Supplementary Information for The Evolution of Homophily

Feng Fu[1,2], Martin A. Nowak[2,3,4], Nicholas A. Christakis[1,5,6], James H. Fowler[7,8]

[1] Department of Health Care Policy, Harvard Medical School, Boston, MA 02115, USA

[2] Program for Evolutionary Dynamics, Harvard University, Cambridge, MA 02138, USA

[3] Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA

[4] Department of Mathematics, Harvard University, Cambridge, MA 02138, USA

[5] Department of Medicine, Harvard Medical School, Boston, MA 02115, USA

[6] Department of Sociology, Faculty of Arts and Sciences, Harvard University, Cambridge, MA 02138, USA

[7] Division of Medical Genetics, University of California, San Diego, La Jolla, CA 92093, USA

[8] Department of Political Science, University of California, San Diego, La Jolla, CA 92093, USA

Classification: BIOLOGICAL SCIENCES – Population Biology

Manuscript information: 42 pages (including figure legends); 15 figures; 2 tables

Corresponding author: JHF, jhfowler@ucsd.edu

Contents

I. MODEL

Let us consider a haploid asexual population of size N. Each individual is marked by a phenotypic label, denoted by $G_i \in \{1, 2, \ldots, M\}$. Such phenotypes could refer to any observable trait, such as body mass, skin color, means of communication, and so on. Although we could also consider multidimensional phenotypic spaces, for simplicity, we assume a finite discrete phenotypic space represented by integers. We further assume that these individual phenotypes are neutral and symmetric to each other. That is, acquiring one of these phenotypes does not naturally grant an advantage or disadvantage in fitness, but instead it is the interactions between the focal individual and its social partners that determine its fitness. Specifically, if two interacting individuals have the same phenotype, they both obtain a; whereas if they have different phenotypes, they both obtain b. Thus the payoff matrix is

	G_i	$G_{j\neq i}$
G_i	a	b
$G_{j\neq i}$	b	a

. (1)

Accordingly the interactions can be categorized into three cases:

- $a = b$. All individuals obtain a constant payoff from each pairwise social interaction, regardless of the phenotypic similarity between them and their social partners.

- $a > b$. Social interactions between individuals yield more benefits to *coordination*. This payoff structure captures many real social interactions, where there are synergistic effects and the individuals involved are better off if they share the same phenotype.

- $a < b$. Social interactions between individuals yield more benefits to *anti-coordination*. This payoff structure also characterizes many real social interactions, where there are gains to specialization and individuals are better off if they complement each other in their phenotypes.

An individual's tendency to initiate interactions with others who have the same phenotype is described by a continuous preference $p_i \in [0, 1]$. With probability p_i, individual i chooses to initiate an interaction with an individual of the same phenotype; otherwise with probability $1 - p_i$, the individual chooses to initiate an interaction with an individual with a different phenotype. For two limiting cases, $p_i \to 1$ individuals choose to interact only with those who share the same phenotype (homophily) and $p_i \to 0$ individuals always choose to interact with individuals with

different phenotypes (heterophily). For other p_i values between 0 and 1, there is a balance between homophily and heterophily. We say an individual shows a tendency towards homophily if $p_i > 1/2$ ('birds of a feather flock together'); otherwise it shows a tendency towards heterophily ('opposites attract').

We assume that in order for an interaction to occur, both individuals must choose it. Thus, the probability of a social interaction between two individuals depends on their preferences and phenotypic similarity. Let us specifically consider two individuals i and j, with preferences p_i and p_j, respectively. If they share the same phenotypic trait, then with probability $p_i p_j$, they mutually select each other as social partners and thus they obtain a payoff a each. Whereas, if they have different phenotypic traits, then with probability $(1 - p_i)(1 - p_j)$, they mutually choose each other as social partners and so they obtain a payoff b each.

An individual's fitness can be expressed as an exponential function of its payoff, i.e., fitness = $\exp[\beta \cdot \text{payoff}]$, where β is the intensity of selection. For weak selection ($\beta \ll 1$), we have fitness $\sim 1 + \beta \cdot$ payoff; that is, individuals' payoffs are just small perturbations to their baseline fitness.

We assume individuals reproduce proportional to their fitness. Reproduction can be genetic (more fit individuals literally reproduce more) or cultural (less fit individuals copy the behavior of fitter individuals). Specifically an individual i is chosen with probability proportional to its fitness f_i. An offspring is being reproduced that replaces one random individual in the population. Reproduction is, however, subject to mutation. With probability v, the offspring chooses at random one of the M available phenotypes; otherwise it inherits the parental phenotype. An offspring inherits its parent's preference with probability $1 - u$; otherwise with probability u, it chooses a novel preference. We consider both global and local mutations. For local mutation, the offspring chooses a preference drawn from a Gaussian distribution with the parental value as mean and with a small standard deviation. For global mutation, the offspring adopts a preference that is randomly and uniformly drawn over the interval $[0, 1]$.

We are interested in the stationary distribution of preferences that is reached in this mutation-selection process. In what follows, we shall derive the analytical conditions for the evolution of homophily and perform extensive agent-based simulations to validate our analytical predictions.

II. THEORETICAL ANALYSIS

In our model, both phenotypes and preferences for homophily are heritable traits. To determine how these coevolve, we must determine which preference is most favored by natural selection (*i.e.*, performs the best) in the long run?

A. Mutation-selection equilibrium

For mathematical tractability and without loss of generality, let us start by analyzing the evolutionary dynamics involving an arbitrary set of $n \geq 2$ discrete preferences. In the limit of large $n \gg 1$, the results will approximate those for continuous preferences.

Denote by x_i the abundance of preference i ($i = 1, 2, \cdots, n$), $\sum_{j=1}^{n} x_j = 1$. Denote by G_i the phenotype of individual i, $G_i = \{1, 2, \cdots, M\}$. Let x_i^l be the abundance of individuals with preference i having the phenotype l, $\sum_{l=1}^{M} x_i^l = x_i$, and let x_*^l be the abundance of individuals with the phenotype l, $x_*^l = \sum_{k=1}^{n} x_k^l$.

Let $\mathbf{A} = \{\pi_{ij}\}$ be the symmetric $n \times n$ payoff matrix, where π_{ij} is the payoff to an individual with preference i interacting with an individual with preference j. In our model, individuals acquire payoffs only from each of their successful interactions. For the purpose of payoff accounting, we here use the expected payoff an individual obtains from each encounter. For notational simplicity, we can formally write

$$\pi_{ij} = \delta_{ij}\pi_{ij}^{\mathrm{S}} + (1 - \delta_{ij})\pi_{ij}^{\mathrm{D}}, \tag{2}$$

where $\delta_{ij} = 1$ if individuals i and j share the same phenotype ($G_i = G_j$), otherwise $\delta_{ij} = 0$, and the expected payoffs are given by

$$\pi_{ij}^{\mathrm{S}} = a p_i p_j, \tag{3}$$

$$\pi_{ij}^{\mathrm{D}} = b(1 - p_i)(1 - p_j). \tag{4}$$

We here consider a simple unbiased matching process, in which each individual is equally likely to meet with everyone else. Under this assumption, individuals meet each other in proportion to the abundance of each phenotype. We shall address matching bias in our analysis later on.

Specifically, the matching process can be described by an $n \times n$ matrix $\mathbf{Q} = \{q_{ij}\}$, where q_{ij} is the total number of meetings between individuals with preferences i and j per unit of time. For

notational convenience, we can formally write

$$q_{ij} = \delta_{ij} q_{ij}^{\mathrm{S}} + (1 - \delta_{ij}) q_{ij}^{\mathrm{D}}, \tag{5}$$

where

$$q_{ij}^{\mathrm{S}} = N^2 \sum_{l=1}^{M} x_i^l x_j^l, \tag{6}$$

$$q_{ij}^{\mathrm{D}} = N^2 \sum_{l=1}^{M} \sum_{r=1, r\neq l}^{M} x_i^l x_j^r. \tag{7}$$

Then the payoff of an individual with preference k can be formally written as $P_k = \sum_j \pi_{kj} q_{kj}/(x_k N)$. The fitness of an individual with preference k is given by $f_k = \exp(\beta P_k)$. The total population fitness is $\sum_k N x_k f_k$, where N is the population size. Evolutionary updating occurs according to a frequency-dependent Moran process (see the model description above). In an update event, the average change of x_k due to selection can be written as:

$$\Delta x_k^{\mathrm{sel}} = x_k \left(\frac{f_k}{\sum_j N x_j f_j} + 1 - \frac{1}{N} \right) - x_k. \tag{8}$$

For weak selection, the above equation can be linearized in the leading order of β:

$$\Delta x_k^{\mathrm{sel}} = \frac{\beta}{N} x_k \left(\sum_j \pi_{kj} q_{kj}/(x_k N) - \sum_i \sum_j x_i \pi_{ij} q_{ij}/(x_i N) \right) + O(\beta). \tag{9}$$

Averaging the above change over all possible population states, we can obtain the expected change in the stationary state in the leading order of β (from now on, for simplicity we will omit higher orders of β):

$$\left\langle \Delta x_k^{\mathrm{sel}} \right\rangle = \frac{\beta}{N^2} \left(\sum_j \langle \pi_{kj} q_{kj} \rangle - \sum_i \sum_j \left\langle x_k \pi_{ij} q_{ij} \right\rangle \right). \tag{10}$$

According to the perturbation method of games developed in refs. [1, 2], in the limit of weak selection the average $\langle \cdot \rangle$ can be taken over the stationary population state in neutral evolution when $\beta = 0$.

