Supporting Information

Peel et al. 10.1073/pnas.1713019115

SI Text

A. Directed Networks

We can easily extend the multiscale mixing measure r_α described in the main text to directed networks. The main change is to incorporate two sets of marginals a and b that describe the proportion of edges starting from and ending at each of the attribute types. Then the directed global assortativity of a network with respect to a particular categorical node attribute y_i is

$$r_{\text{global}} = \frac{\sum_g e_{gg} - \sum_g a_g b_g}{1 - \sum_g a_g b_g}, \tag{S1}$$

where a_g and b_g represent the total number of outgoing and incoming links of all nodes of type g,

$$a_g = \sum_h e_{gh}, \qquad b_h = \sum_g e_{gh}. \tag{S2}$$

Then we can update our definition of local assortativity accordingly,

$$r(\ell) = \frac{1}{Q_{\max}} \sum_g \left(e_{gg}(\ell) - a_g b_g\right). \tag{S3}$$

B. Scalar Attributes

For scalar attributes, we can simply calculate the Pearson's correlation across edges. Using x_i and x_j to indicate the scalar attribute value of the nodes in edge A_{ij}, then we can write the global assortativity as

$$r_{\text{global}} = \frac{\text{cov}(x_i, x_j)}{\sigma_i \sigma_j} \tag{S4}$$

$$= \frac{\sum_{ij} A_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i k_i (x_i - \bar{x})^2}, \tag{S5}$$

where $\bar{x} = 1/2m \sum_i k_i x_i$ is the mean value of x weighted by node degree k and σ_i is the SD of the attribute values. If we standardize the scalar values using the linear transformation $\tilde{x}_i = \frac{x_i - \bar{x}}{\sigma_i}$, then we can simplify this further as

$$r_{\text{global}} = \sum_{ij} \frac{A_{ij}}{2m} \tilde{x}_i \tilde{x}_j. \tag{S6}$$

Then we can calculate the local assortativity $r_\alpha(\ell)$ for scalar variables as

$$r_\alpha(\ell) = \sum_{ij} w_\alpha(i;\ell) \frac{A_{ij}}{k_i} \tilde{x}_i \tilde{x}_j. \tag{S7}$$

Fig. S1 gives some examples of distributions of r_{multi} for scalar attributes in the food web network.

C. Categorical Assortativity as a Correlation

The assortativity coefficient r_{global} for categorical attributes can be interpreted as a normalized Pearson's correlation. To see this, we start by observing that the Pearson's correlation of two binary variables is equivalent to the Phi coefficient for binary contingency tables (31). Table S1 shows a contingency table using the same notation as the directed assortativity, i.e., a and b give the marginal proportions and e gives the joint proportions.

Then the Pearson product–moment correlation of these variables is known as ϕ, which we derive using the moments of a Bernoulli distribution,

$$\phi = \frac{\mathbb{E}[y_i, y_j] - \mathbb{E}[y_i]\mathbb{E}[y_j]}{\sigma_{y_i}\sigma_{y_j}} \tag{S8}$$

$$= \frac{e_{11} - a_1 b_1}{\sqrt{a_1 a_0}\sqrt{b_1 b_0}}. \tag{S9}$$

Note that it is only necessary to calculate this in terms of e_{11}, since $e_{11} - a_1 b_1 = e_{00} - a_0 b_0$. We can see this using the identity $e_{00} = b_0 - a_1 + e_{11}$,

$$e_{00} - a_0 b_0 = b_0 - a_1 + e_{11} - (1 - a_1)(1 - b_1) \tag{S10}$$

$$= (1 - b_1) - a_1 + e_{11} - (1 - a_1 - b_1 + a_1 b_1) \tag{S11}$$

$$= e_{11} - a_1 b_1. \tag{S12}$$

A well-known issue with ϕ is that the extreme values of $+1$ and -1 are typically unobtainable, which can cause issues with its interpretation. In fact, $\phi = 1$ can only occur if $a_1 = b_1$, e.g., when the network is undirected, while $\phi = -1$ can only occur if $a_1 = b_2 = 0.5$ (32, 33). To address this issue, there have been a number of proposed normalizations to ensure the $\phi = 1$ is obtainable (34). One such normalization is the $\phi/\phi_{\max}$ proposed by Cureton (35),

