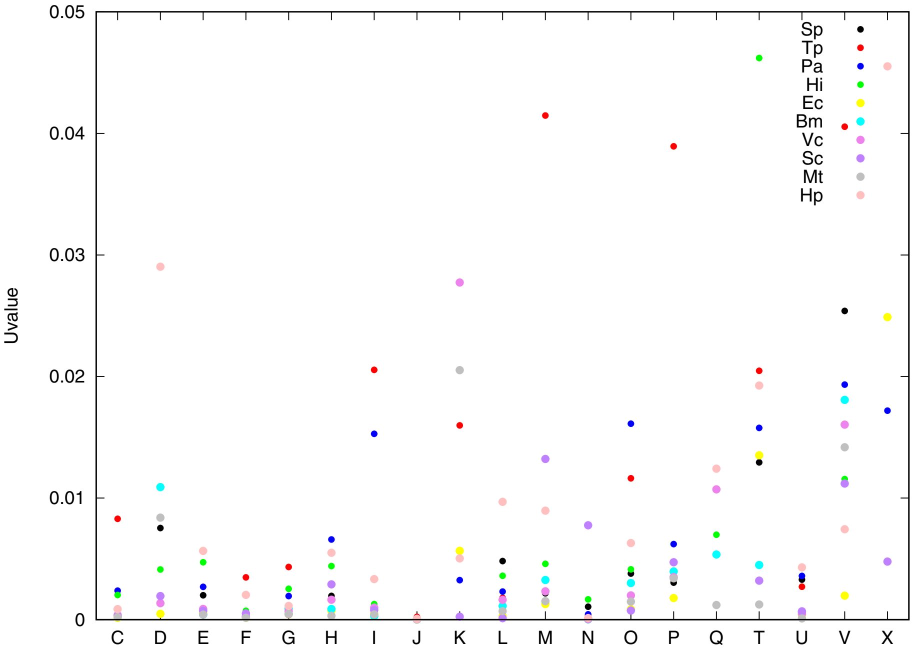

Figure 8. Diagonal entries of U-value arrays. The x-axis is labelled by functional classes and each color represents a different organism as indicated in the top right legend of the plot.

multiple testing or via the significance level of the Euclidean distance between the observed intra-community densities and the expected ones under the random coloring model. The idea of random coloring is implicit in[14] from which we also borrowed terminology. As a result, we extended their model to an arbitrary number of colors and made it computationally efficient and also quantitative (via the **z**-score). The method is clearly applicable to any kind of network and to any of its functional description. Different networks with the same functional description can also be compared directly. Moreover, we noticed that the coefficients of variations of the $M^{i,j}$'s and L^{i}'s are invariant for the pair (G, c), where G is the network and c is the profile of the functional description g of G.

Obtained results provide evidence of the strong homophilic nature of PPIs, in terms of protein function, and of Pokec social network, in terms of age classes, making our method reliable and affordable since homophilic nature of PPIs and social networks is something expected and known to some extent.

Network homophily is directly linked to network communities and to the paradigm of Guilt By Association (GAS)[22]. According to this paradigm, attribute of a given node can be inferred by analyzing the attributes of its neighbours[23,24]. In this view assessing and measuring network homophily can be extremely significant for the applicability of the GAS paradigm, allowing to classify nodes according to neighbor attributes. The analysis of **Z** matrix in Pokec network can provide an example of how GAS paradigm can be concretely applied. Users belonging to X class (age not reported or non reliable) are significantly close (according to the values of entries of **Z** matrix) to classes C and D, showing an heterophilic behavior while they are not close to users of classes D and E. This leads to hypothesize that users of class X could have, even if they did not report it, an age associated to class C or D. It is worth noting anyway that in some networks, in particular in PPIs, node attributes can be already classified through GAS paradigm, leading to a bias or to a tautological analysis, generating a circular argument.

Concerning PPI networks, comparison of **Z** matrices shows that the homophilic behavior is not linked to evident stronger similarity among close related species (also *Saccharomyces cerevisiae* and *Pyrococcus abyssi* show similar homophilic/heterophilic **z**-scores), so that homophilic behavior can be considered as an intrinsic characteristic of PPIs. Interestingly, some functional classes are more associated than expected showing an heterophilic behavior, especially classes J, K and L, that can be grouped in the higher category "Information, storage and processing". Another significant **z**-score highlights heterophily in most of organism networks with respect to classes J and U representing "Translation, ribosomal structure and biogenesis" (class J) and "Intracellular trafficking, secretion, and vesicular transport" (class U) respectively.

The model has been implemented in Python, and experimental results confirm that the computational complexity of the proposed model is optimal for edge density computation, requiring $O(n + m)$ time to compute the **Z** matrix. Computing the **z**-score of the number of i-isolated nodes is more time consuming, requiring $O(nm)$ time, but experiments show that it is still efficient in practice for sparse large networks.

In conclusion we are confident that this work can provide a significant contribution allowing to assess and measure, through a robust statistical method, homophily in networks.