The average change of x_k due to preference mutation is given by $[(1 - x_k)/n - x_k(n-1)/n]/N = (1/n - x_k)/N$. Thus the total expected change $\Delta x_k^{\mathrm{tot}}$ in the mutation-selection equilibrium is:

$$\langle \Delta x_k^{\mathrm{tot}} \rangle = (1 - u) \langle \Delta x_k^{\mathrm{sel}} \rangle + \frac{u}{N} \langle (\frac{1}{n} - x_k) \rangle = 0. \tag{11}$$

It follows that the stationary frequency $\langle x_k \rangle$ is given by:

$$\langle x_k \rangle = \frac{1}{n} + \frac{\beta(1-u)}{Nu} \left(\sum_j \langle \pi_{kj} q_{kj} \rangle - \sum_i \sum_j \left\langle x_k \pi_{ij} q_{ij} \right\rangle \right). \tag{12}$$

B. Selection criteria

We say natural selection favors a preference k if its abundance is greater than $1/n$ in the stationary state. We thus obtain the condition for preference k to be favored by natural selection:

$$\sum_j \langle \pi_{kj} q_{kj} \rangle - \sum_i \sum_j \left\langle x_k \pi_{ij} q_{ij} \right\rangle > 0. \tag{13}$$

Following prior work [2, 3], we can simplify the calculations using the symmetry condition. Under neutral evolution, different preferences can be seen as different types of individuals, and hence index permutations do not result in any changes. Here we need to consider only five cases [3]: $i = j = k$, $i = j \neq k$, $i = k \neq j$, $j = k \neq i$, and $i \neq j \neq k \neq i$.

We can further rewrite the condition (13) as

$$\begin{aligned}
&\sum_j \langle \pi_{kj} q_{kj} \rangle - \sum_i \sum_j \left\langle x_k \pi_{ij} q_{ij} \right\rangle \\
=\ & \sum_j \langle \pi_{kj} q_{kj} \rangle - \pi_{kk} \langle x_k q_{kk} \rangle - \sum_{i \neq k} \pi_{ii} \langle x_k q_{ii} \rangle - \sum_{j \neq k} \pi_{kj} \langle x_k q_{kj} \rangle - \sum_{i \neq k} \pi_{ik} \langle x_k q_{ik} \rangle - \sum_{i,j \neq k} \pi_{ij} \langle x_k q_{ij} \rangle \\
=\ & \pi_{kk} \langle q_{kk} \rangle + \langle q_{kj} \rangle \left(\sum_j \pi_{kj} - \pi_{kk} \right) - \pi_{kk} \langle x_k q_{kk} \rangle - \langle x_k q_{ii} \rangle \left(\sum_i \pi_{ii} - \pi_{kk} \right) - \langle x_k q_{kj} \rangle \left(\sum_j \pi_{kj} - \pi_{kk} \right) \\
& - \langle x_k q_{ik} \rangle \left(\sum_i \pi_{ik} - \pi_{kk} \right) - \langle x_k q_{ij} \rangle \left(\sum_{i,j} \pi_{ij} - \sum_i \pi_{ki} - \sum_i \pi_{ik} - \sum_i \pi_{ii} + 2\pi_{kk} \right)
\end{aligned} \tag{14}$$

Recollecting the terms in the above equation leads to

$$\begin{aligned}
&\sum_j \langle \pi_{kj} q_{kj} \rangle - \sum_i \sum_j \left\langle x_k \pi_{ij} q_{ij} \right\rangle \\
=\ & \pi_{kk} \underbrace{\left(\langle q_{kk} \rangle - \langle q_{kj} \rangle - \langle x_k q_{kk} \rangle + \langle x_k q_{ii} \rangle + \langle x_k q_{kj} \rangle + \langle x_k q_{ik} \rangle - 2\langle x_k q_{ij} \rangle \right)}_{\alpha} \\
& + \frac{\sum_i \pi_{ii}}{n} \underbrace{n \left(\langle x_k q_{ij} \rangle - \langle x_k q_{ii} \rangle \right)}_{\delta} \\
& + \frac{\sum_j \pi_{kj}}{n} \underbrace{n \left(\langle q_{kj} \rangle - \langle x_k q_{kj} \rangle + \langle x_k q_{ij} \rangle \rangle \right)}_{\beta} \\
& + \frac{\sum_j \pi_{jk}}{n} \underbrace{n \left(\langle x_k q_{ij} \rangle - \langle x_k q_{jk} \rangle \right)}_{\gamma} + \frac{\sum_{i,j} \pi_{ij}}{n^2} \underbrace{\left(-n^2 \langle x_k q_{ij} \rangle \right)}_{\epsilon}.
\end{aligned} \tag{15}$$

Using the symmetry condition we can further show that

$$\begin{aligned}\alpha+\delta &= -(n-1)\langle x_k q_{ii}\rangle + (n-2)\langle x_k q_{ij}\rangle + \langle q_{kk}\rangle - \langle q_{kj}\rangle - \langle x_k q_{kk}\rangle + \langle x_k q_{kj}\rangle + \langle x_k q_{jk}\rangle \\ &= -(n-1)\langle x_k q_{ii}\rangle + (n-2)\langle x_k q_{ij}\rangle + \langle q_{kk}\rangle - \langle q_{kj}\rangle - \langle q_{kk}\rangle + (n-1)\langle x_k q_{ii}\rangle \\ &\quad + \langle q_{kj}\rangle - \langle x_k q_{ik}\rangle - (n-2)\langle x_k q_{ij}\rangle + \langle x_k q_{ik}\rangle = 0, \end{aligned} \tag{16}$$

and also

$$\begin{aligned}\beta+\gamma+\epsilon &= n\Big(\langle q_{kj}\rangle + 2\langle x_k q_{ij}\rangle - \langle x_k q_{kj}\rangle - \langle x_k q_{jk}\rangle - n\langle x_k q_{ij}\rangle\Big) \\ &= n\left(\sum_i x_i\langle q_{kj}\rangle + 2\langle x_k q_{ij}\rangle - \langle x_k q_{kj}\rangle - \langle x_k q_{jk}\rangle - n\langle x_k q_{ij}\rangle\right) \\ &= n\Big((n-2)\langle x_k q_{ij}\rangle + \langle x_k q_{kj}\rangle + \langle x_k q_{jk}\rangle - \langle x_k q_{kj}\rangle - \langle x_k q_{jk}\rangle + (2-n)\langle x_k q_{ij}\rangle\Big) = 0. \end{aligned} \tag{17}$$

Denote $\overline{\pi_{k*}} = \sum_j \pi_{kj}/n$, $\overline{\pi_{*k}} = \sum_j \pi_{jk}/n$, $\overline{\pi_{**}} = \sum_j \pi_{jj}/n$, and $\bar{\pi} = \sum_i \sum_j \pi_{ij}/n^2$. The preference selection condition now has the following form [3]:

$$\lambda_1(\pi_{kk} - \overline{\pi_{**}}) + \lambda_2(\overline{\pi_{k*}} - \overline{\pi_{*k}}) + \lambda_3(\overline{\pi_{k*}} - \bar{\pi}) > 0, \tag{18}$$

where we have

$$\lambda_1 \propto \langle x_k q_{ii}\rangle - \langle x_k q_{ij}\rangle, \tag{19}$$

$$\lambda_2 \propto \langle x_k q_{jk}\rangle - \langle x_k q_{ij}\rangle, \tag{20}$$

$$\lambda_3 \propto n\langle x_k q_{ij}\rangle. \tag{21}$$

Substituting Eqs. (2) and (5) into the above condition, we finally arrive at the preference selection condition