$$\frac{\phi}{\phi_{\max}} = \frac{e_{11} - a_1 b_1}{\beta - a_1 b_1}, \tag{S13}$$

where β is the maximum possible value that e_{11} can take, i.e., $\min(a_1 b_1)$. Note that, for undirected networks,

$$\sqrt{a_1 b_1 a_2 b_2} = \sqrt{a_1^2 a_2^2} \tag{S14}$$

$$= a_1 a_2 \tag{S15}$$

$$= a_1(1 - a_1) \tag{S16}$$

$$= a_1 - a_1^2, \tag{S17}$$

which equals $\phi_{\max}$ when $a_1 \leq a_2$.

Then we can generalize $\phi/\phi_{\max}$ from binary to multicategory variables by treating each distinct value as a binary variable and taking their sum. If we set $\beta = 1$, then we obtain Eq. **S1**, and thus we recover Newman's assortativity (13). We also note that Eq. **S1** also corresponds to Cohen's κ that is frequently used to assess interrater agreement (36).

The normalization of the assortativity coefficient means that $r_{\min} \leq r \leq 1$ and

$$r_{\min} = -\frac{\sum_g a_g b_g}{1 - \sum_g a_g b_g}, \tag{S18}$$

which lies in the range $-1 \leq r_{\min} < 0$.

D. Assortativity as Autocorrelation of a Time Series

Assume a scalar attribute x_i on each node i of an undirected network. As mentioned in the main text, the probability of being at node i is stationary and proportional to the degree, $\pi_i = k_i/2m$. Given that a random walker is currently at node i, it moves to node j with probability A_{ij}/k_i.

We define a random time series, using the simple random walker, as the sequence of attributes of the nodes visited in the random walk, i.e., the value of the time series at time t is the attribute value x of the node visited at time t in the random walk. Asymptotically, the average value observed by the random walker is $\bar{x} = \sum_i \pi_i x_i = \sum k_i x_i / 2m$, and the variance is $\sigma^2 = \sum_i \pi_i x_i^2 - \bar{x}^2$.

Likewise, the autocovariance between the attribute observed at two consecutive steps (time lag of 1) is $R_x = \sum_{ij} \pi_i A_{ij}/k_i x_i x_j - \bar{x}^2$. Replacing x by $\tilde{x} = \frac{x-\bar{x}}{\sigma}$, we obtain the autocorrelation $R_{\tilde{x}} = \sum_{ij} \pi_i A_{ij}/k_i \tilde{x}_i \tilde{x}_j$, which coincides with r_{global} as defined in Eq. **S6**.

When faced with categorical data, we proceed as in *SI Text*, section C. We consider, for each type g of nodes, the scalar attribute x_g valued at 1 for nodes with type g and zero elsewhere. The modularity Q is therefore the sum for each type g of the autocovariance, $Q = \sum_g R_g$. As in *SI Text*, section C, this can be normalized in various ways, one of which is Newman's global assortativity as used in this article, which therefore represents a sort of categorical autocorrelation of the time series process of the categorical attributes observed by the stationary simple random walker.

E. Disconnected Networks

By using the personalized PageRank as a neighborhood function, it means that only nodes within the same connected component contribute to r_{multi}. Consequently, r_{multi} for each node is insensitive to whether multiple connected components are included.