Received: 11 August 2021; Accepted: 6 May 2022
Published online: 13 June 2022

References

1. McPherson, J. M., Smith-Lovin, L. & Cook, J. M. Birds of a feather: Homophily in social networks. *Ann. Rev. Sociol.* **27**, 415–444 (2001).
2. Aukett, R., Ritchie, J. & Mill, K. Gender differences in friendship patterns. *Sex Roles* **19**(1–2), 57–66 (1988).
3. Cheadle, J. E. & Schwadel, P. The friendship dynamics of religion, or the religious dynamics of friendship? A social network analysis of adolescents who attend small schools. *Soc. Sci. Res.* **41**, 1198–1212 (2012).
4. Karimi, F., Génois, M., Wagner, C., Singer, P. & Strohmaier, M. Homophily influences ranking of minorities in social networks. *Sci. Rep.* **8**(1), 1–12 (2018).
5. Kossinets, G. & Watts, D. J. Origins of homophily in an evolving social network. *Am. J. Sociol.* **115**, 405–450 (2009).
6. McPherson, J. M. & Smith-Lovin, L. Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *Am. Sociol. Rev.* **52**, 370–379 (1987).
7. Shrum, W., Cheek, N. H. Jr. & Hunter, S. M. Friendship in school: Gender and racial homophily. *Sociol. Educ.* **25**, 227–239 (1988).
8. Easley, D. & Kleinberg, J. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World* (Cambridge University Press, Cambridge, 2010).
9. Newman, M. E. J. Mixing patterns in networks. *Phys. Rev. E* **67**, 026126 (2003).
10. Kibae, K. & Altmann, J. Effect of homophily on network formation. *Commun. Nonlinear Sci. Numer. Simul.* **44**, 48249–4 (2017).
11. Lancichinetti, A., Kivelä, A., Saramäki, J. & Fortunato, S. Characterizing the community structure of complex networks. *PLoS ONE* **5**(8), e11976 (2010).
12. Gulbache, N. & Lehman, S. The art of community detection. *BioEssays* **30**, 934–938 (2008).
13. Yang, J. & Leskove, J. Defining and evaluating network communities based on ground-truth. *Knowl. Inf. Syst.* **42**, 181–213 (2015).
14. Park, J. & Barabasi, A. L. Distribution of node characteristics in complex networks. *Proc. Natl. Acad. Sci. USA* **104**(46), 17916–17920 (2007).
15. Newman, M. Modularity and community structure in networks. *Proc. Nat. Acad. Sci. USA* **103**(23), 8577–8582 (2006).
16. Knuth, D. *The Art of Computer Programming, Vol. 1: Fundamental Algorithms* 3rd edn. (Addison-Wesley, Reading, 1997).
17. Ferentinos, K. On Tchebycheff type inequalities. *Trabajos Estadıst. Investig. Oper.* **33**, 125–132 (1982).
18. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* **45**, D362-68 (2017).
19. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607-613 (2019).
20. Takac, L., Zabovsky, M. Data analysis in public social networks. In *International Scientific Conference & International Workshop Present Day Trends of Innovations* Lomza, Poland (2012).
21. Santoni, D. & Romano-Spica, V. Comparative genomic analysis by microbial COGs self-attraction rate. *J. Theor. Biol.* **258**, 513–520 (2009).
22. Oliver, S. Guilt-by-association goes global. *Nature* **403**, 601–603 (2000).
23. Deng, M., Zhang, K., Mehta, S., Chen, T. & Sun, F. Prediction of protein function using protein-protein interaction data. *J. Comput. Biol.* **10**(6), 947–960 (2003).
24. Piovesan, D., Giollo, M., Ferrari, C. & Tosa, S. C. E. Protein function prediction using guilty by association from interaction networks. *Amino Acids* **47**, 2583–2592 (2015).
25. Chowdhary, R., Zhang, J. & Liu, J. S. Bayesian inference of protein-protein interactions from biological literature. *Bioinformatics* **25**(12), 1536–1542 (2009).
26. Jansen, R. *et al.* A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**(5644), 449–453 (2003).
27. Jeong, H., Mason, S. P., Barabási, A. L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
28. Keshava Prasad, T. S. *et al.* Human protein reference database-2009 update. *Nucleic Acids Res.* **37**, D767–D772 (2009).
29. von Mering, C. *et al.* STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* **31**, 258–61 (2002).
30. Von Mering, C. *et al.* Comparative assessment of large-scale datasets of protein-protein interactions. *Nature* **417**, 399–403 (2002).

Acknowledgements

The authors are grateful to an anonymous referee for suggesting a deep revision of the article aimed at putting our methodology at the center, for suggesting one of the methods in the new "Assessing and measuring homophily" section, and also for inspiring the entire section. We thank both referees for their thorough reading and suggestions that substantially improved the presentation. Finally, we thank Margherita Notarantonio for her help in result visualization.

Author contributions

All the authors collaborated in devising the proposed method, in evaluating experimental results, and in producing the final version of this manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-12710-7.

Correspondence and requests for materials should be addressed to P.G.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.