$$\begin{aligned}&\left[\lambda_1^{\mathrm{S}}(\pi_{kk}^{\mathrm{S}} - \overline{\pi_{**}^{\mathrm{S}}}) + \lambda_2^{\mathrm{S}}(\overline{\pi_{k*}^{\mathrm{S}}} - \overline{\pi_{*k}^{\mathrm{S}}}) + \lambda_3^{\mathrm{S}}(\overline{\pi_{k*}^{\mathrm{S}}} - \overline{\pi^{\mathrm{S}}})\right] \\ +\ &\left[\lambda_1^{\mathrm{D}}(\pi_{kk}^{\mathrm{D}} - \overline{\pi_{**}^{\mathrm{D}}}) + \lambda_2^{\mathrm{D}}(\overline{\pi_{k*}^{\mathrm{D}}} - \overline{\pi_{*k}^{\mathrm{D}}}) + \lambda_3^{\mathrm{D}}(\overline{\pi_{k*}^{\mathrm{D}}} - \overline{\pi^{\mathrm{D}}})\right] > 0, \end{aligned} \tag{22}$$

where, up to a positive common factor, these structural coefficients are given by

$$\lambda_1^{\mathrm{S}} \propto \langle x_k x_i^l x_i^l\rangle - \langle x_k x_i^l x_j^l\rangle, \tag{23}$$

$$\lambda_2^{\mathrm{S}} \propto \langle x_k x_k^l x_j^l\rangle - \langle x_k x_i^l x_j^l\rangle, \tag{24}$$

$$\lambda_3^{\mathrm{S}} \propto n\langle x_k x_i^l x_j^l\rangle, \tag{25}$$

and

$$\lambda_1^{\mathrm{D}} \propto \langle x_k x_i^l x_i^r \rangle - \langle x_k x_i^l x_j^r \rangle, \tag{26}$$

$$\lambda_2^{\mathrm{D}} \propto \langle x_k x_k^l x_j^r \rangle - \langle x_k x_i^l x_j^r \rangle \tag{27}$$

$$\lambda_3^{\mathrm{D}} \propto n\langle x_k x_i^l x_j^r \rangle. \tag{28}$$

Note that, in the above formula, we write $\sum_{l=1}^{M} x_i^l x_j^l$ as $x_i^l x_j^l$ and $\sum_{l=1}^{M} \sum_{r=1, r\neq l}^{M} x_i^l x_j^r$ as $x_i^l x_j^r$ for notational simplicity.

C. Formula for continuous preferences

The selection condition derived for multiple discrete preferences can be extended to the case of continuous preferences [4]. To do this, we partition the interval $[0,\ 1]$ into n small segments of equal length. For sufficiently large number of n, the sums in the condition (22) can be replaced by integrals. Note that in this continuum limit, the stationary abundance of preferences takes the form of the probability density function of the preference distribution. The condition for preference p to be selected becomes

$$\begin{aligned}
\mathcal{D}(p) \;=\; & \lambda_1^{\mathrm{S}}[A^{\mathrm{S}}(p,p) - \int_0^1 A^{\mathrm{S}}(q,q)dq] + \lambda_2^{\mathrm{S}}[\int_0^1 A^{\mathrm{S}}(p,q)dq - \int_0^1 A^{\mathrm{S}}(q,p)dq] \\
+\; & \lambda_3^{\mathrm{S}}[\int_0^1 A^{\mathrm{S}}(p,q)dq - \int_0^1 \int_0^1 A^{\mathrm{S}}(p,q)dpdq] + \lambda_1^{\mathrm{D}}[A^{\mathrm{D}}(p,p) - \int_0^1 A^{\mathrm{D}}(q,q)dq] \\
+\; & \lambda_2^{\mathrm{D}}[\int_0^1 A^{\mathrm{D}}(p,q)dq - \int_0^1 A^{\mathrm{D}}(q,p)dq] + \lambda_3^{\mathrm{D}}[\int_0^1 A^{\mathrm{D}}(p,q)dq - \int_0^1 \int_0^1 A^{\mathrm{D}}(p,q)dpdq] \\
& > 0,
\end{aligned} \tag{29}$$

where $A^{\mathrm{S}}(p,q) = apq$ and $A^{\mathrm{D}}(p,q) = b(1-p)(1-q)$, and these structural coefficients λ_i^{S}'s and λ_i^{D}'s are the same as given for multiple discrete preferences.

D. Triplet correlations at neutrality

Using coalescent theory [5], we are able to calculate these structural coefficients λ_i^{S}'s and λ_i^{D}'s. For brevity, we here only sketch the calculation procedures. For a more detailed description of coalescent theory and its recent combination with evolutionary game theory, we refer to refs. [1–3, 5].

Let us first interpret these triplet correlations. Take the average $\langle x_k x_i^l x_i^l \rangle$ for an example. It is the expected probability that when we randomly choose three individuals from the population, two of them have the same preference and phenotype, while the third one has a different preference. Other triplet correlations can be explained similarly.

It is helpful to compute coalescence in the continuous time limit, τ, by rescaling discrete time with $\tau = 2t/N^2$. In this new time scale, we adopt the rescaled mutation rates, $\mu = Nu$ for preference mutations, and $\nu = Nv$ for phenotypic mutations. The trick of coalescent theory is tracing any two individual lineages backward in time. After a certain time τ_2 (the 'coalescent time'), we can always find their most recent common ancestor. The probability density of τ_2 is given by [5]

$$T_2(\tau_2) = e^{-\tau_2}. \tag{30}$$

We can also obtain the coalescent time density function $T_3(\tau_2, \tau_3)$ for three randomly chosen individuals. In this case, any two of them coalesce first back at time τ_3, and then this lineage coalesces with the remaining one back at time τ_2. We thus have [5]

$$T_3(\tau_2, \tau_3) = 3e^{-3\tau_3} e^{-\tau_2}. \tag{31}$$

We identify two identical individuals (*i.e.*, they both have the same preference and phenotype) immediately after the coalescence of the two chosen lineages. For the Moran process in our model, the preference along each lineage mutates with rate $Nu/2 = \mu/2$ and the phenotype with rate $Nv/2 = \nu/2$. If at least one preference mutation occurs along the two lineages after their coalescence, the two individuals have the same preference with probability $1/n$. Similarly, if at least one phenotypic mutation happens along the two lineages after their coalescence, the two individuals still have the same phenotypes with probability $1/M$. We can obtain the probability that for two randomly chosen individuals, they still have the same preference (have the same phenotype) after their coalescent time τ_2, respectively:

$$s_2(\tau_2) = e^{-\mu\tau_2} + \frac{1 - e^{-\mu\tau_2}}{n}, \tag{32}$$

$$g_2(\tau_2) = e^{-\nu\tau_2} + \frac{1 - e^{-\nu\tau_2}}{M}. \tag{33}$$

In the same vein, the probability $s_3(\tau_2, \tau_3)$ that three randomly chosen individuals all have the same preference after their coalescent time $\tau_2 + \tau_3$ is given by [2]

$$\begin{aligned} s_3(\tau_2, \tau_3) = & \frac{1}{n^2}\Big[s_2(\tau_2)\Big(1 + 3(n-1)e^{-\mu\tau_3} + (n-1)(n-2)e^{-3/2\mu\tau_3}\Big) \\ & + (1 - s_2(\tau_2))\Big(1 + (n-3)e^{-\mu\tau_3} - (n-2)e^{-3/2\mu\tau_3}\Big)\Big] \end{aligned} \tag{34}$$

Based on these coalescence quantities, we now can calculate the following pair/triplet correlations associated with λ_i^{S}:

$$\langle x_k^l x_k^l \rangle = \frac{1}{n}\int_0^\infty T_2(\tau_2) s_2(\tau_2) g_2(\tau_2) d\tau_2, \tag{35}$$

$$\langle x_k x_k^l x_k^l \rangle = \frac{1}{n}\frac{1}{3}\int_0^\infty \int_0^\infty T_3(\tau_2,\tau_3) s_3(\tau_2,\tau_3)[g_2(\tau_3) + g_2(\tau_2+\tau_3) + g_2(\tau_2+\tau_3)] d\tau_2 d\tau_3, \tag{36}$$

$$\begin{aligned}\langle x_k x_k^l x_*^l \rangle &= \frac{1}{n}\frac{1}{3}\int_0^\infty \int_0^\infty T_3(\tau_2,\tau_3)[s_2(\tau_3) g_2(\tau_2+\tau_3) + s_2(\tau_2+\tau_3) g_2(\tau_3) \\ &+ s_2(\tau_2+\tau_3) g_2(\tau_2+\tau_3)] d\tau_2 d\tau_3.\end{aligned} \tag{37}$$

The factor $1/n$ in the right hand side of above equations accounts for the probability of randomly selecting an individual with preference k. This probability is given by $1/n$ under neutral evolution.