F. Missing Values

It is common, when dealing with real datasets, that some values may be missing. This is the case for the Facebook 100 data, where a number of node attributes are missing. When considering the global assortativity, previous work has simply ignored contributions from missing data values (3). That is, only edges that connect nodes for which both the attribute values are known are considered when calculating e_{gh}. This treatment works fine for the global assortativity, because each edge counts equally. However, simply omitting missing values when calculating the local assortativity can cause a bias in the distribution. For example, consider the case when node ℓ and its immediate neighbors have missing values but, beyond those, the attribute values are known. For small values of α, the weight $w_\alpha(i;\ell)$ is largest for nodes with missing attribute values. Simply ignoring their edges would mean reassigning more weight to edges farther away from ℓ when normalizing to ensure that $\sum_{gh} e_{gh}(\ell) = 1$, a necessary step in calculating the assortativity. Then, when we examine the distribution of $r(\ell)$ across all nodes in the network, the resulting distribution will be a biased representation. To deal with this issue, we calculate each of the local assortativities as normal, but assign each a weight $z_\ell = \sum_{gh} e_{gh}(\ell)$, i.e., the sum of local edge counts before normalization. The weight z_ℓ describes our confidence in the local assortativity estimate from $z_\ell = 0$, indicating no confidence, to $z_\ell = 1$ when all node attributes within the neighborhood are known. We adjust for these weights when plotting the histograms in the main text.

G. Calculating the Personalized PageRank Vector

The personalized PageRank vector is the stationary distribution of a random walk with restarts. We calculate it by direct simulation of the random walk process using the power method:

$$w_\alpha(i;\ell)_{s+1} = \alpha \sum_j \frac{A_{ij}}{k_i} w_\alpha(j;\ell)_s + (1-\alpha)\delta_{i,\ell}, \qquad \textbf{[S19]}$$

and, at convergence, yields a distribution $w(i;\ell)$ with a mode at ℓ.

H. Integrating over α

To integrate over all values of α, we take advantage of the fact that we can equivalently write the ηth approximation the power method in Eq. **S19** as the ηth degree truncation of the power series (37),

$$w_\alpha(i;\ell)_\eta = \delta_{i,\ell} + \sum_{s=1}^{\eta} \alpha^s \left[\left(\frac{A_{i\ell}}{k_i}\right)^s - \left(\frac{A_{i\ell}}{k_i}\right)^{(s-1)}\right]. \qquad \textbf{[S20]}$$

By taking advantage of the relationship between α and the sequence of approximations computed by the power method, we can calculate the distribution $w_\alpha(i;\ell)$ for a given $\alpha = \alpha_0$ and use the sequence of approximations to calculate the distribution for any other α (37),

$$w_\alpha(i;\ell)_\eta = \delta_{i,\ell} + \sum_{s=1}^{\eta} \frac{\alpha^s}{\alpha_0^s} \left(w(i;\ell,\alpha_0)_s - w(i;\ell,\alpha_0)_{s-1}\right). \qquad \textbf{[S21]}$$

We can then integrate over all possible values of α (23),

$$w_{\text{multi}}(i;\ell)_\eta = \int_0^1 w_\alpha(i;\ell)_\eta \, \mathrm{d}\alpha \qquad \textbf{[S22]}$$

$$= \delta_{i,\ell} + \sum_{s=1}^{\eta} \frac{\left(w_{\alpha_0}(i;\ell)_s - w_{\alpha_0}(i;\ell)_{s-1}\right)}{(s+1)\alpha_0^s}. \qquad \textbf{[S23]}$$

I. Null Model Network Generation

We created a null model to generate networks with the same global assortativity as the observed network to compare the distributions of r_{multi}. For a fair comparison, we decided to keep the node degree and metadata label fixed while randomly rewiring the network. We do so using a modified version of the Markov chain Monte Carlo (MCMC) sampling of the configuration model for stub-labeled simple graphs (38) [for simple graphs, sampling from the space of stub-labeled graphs is equivalent to sampling from the space of vertex-labeled graphs (38)]. The modification is to ensure that we sample a graph with (approximately) the same global assortativity as the observed network. We achieve this by adding a rejection sampling step based on the binomial likelihood of observing the number of edges between nodes of the same type $m_{\text{in}} = m \sum_g e_{gg}$ given the proportion of edges required to maintain the global assortativity $\omega_{\text{in}} = \sum_g e_{gg}$,

$$L(G_i) = \log \binom{m}{m_{\text{in}}} (\omega_{\text{in}})^{m_{\text{in}}} (1-\omega_{\text{in}})^{m-m_{\text{in}}}. \qquad \textbf{[S24]}$$