Using symmetry, we can directly calculate other correlations as listed below:

$$\langle x_i^l x_j^l \rangle = \frac{1}{n-1}\left(\langle x_i^l x_*^l \rangle - \langle x_i^l x_i^l \rangle\right), \tag{38}$$

$$\langle x_k x_k^l x_j^l \rangle = \frac{1}{n-1}\left(\langle x_k x_k^l x_*^l \rangle - \langle x_k x_k^l x_k^l \rangle\right), \tag{39}$$

$$\langle x_k x_i^l x_j^l \rangle = \frac{1}{n-2}\left(\langle x_i^l x_j^l \rangle - 2\langle x_k x_k^l x_j^l \rangle\right), \tag{40}$$

$$\langle x_k x_i^l x_i^l \rangle = \frac{1}{n-1}\left(\langle x_i^l x_i^l \rangle - \langle x_k x_k^l x_k^l \rangle\right), \tag{41}$$

where we use $x_*^l = x_i^l + \sum_{j\neq i} x_j^l$ to get the first two identities, and for the last two we use $x_k = 1 - \sum_{j\neq k} x_j$.

Analogously, we calculate the triplet correlations associated with λ_i^{D}. The probability $z_2(\tau_2)$ of two randomly chosen individuals having different phenotypes since their coalescent time τ_2 is given by:

$$z_2(\tau_2) = \frac{M-1}{M}\left(1 - e^{-\nu\tau_2}\right). \tag{42}$$

Then we can calculate the following pair/triplet correlations:

$$\langle x_i^l x_i^r \rangle = \frac{1}{n}\int_0^\infty T_2(\tau_2) s_2(\tau_2) z_2(\tau_2) d\tau_2, \tag{43}$$

$$\langle x_k x_k^l x_k^r \rangle = \frac{1}{n}\frac{1}{3}\int_0^\infty \int_0^\infty T_3(\tau_2,\tau_3) s_3(\tau_2,\tau_3)[z_2(\tau_3) + z_2(\tau_2+\tau_3) + z_2(\tau_2+\tau_3)] d\tau_2 d\tau_3, \tag{44}$$

$$\begin{aligned}\langle x_k x_k^l x_*^r \rangle &= \frac{1}{n}\frac{1}{3}\int_0^\infty \int_0^\infty T_3(\tau_2,\tau_3)[s_2(\tau_3) z_2(\tau_2+\tau_3) + s_2(\tau_2+\tau_3) z_2(\tau_3) \\ &+ s_2(\tau_2+\tau_3) z_2(\tau_2+\tau_3)] d\tau_2 d\tau_3,\end{aligned} \tag{45}$$

$$\langle x_i^l x_j^r \rangle = \frac{1}{n(n-1)}\int_0^\infty T_2(\tau_2)(1 - s_2(\tau_2)) z_2(\tau_2) d\tau_2. \tag{46}$$

We should note that the factor $1/[n(n-1)]$ in Eq. (46) takes into account the fact that there are $n(n-1)$ different combinations of the preference index satisfying $i \neq j$.

Again, we can calculate the following correlations using symmetry:

$$\langle x_k x_k^l x_j^r \rangle = \frac{1}{n-1}\left(\langle x_k x_k^l x_*^r \rangle - \langle x_k x_k^l x_k^r \rangle\right), \tag{47}$$

$$\langle x_k x_i^l x_j^r \rangle = \frac{1}{n-2}\left(\langle x_i^l x_j^r \rangle - 2\langle x_k x_k^l x_j^r \rangle\right), \tag{48}$$

$$\langle x_k x_i^l x_i^r \rangle = \frac{1}{n-1}\left(\langle x_i^l x_i^r \rangle - \langle x_k x_k^l x_k^r \rangle\right). \tag{49}$$

Omitting some intermediate calculations, below we give the explicit expressions of λ_i^{S} and λ_i^{D} in terms up to a same positive common factor:

$$\lambda_1^{\mathrm{S}} \propto (1+\nu)(3+\mu+\nu)(M(2+\mu)(3+3\mu+2\nu)+\nu(4+3\mu+2\nu)), \tag{50}$$

$$\begin{aligned}\lambda_2^{\mathrm{S}} \propto\ & M(2+\mu)\left(9+3\mu(4+\mu)+6\nu+5\mu\nu+\nu^2\right)\\ &+\ \nu\left(3\mu^3+2(2+\nu)(3+\nu)^2+\mu^2(21+8\nu)+\mu(49+\nu(38+7\nu))\right),\end{aligned} \tag{51}$$

$$\begin{aligned}\lambda_3^{\mathrm{S}} \propto\ & \mu\Big[M(2+\mu)\left(9+3\mu(4+\mu)+7\nu+5\mu\nu+2\nu^2\right)\\ &+\ \nu\left(34+3\mu^3+40\nu+2\nu^2(8+\nu)+\mu(3+\nu)(16+7\nu)+\mu^2(21+8\nu)\right)\Big].\end{aligned} \tag{52}$$

$$\lambda_1^{\mathrm{D}} \propto (M-1)\nu(1+\nu)(3+\mu+\nu)(4+3\mu+2\nu), \tag{53}$$

$$\lambda_2^{\mathrm{D}} \propto (M-1)\nu\left(3\mu^3+2(2+\nu)(3+\nu)^2+\mu^2(21+8\nu)+\mu(49+\nu(38+7\nu))\right), \tag{54}$$

$$\lambda_3^{\mathrm{D}} \propto (M-1)\mu\nu\left(34+3\mu^3+40\nu+2\nu^2(8+\nu)+\mu(3+\nu)(16+7\nu)+\mu^2(21+8\nu)\right). \tag{55}$$

We can see that these λ expressions indeed do not depend on the number of preferences [3], but are jointly determined by the number of phenotypes M, the preference mutation rate μ, and the phenotypic mutation rate ν.

Although we derive the selection condition specifically for our model, the same derivations work for other evolutionary updating rules, for example, the Wright-Fisher process, as the formula using the coalescent theory is robust to such variations [5].

Having calculated the structural coefficients λ_i^{S}'s and λ_i^{D}'s, we can derive the relative abundance (probability density) of any preference p. Based on Eq. (12), we can also calculate its equilibrium frequency (probability density).

E. Conditions for the evolution of homophily

After some algebra, we obtain the following condition. Preference p is favored by natural selection (its stationary frequency is greater than the neutral average) if and only if

$$\mathcal{D}(p) = C_2(a,b,M,\mu,\nu)p^2 + C_1(a,b,M,\mu,\nu)p + C_0(a,b,M,\mu,\nu) > 0, \tag{56}$$

where the coefficients C_i's are

$$\begin{aligned}
C_2 &= (1+\nu)(3+\mu+\nu)(b(-1+M)\nu(4+3\mu+2\nu) + a(M(2+\mu)(3+3\mu+2\nu) + \nu(4+3\mu+2\nu))),\\
C_1 &= \frac{1}{2}\Big(-b(-1+M)\nu\Big(3\mu^4 + 8(1+\nu)(2+\nu)(3+\nu) + \mu^3(21+8\nu) + \mu^2(60+7\nu(7+\nu)) +\\
&\quad +2\mu(43+\nu(4+\nu)(14+\nu))\Big) + a\mu\Big(M(2+\mu)\Big(9+3\mu(4+\mu)+7\nu+5\mu\nu+2\nu^2\Big) +\\
&\quad +\nu\Big(34+3\mu^3+40\nu+2\nu^2(8+\nu)+\mu(3+\nu)(16+7\nu)+\mu^2(21+8\nu)\Big)\Big)\Big),\\
C_0 &= \frac{1}{24}\Big(-a\Big(M(2+\mu)\Big(9\mu^3+4(1+\nu)(3+\nu)(3+2\nu)+3\mu^2(16+9\nu)+\mu(3+2\nu)(25+13\nu)\Big) +\\
&\quad +\nu\Big(9\mu^4+8(1+\nu)(2+\nu)(3+\nu)+3\mu^2(4+\nu)(13+7\nu)+3\mu^3(21+8\nu) +\\
&\quad +2\mu(7+3\nu)(11+\nu(9+\nu))\Big)\Big) + b(-1+M)\nu\Big(9\mu^4+16(1+\nu)(2+\nu)(3+\nu) +\\
&\quad +3\mu^3(21+8\nu)+3\mu^2(56+\nu(45+7\nu))+\mu(206+2\nu(132+\nu(44+3\nu)))\Big)\Big).
\end{aligned} \tag{57}$$

Natural selection acts against preference p if $\mathcal{D}(p) < 0$. The quadratic function $\mathcal{D}(p)$ represents a series of parabolas, depending on the model parameters (Figs. S1 and S2). We can prove that there always exists at least one interior root $p_c \in (0,1)$ satisfying $\mathcal{D}(p) = 0$. We call such p_c the critical preference for homophily.

The population tends to evolve homophilic preferences, *i.e.*, the population average $\langle p \rangle > 1/2$, if and only if

$$\int_0^1 \mathcal{D}(p)dp > 0, \tag{58}$$

which leads to the following simplified condition:

$$a > Kb, \tag{59}$$

where the term K is given by

$$K = \frac{\nu(\mu+\nu+2)(M-1)}{\nu(\mu+\nu+2)+(\mu+2\nu+3)M} \tag{60}$$

We note that the function K is always positive and monotonically increases with the preference mutation rate, μ, the phenotypic mutation rate, ν, and the number of phenotypes, M.

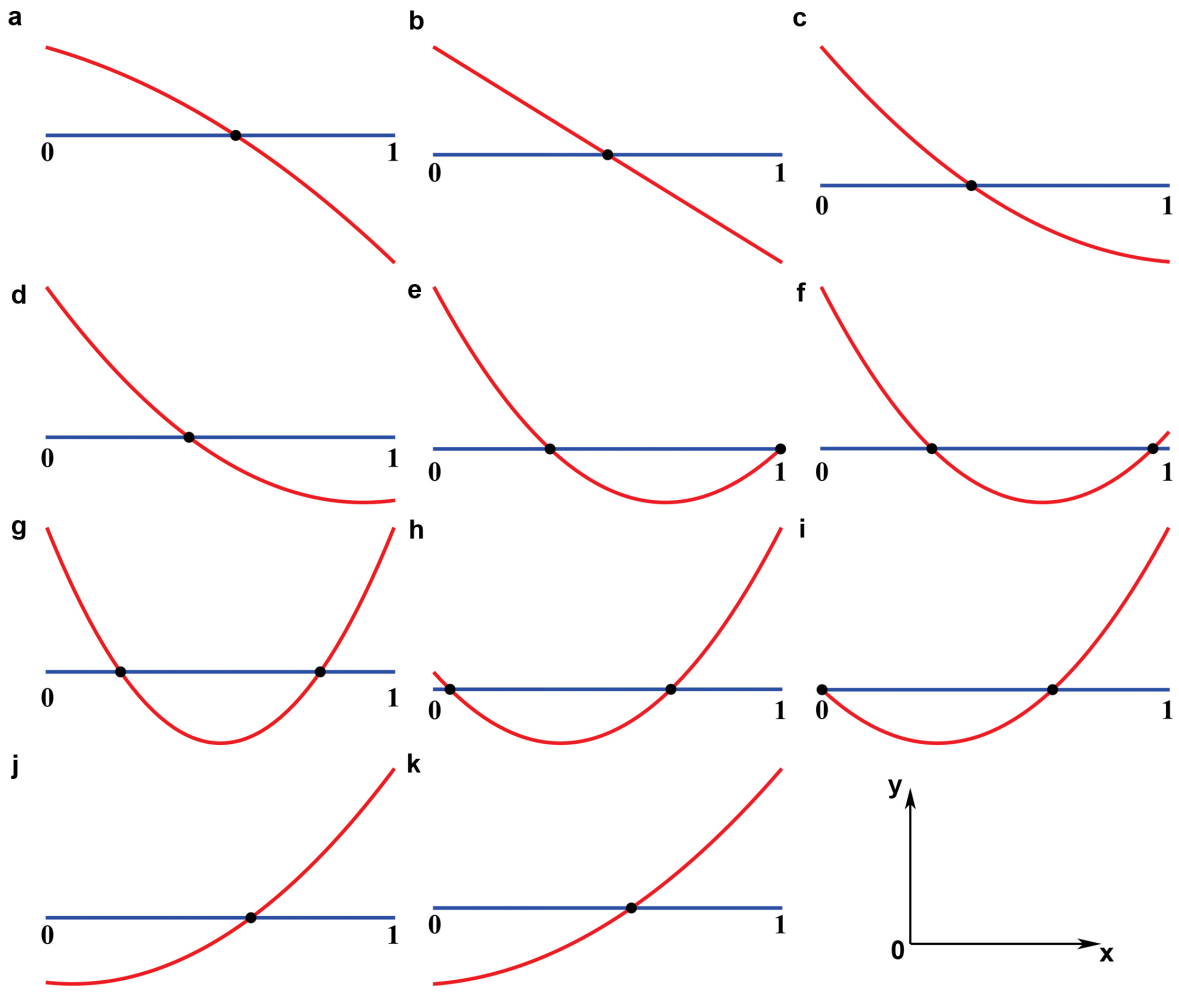

FIG. S1: Plots of the selection condition function $\mathcal{D}(p)$ for increasing payoff values a at fixed payoff value b ($b > 0$). Parameters: $\mu = 1.2$, $\nu = 1.8$, $M = 3$, $b = 1$, (**a**–**k**), respectively, $a = -3, -14/41, 100/683, 150/683, 562/1475, 24/59, 5/9, 700/901, 758/901, 220/109, 5$.

In the main text, we focus on the results for positive payoff values a and b. In what follows, we shall provide general results for the whole parameter space (b, a) for completeness. Positive payoff values suggest synergistic (beneficial) interactions, while negative ones suggest antagonistic interactions.

When both types of associations are beneficial ($a, b > 0$), increases in μ, ν, and M mean that the benefit to coordination must be higher in order for homophily to evolve. In addition to the results ($a, b > 0$) shown in the main text, Fig. S3 displays the results with the full set of (b, a) points in $[-0.5, 0.5]^2$. Again, we find excellent agreement between our theoretical predictions and simulation results.

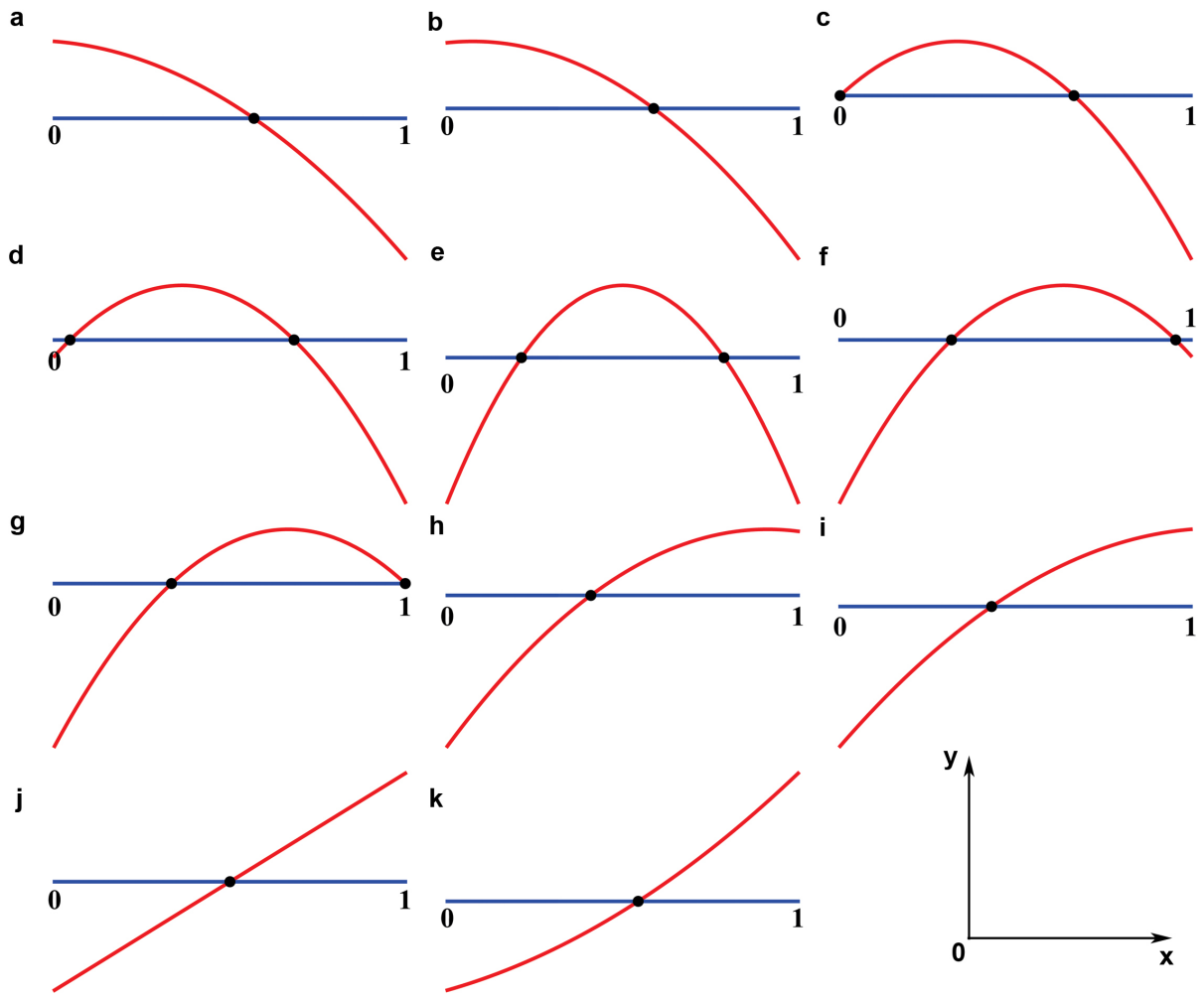

FIG. S2: Plots of the selection condition function $\mathcal{D}(p)$ for increasing payoff value a at fixed payoff value b ($b < 0$). These plots are the 'mirror' cases with respect to Fig. S1. Parameters: $\mu = 1.2$, $\nu = 1.8$, $M = 3$, $b = -1$, (**a**–**k**), respectively, $a = -5, -220/109, -758/901, -700/901, -5/9, -24/59, -562/1475, -150/683, -100/683, 14/41, 3$.