The modified MCMC algorithm is shown in Algorithm 1.

```
Algorithm 1: stub-labeled MCMC.
Require: initial simple graph $G_0$, initial temp. $t_0$
Ensure: sequence of graphs $G_i$
  for $i <$ number of graphs to sample do
    choose two edges at random
    randomly choose one of the two possible swaps
    if edge swap would create a self-loop or multiedge then
      resample current graph: $G_i \leftarrow G_{i-1}$
    else
      if $Unif(0,1) < \exp\left(L(G_i) - L(G_{i-1})/t_i\right)$ then
        swap the chosen edges, producing $G_i$
      else
        reject $G_i$
    $t_{i+1} \leftarrow$ update$(t_i)$.
```

J. Datasets

Weddell Sea Food Web. The food web of the Antarctic Weddell Sea (4) consists of 488 species and 15,885 consumer relations. For each of the nodes in this network, we have five categorical attributes: Metabolic Category {*Plant*, *Ectotherm vertebrate*, *Endotherm vertebrate*, *Invertebrate*}, Feeding Type {*Carnivorous/necrovorous*, *Herbivorous/detrivorous*, *Detrivorous*, *Omnivorous*, *Primary producer*, *Carnivorous*}, FeedingMode {*Pelagic predator*, *Predator/scavenger*, *Primary producer*, *Predator*, *Deposit-feeder*, *Grazer*, *Suspension-feeder*}, Mobility {1, 2, 3, 4}, Environment {*Bathydemersal*, *Land-based*, *Resource*, *Pelagic*, *Benthopelagic*, *Benthic*, *Demersal*}. For scalar attributes, we use the mean mass of the species, mobility (although discrete, the values are ordinal), and node degree.

Facebook 100. The Facebook 100 dataset (3) contains an anonymized snapshot of the friendship connections among 1,208,316 users affiliated with the first 100 colleges admitted to Facebook. The dataset contains a total of 93,969,074 friendship edges between users of the same college. Each node has a set of categorical social variables: status {*undergraduate*, *graduate student*, *summer student*, *faculty*, *staff*, *alumni*}, dorm, major, gender {*male*, *female*}, and graduation year.

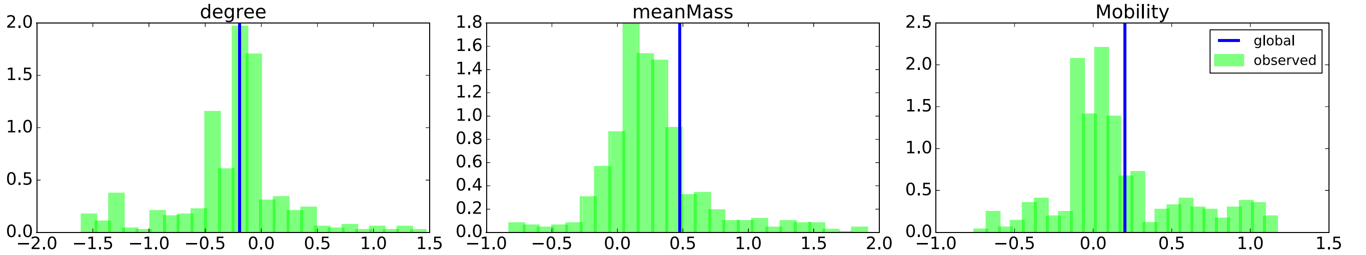

Fig. S1. Multiscale assortativity for different scalar attributes in the Weddell Sea Food Web: node degree, average species mass, and mobility. Note that mobility is a discrete ordinal variable (taking integer values in Eqs. **1** and **4**), and, in the main text, we treat it as an unordered discrete variable.