We can understand these results as follows. When natural selection would otherwise favor homophily, any increase in (unbiased) mutations in preferences tends to brings down the population equilibrium to $1/2$. Increasing the phenotypic mutation raises the chance of meeting individuals of different phenotypes and, consequently, the overall benefit of heterophilic interactions. Likewise, increasing the number of phenotypic variations, M, makes the population more diverse, and in effect reduces the average number of individuals having the same phenotype. As a result, increasing M also requires a larger advantage to homophilic interactions to offset the potential benefit from interacting with numerous individuals of different phenotypes.

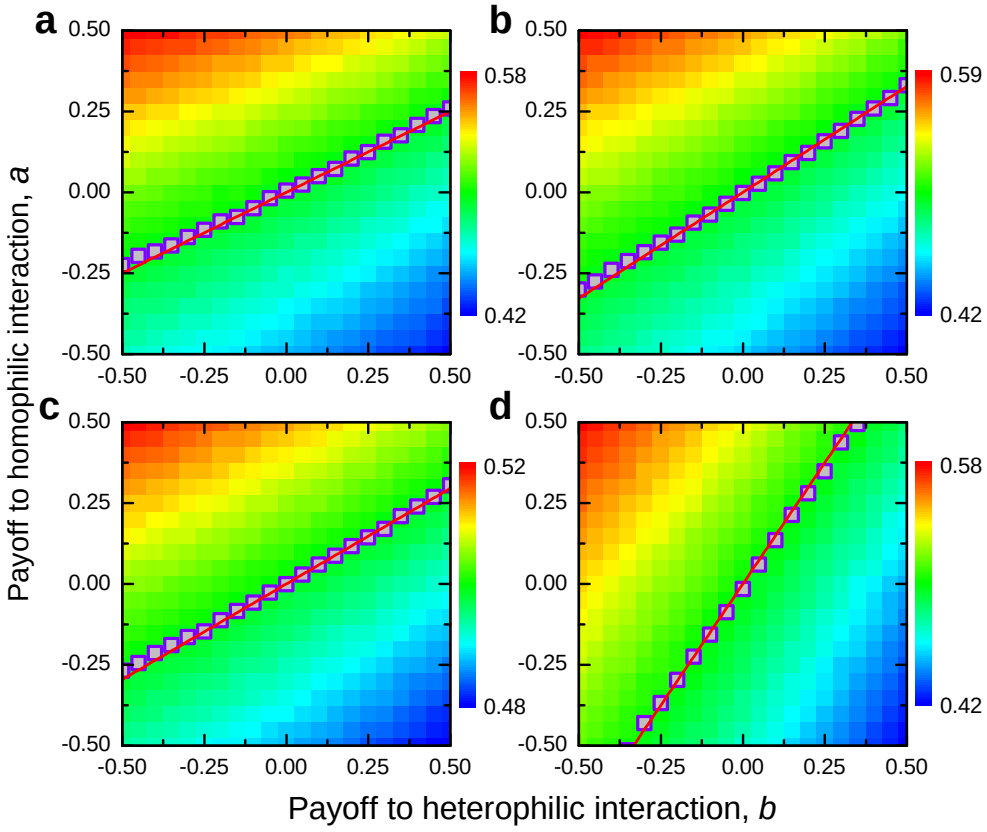

FIG. S3: Population average $\langle p \rangle$ as a function of $(b,\ a)$. Parameters: $N = 50$, $\beta = 5 \times 10^{-3}$, (**a**) $M = 5$, $u = 0.04$, $v = 0.02$, (**b**)$M = 20$, $u = 0.04$, $v = 0.02$, (**c**) $M = 5$, $u = 0.2$, $v = 0.02$, (**d**) $M = 5$, $u = 0.04$, $v = 0.1$. Results are averaged over $T = 10^8$ time steps.

If, on the other hand, anti-coordination payoffs are negative ($b < 0$), then increases in μ, ν, and M all increase heterophilic interactions and therefore decrease their payoffs, making it easier for homophily to evolve. In fact, we retain the same condition – homophily evolves provided that $a > Kb$.

If the payoff values a, b have different signs such that $ab < 0$, natural selection favors the evolution of homophily if $a > 0 > b$; that is, the homophilic interactions are beneficial ($a > 0$) while the heterophilic interactions are harmful ($b < 0$). For other special cases with $ab = 0$, natural selection trivially favors the evolution of homophily if $a > b = 0$ or $b < a = 0$.

1. Critical slopes

The critical condition (59) can be seen as a critical line in the two-dimensional (b, a) plane. For a values above this critical line ($a > Kb$), natural selection favors the evolution of homophily. For a values below this line ($a < Kb$), natural selection favors the evolution of heterophily.

Notably, the condition $a = Kb$ represents a symmetry breaking point. For $a = Kb$, the preference selection function $\mathcal{D}(p)$ is symmetric with respect to $p = 1/2$ (Figs. S1g and S1e). If two preferences are at equal distance from $p = 1/2$, they are equally abundant in the population. For $a \neq Kb$, the distribution is shifted to favor one of the two extreme trait values over the other. For example, the preference $p = 1$ (perfect homophily) is more abundant than $p = 0$ (perfect heterophily) in the mutation-selection equilibrium if $a > Kb$ (Figs. S1h and S1f).

Interestingly, when the condition $a = Kb$ holds, there exist two critical preferences for homophily, $p_c = \frac{3\pm\sqrt{3}}{6}$, which do not depend on the model parameters. At this symmetry breaking point with positive interactions ($a = Kb > 0$), natural selection favors extreme preferences close to the boundary $p < \frac{3-\sqrt{3}}{6}$ and $p > \frac{3+\sqrt{3}}{6}$ (Fig. S1g). In contrast, for antagonistic interactions $a = Kb < 0$, natural selection favors intermediate preferences $\frac{3-\sqrt{3}}{6} < p < \frac{3+\sqrt{3}}{6}$ (Fig. S1e).

In the two-dimensional $(b,\ a)$ plane, there exist other critical slopes. If $C_2 = 0$, the preference selection function $\mathcal{D}(p)$ degenerates to a linear function. In this case, the critical preference for homophily is $p_c = 1/2$, and this does not depend on other model parameters (Fig. S4a). The condition $C_2 = 0$ gives the critical slope K_2:

$$K_2 = \frac{\nu(3\mu + 2\nu + 4)(1 - M)}{(\mu + 2)(3\mu + 2\nu + 3)M + \nu(3\mu + 2\nu + 4)}. \tag{61}$$

We can see that K_2 is always negative, $K_2 < 0$. If $a > K_2 b$, we have $C_2 > 0$ and therefore the parabola $\mathcal{D}(p)$ opens upward. Otherwise, if $a < K_2 b$, then $C_2 < 0$ and therefore the parabola $\mathcal{D}(p)$ opens downward. In other words, the convexity of the preference selection function $\mathcal{D}(p)$ changes from convex upward to convex downward across the critical line $a = K_2 b$.

Whether there exist two or one single critical $p_c \in (0,\ 1)$ is determined by two other critical slopes, denoted by K_3 and K_4, respectively. By setting $\mathcal{D}(0) = C_0 = 0$ and $\mathcal{D}(1) = C_2 + C_1 + C_0 = 0$, respectively, we can obtain the explicit expressions of K_3 and K_4. Moreover, we can show that $K_3 > K > K_4 > 0$. There exist two critical $p_c \in (0,\ 1)$ if $K_4|b| < |a| < K_3|b|$ (Fig. S4b). Otherwise, there is only one single critical $p_c \in (0,\ 1)$.

Whether the vertex (tip) of the parabola $\mathcal{D}(p)$ is located within the region $[0,\ 1]$ is determined by another two critical slopes, denoted by K_5 and K_6, respectively. The axis of symmetry is given

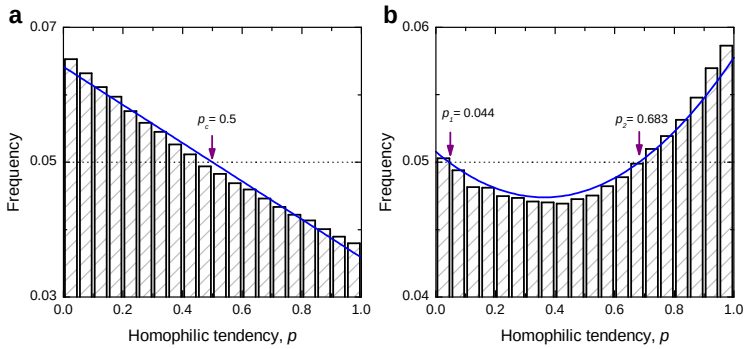

FIG. S4: Stationary distributions of preferences for (**a**) $a = -14/41$ and (**b**) $a = 700/901$. These two plots of simulation results correspond to the theoretical predictions in fig S1b and S1i, respectively. Parameters: $N = 30$, $M = 3, \beta = 0.005$, $u = 0.04$, $v = 0.06$, $b = 1$. Results are averaged over $T = 10^9$ time steps.

by $p = -\frac{C_1}{2C_2}$. Therefore, K_5 and K_6 can be obtained, respectively, by setting $-\frac{C_1}{2C_2} = 0$ and $-\frac{C_1}{2C_2} = 1$. Furthermore, we can show that $K_5 > K_3 > K > K_4 > K_6 > 0$. The preference selection function $\mathcal{D}(p)$ has one maximum or minimum point within $[0,\ 1]$ if $K_6|b| \leq |a| \leq K_5|b|$. Otherwise, $\mathcal{D}(p)$ is monotonic with p within $[0,\ 1]$.

In the limit of low preference mutation, $\mu \to 0$, these critical slopes have simple forms and can be given as follows:

$$K = \frac{(M-1)v(2+v)}{v(2+v)+M(3+2v)} \tag{62}$$

$$K_2 = \frac{(1-M)v(4+2v)}{2M(3+2v)+v(4+2v)} \tag{63}$$

$$K_3 = \frac{2(M-1)v(4+2v)}{2M(3+2v)+v(4+2v)} \tag{64}$$

$$K_4 = \frac{(M-1)v(2+v)}{2M(3+2v)+v(4+2v)} \tag{65}$$

$$K_5 \to \infty \tag{66}$$

$$K_6 \to 0 \tag{67}$$

As shown in Fig. S5, in the limit of low preference mutation ($\mu \to 0$) and high phenotypic

mutation ($\nu \to \infty$), these critical slopes become

$$K = M - 1 \tag{68}$$

$$K_2 = -(M-1) \tag{69}$$

$$K_3 = 2(M-1) \tag{70}$$

$$K_4 = \frac{M-1}{2} \tag{71}$$

$$K_5 \to \infty \tag{72}$$

$$K_6 \to 0 \tag{73}$$

In the limit of high preference mutation, $\mu \to \infty$, these critical slopes become:

$$K_2 = 0, \tag{74}$$

$$K = K_3 = K_4 = K_5 = K_6 = \frac{\nu(M-1)}{\nu + M}. \tag{75}$$

To validate these analytical results, we show the simulation results about two additional pairs of payoff values a, b satisfying $a = K_2 b$ and $Kb < a < K_3 b$ in Fig. S4 (corresponding to Figs. S1b and S1i). We find excellent agreement between our analytical predictions and simulations.

2. *Critical preferences for homophily*

The critical preference for homophily, p_c, can be obtained by solving the roots of the quadratic function $\mathcal{D}(p) = 0$ ($a^2 + b^2 \neq 0$, and $a \neq K_2 b$):

$$p_{1,2} = \frac{-C_1 \mp \sqrt{C_1^2 - 4C_2C_0}}{2C_2}. \tag{76}$$

It is easy to prove by contradiction to show that at least one of p_1 and p_2 lies in the interval $(0,\ 1)$. We have $p_c = \{p_i \mid 0 < p_i < 1, i = 1, 2\}$. As shown in the above analysis, the condition $K_4|b| < |a| < K_3|b|$ guarantees both preferences $p_{1,2}$ fall in $(0,\ 1)$. Otherwise, only one of them lies in $(0,\ 1)$. The ranking order of $p_{1,2}$ changes across the critical line $a = K_2 b$. For $a > K_2 b$, we have $p_1 < p_2$, while for $a < K_2 b$ we have $p_1 > p_2$.

Figure S5 shows in detail how the two roots $p_{1,2}$ are located with respect to the interval $(0,\ 1)$ for different regions of the $(b,\ a)$-plane divided by the critical slopes. In the limits of μ and ν, the explicit expressions of p_c have some interesting extremes. We summarize our results as follows:

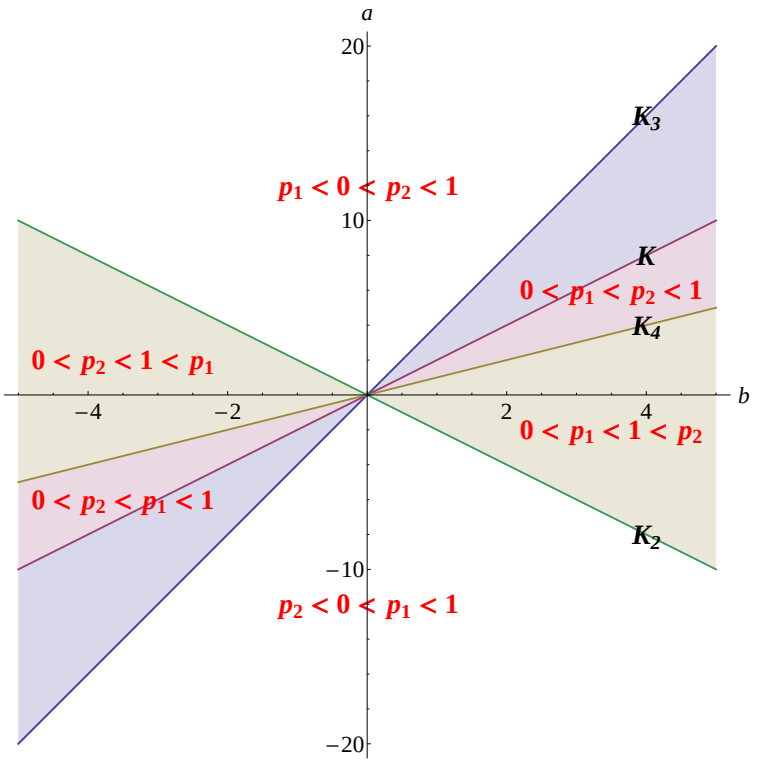

FIG. S5: Critical slopes. These critical lines $a = K_i b$ $(i = 2, 3, 4)$ divide the 2-D $(b,\ a)$-plane into 6 regions that determine the critical preferences p_c. Shown is the limiting case with $\mu \to 0$ and $\nu \to \infty$: $K = M - 1$, $K_2 = -(M-1)$, $K_3 = 2(M-1)$, $K_4 = (M-1)/2$.

- $\mu \to 0$ and $\nu \to 0$. There exists only one critical preference $p_c = \frac{1}{\sqrt{3}}$. Regardless of the payoff value b, natural selection favors $p > \frac{1}{\sqrt{3}}$ if $a > 0$; natural selection favors $p < \frac{1}{\sqrt{3}}$ if $a < 0$.

- $\mu \to 0$ and $\nu \to \infty$. These exist two possible critical preferences $p_c = \frac{3b(M-1) \pm \sqrt{3}\sqrt{a^2+b^2(M-1)^2+ab(1-M)}}{3(a+b(M-1))}$. For positive payoff b values $(b > 0)$, natural selection favors $p > \frac{3b(M-1)+\sqrt{3}\sqrt{a^2+b^2(M-1)^2+ab(1-M)}}{3(a+b(M-1))}$ if $a > 2(M-1)b$; natural selection favors $p < \frac{3b(M-1)-\sqrt{3}\sqrt{a^2+b^2(M-1)^2+ab(1-M)}}{3(a+b(M-1))}$ and $p > \frac{3b(M-1)+\sqrt{3}\sqrt{a^2+b^2(M-1)^2+ab(1-M)}}{3(a+b(M-1))}$ if $(M-1)b/2 < a < 2(M-1)b$; natural selection favors $p < \frac{3b(M-1)-\sqrt{3}\sqrt{a^2+b^2(M-1)^2+ab(1-M)}}{3(a+b(M-1))}$ if $a < (M-1)b/2$. Similar results can be obtained for negative payoff b values, $b < 0$ (Fig. S5).