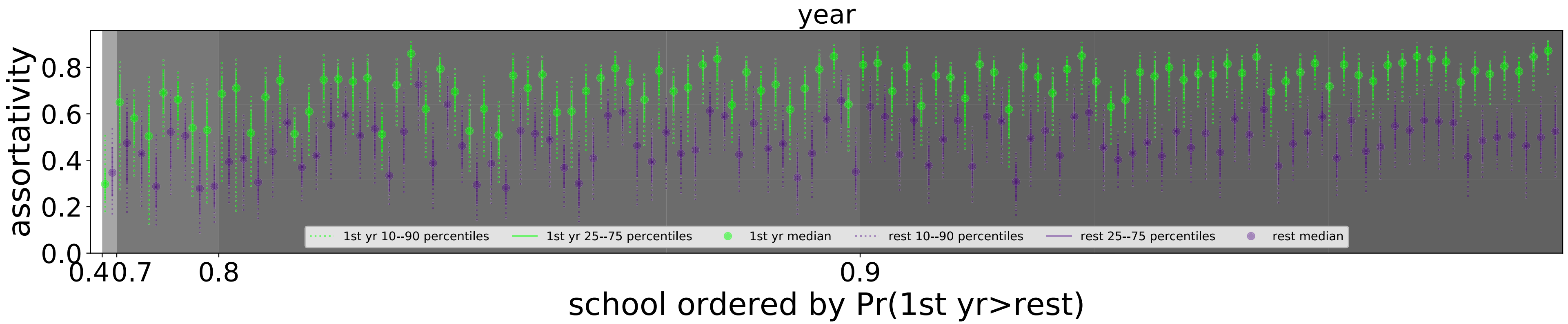

Fig. S2. Distributions of the local assortativity by year separated into first years and the rest of the students for each school. Schools are ordered by increasing proportion of first years that are more assortative than the rest of the students. First-year students are more assortative than the rest for all schools except for one.

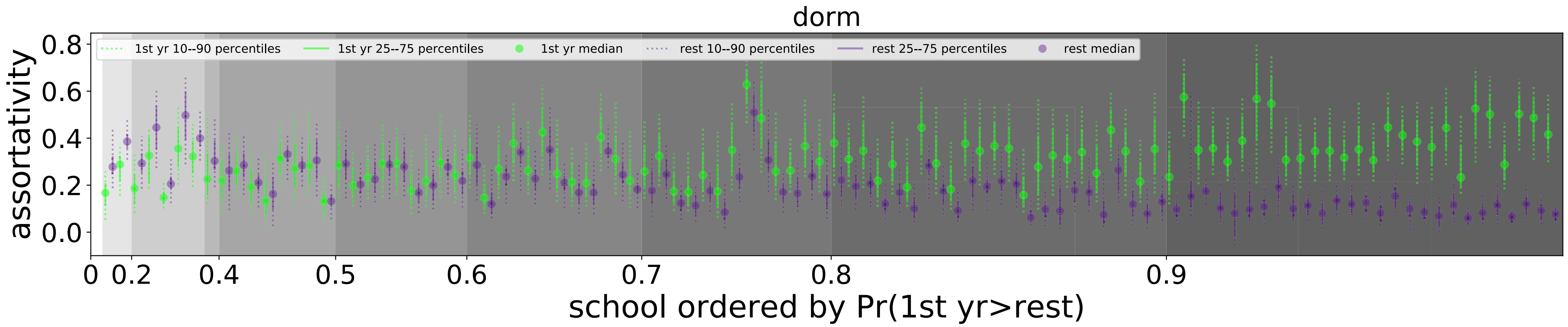

Fig. S3. Distributions of the local assortativity by residence (dorm) separated into first years and the rest of the students for each school. Schools are ordered by increasing proportion of first years that are more assortative than the rest of the students. In general, first-year students are more assortative than the rest; however, there are some schools in which the difference between first year and the rest is negligible. In a few schools, we observe that the first-year students are less assortative than the rest.

Table S1. Binary contingency table

	$y_j = 0$	$y_j = 1$	
$y_i = 0$	e_{00}	e_{01}	a_0
$y_i = 1$	e_{10}	e_{11}	a_1
	b_0	b_1	

Table S2. List of schools ordered by global assortativity

No.	School	r_{global}	No.	School	r_{global}
1	Amherst 41	0.081	51	William 77	0.203
2	Princeton 12	0.087	52	Emory 27	0.205
3	Trinity 100	0.106	53	UCLA 26	0.208
4	Stanford 3	0.109	54	Tennessee 95	0.209
5	Swarthmore 42	0.109	55	Wake 73	0.212
6	Johns Hopkins 55	0.110	56	MIT 8	0.219
7	Hamilton 46	0.113	57	UMass 92	0.222
8	Bowdoin 47	0.118	58	Berkeley 13	0.222
9	Harvard 1	0.120	59	USC 35	0.224
10	Brown 11	0.120	60	Temple 83	0.228
11	Dartmouth 6	0.126	61	UVA 16	0.230
12	Wellesley 22	0.127	62	Penn 94	0.231
13	Haverford 76	0.128	63	Northwestern 25	0.234
14	Wesleyan 43	0.128	64	Rutgers 89	0.235
15	UConn 91	0.129	65	UPenn 7	0.235
16	Tufts 18	0.130	66	Michigan 23	0.236
17	Williams 40	0.133	67	FSU 53	0.238
18	Reed 98	0.134	68	Cornell 5	0.238
19	Columbia 2	0.136	69	UC 64	0.251
20	BC 17	0.136	70	American 75	0.253
21	Duke 14	0.144	71	Notre Dame 57	0.255
22	Virginia 63	0.149	72	Rochester 38	0.256
23	Oberlin 44	0.151	73	Vassar 85	0.256
24	Villanova 62	0.158	74	Lehigh 96	0.258
25	Howard 90	0.159	75	Texas 80	0.261
26	WashU 32	0.162	76	USFCA 72	0.263
27	Georgetown 15	0.162	77	UC 61	0.265
28	Colgate 88	0.164	78	Syracuse 56	0.270
29	UF 21	0.165	79	Yale 4	0.273
30	BU 10	0.167	80	UCSB 37	0.277
31	Carnegie 49	0.171	81	Cal 65	0.279
32	GWU 54	0.171	82	Texas 84	0.291
33	Bingham 82	0.176	83	UChicago 30	0.291
34	NYU 9	0.182	84	Smith 60	0.292
35	UNC 28	0.185	85	Mississippi 66	0.297
36	Simmons 81	0.186	86	Baylor 93	0.297
37	USF 51	0.187	87	UIllinios 20	0.297
38	JMU 79	0.187	88	MU 78	0.306
39	UCF 52	0.187	89	Tulane 29	0.313
40	Santa 74	0.188	90	Mich 67	0.322
41	Northeastern 19	0.190	91	UGA 50	0.336
42	Maine 59	0.190	92	Wisconsin 87	0.338
43	Middlebury 45	0.190	93	UCSD 34	0.355
44	Brandeis 99	0.193	94	Indiana 69	0.356
45	Bucknell 39	0.194	95	UC 33	0.361
46	MSU 24	0.195	96	Auburn 71	0.370
47	Pepperdine 86	0.198	97	Oklahoma 97	0.397
48	Vermont 70	0.199	98	Caltech 36	0.426
49	Maryland 58	0.199	99	UCSC 68	0.480
50	Vanderbilt 48	0.201	100	Rice 31	0.504

The number given after each university name is the School Index and indicates the order in which they joined Facebook.