- $\mu \to \infty$ and any given ν. These exists only one critical preference $p_c = 1/2$. Natural selection favors $p > 1/2$ if $a > \frac{\nu(M-1)}{\nu+M} b$; otherwise, natural selection favors $p < 1/2$ if $a < \frac{\nu(M-1)}{\nu+M} b$.

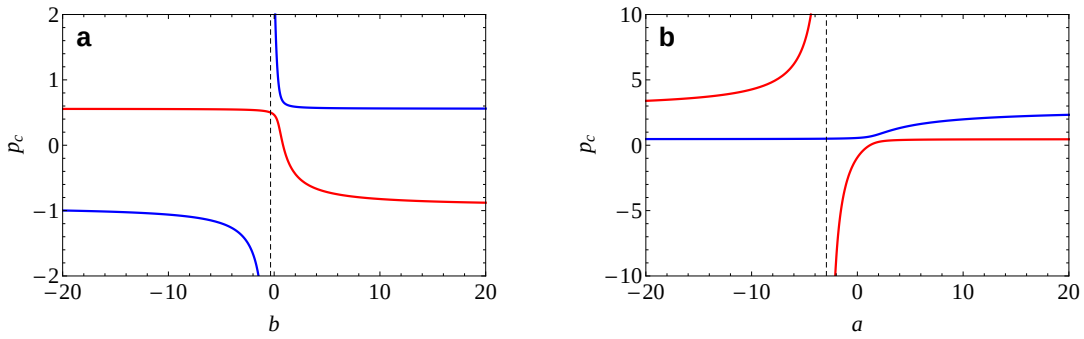

FIG. S6: Critical p_c as a function of **a** the payoff value a for fixed b and **b** the payoff value b for fixed a. Note that the singularity ($a = K_2 b$) in (**a**) $a = -14/41$ and (**b**) $b = -41/14$. Parameters: $M = 3$, $\mu = 1.2$, $\nu = 1.8$, (**a**) $b = 1$, (**b**) $a = 1$. The red and blue lines respectively denote the two zeros of $\mathcal{D}(p)$ as given in Eq. (76).

Figure S6 shows how the values of $p_{1,2}$ change with the payoff values a, b, respectively. We can see that when either of the payoff values a, b goes to $\pm\infty$, both $p_{1,2}$ approach some extreme values and only one of $p_{1,2}$'s limiting values (that is, p_c) lies in $(0,\ 1)$. Moreover, we find that the expressions of the critical preference p_c can be further simplified if the mutation rate μ approaches 0 or ∞. Interestingly, for these limits, p_c does not depend on the number of phenotypes, M, or the phenotypic mutation ν:

- Any given b and $a \to \infty$. For $\mu \to 0$, natural selection favors $p > p_c = \frac{1}{\sqrt{3}}$; for $\mu \to \infty$, natural selection favors $p > p_c = \frac{1}{2}$.
- Any given b and $a \to -\infty$. For $\mu \to 0$, natural selection favors $p < p_c = \frac{1}{\sqrt{3}}$; for $\mu \to \infty$, natural selection favors $p < p_c = \frac{1}{2}$.
- Any given a and $b \to \infty$. For $\mu \to 0$, natural selection favors $p < p_c = 1 - \frac{1}{\sqrt{3}}$; for $\mu \to \infty$, natural selection favors $p < p_c = \frac{1}{2}$.
- Any given a and $b \to -\infty$. For $\mu \to 0$, natural selection favors $p > p_c = 1 - \frac{1}{\sqrt{3}}$; for $\mu \to \infty$, natural selection favors $p > p_c = \frac{1}{2}$.

3. *Special cases*

To better understand the model, we study some of the simplest special cases ($b = 0, a \neq 0$ or $a = 0, b \neq 0$). Assuming zero payoffs for heterophilic or homophilic interactions greatly simplifies the theoretical analysis. Here we focus on the special case $a > b = 0$. Other special cases can be analyzed in the same manner.

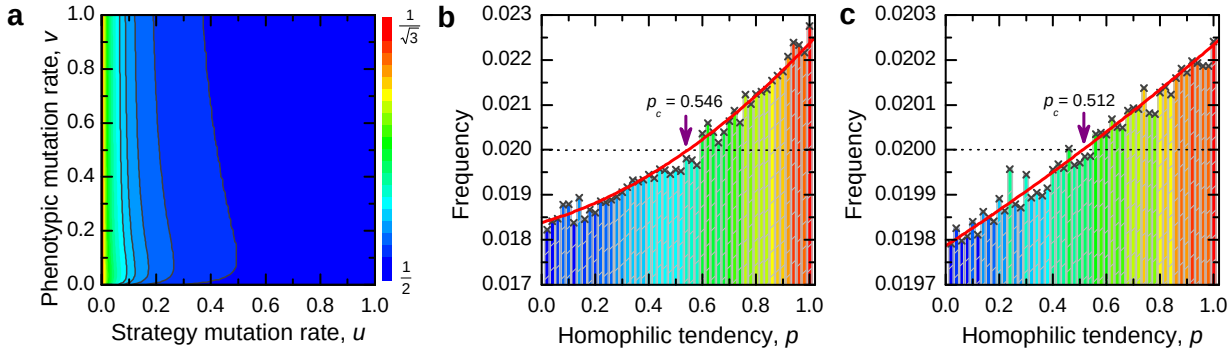

FIG. S7: Special case with $a > b = 0$. **a** shows the theoretical critical preference p_c as a function of the preference mutation rate, u, and the phenotypic mutation rate, v. **b**, **c** show the stationary distributions of preferences for low preference mutation $u = 0.02$ and high preference mutation $u = 0.2$, respectively. The solid red lines in (**b**) and (**c**) are the theoretical distributions. Parameters: $N = 100$, $M = 2$, $\beta = 0.002$, $v = 0.06$, $a = 0.1$. Results are averaged over $T = 2 \times 10^9$ time steps.

Fig. S7 shows that natural selection favors the evolution of homophily. The preference selection criterion function $\mathcal{D}(p)$ monotonically increases with p within the interval $[0,\ 1]$. Thus natural selection always favors perfect homophily ($p = 1$) most and perfect heterophily ($p = 0$) least. There exists only one critical preference $p_c \in (0,\ 1)$, such that all preferences $p > p_c$ are selected. The specific value of p_c depends on the model parameters: M, μ and ν. As shown in Fig. S7a, p_c decreases with the preference mutation μ while these exists an optimal phenotypic mutation ν that maximizes p_c (holding other parameters fixed). In the limit of low mutation $\mu \to 0$, we find a surprisingly simple expression of p_c, which does not depend on the number of phenotypic variations M or the phenotypic mutation rate ν:

$$\text{Preference } p \text{ is favored by natural selection if and only if } p > \frac{1}{\sqrt{3}}. \tag{77}$$

As the preference mutation rate μ increases, this critical value p_c approaches $1/2$. These analytical results are confirmed by agent-based simulations (Figs. S7b and S7c).

In the following, we briefly summarize the results for other special cases:

- $a < b = 0$. Natural selection favors the evolution of heterophily, that is, $\langle p \rangle < 1/2$. For $\mu \to 0$, natural selection favors preferences $p < p_c = \frac{1}{\sqrt{3}}$. For $\mu \to \infty$, natural selection favors preferences $p < p_c = \frac{1}{2}$.

- $b > a = 0$. Natural selection favors the evolution of heterophily, that is, $\langle p \rangle < 1/2$. For $\mu \to 0$, natural selection favors preferences $p < p_c = 1 - \frac{1}{\sqrt{3}}$. For $\mu \to \infty$, natural selection favors preferences $p < p_c = \frac{1}{2}$.

- $b < a = 0$. Natural selection favors the evolution of homophily, that is, $\langle p \rangle > 1/2$. For $\mu \to 0$, natural selection favors preferences $p > p_c = 1 - \frac{1}{\sqrt{3}}$. For $\mu \to \infty$, natural selection favors preferences $p > p_c = \frac{1}{2}$.

We summarize analytical results for the critical preference p_c at different limiting cases in Table I.

F. Biased matching process

So far, we have considered the matching process that governs whether two individuals meet is proportional to the respective abundance of every phenotype. It is, however, not implausible that individuals are more likely to meet similar others. To account for such meeting bias, let us introduce a new parameter $\phi \in [0,\ 1]$: with probability ϕ individuals are matched with a same-phenotype individual; otherwise with probability, $1 - \phi$, individuals are uniformly and randomly paired up. Hence, the parameter ϕ characterizes the assortativity in the matching process. For $\phi = 0$, we have the unbiased matching scenario as studied before. For nonzero ϕ values, the total number of matches between individuals with preference i and j, q_{ij}, now becomes

$$q_{ij} = \delta_{ij} q_{ij}^{\mathrm{S}} + (1 - \delta_{ij}) q_{ij}^{\mathrm{D}}, \tag{78}$$

where we have

$$q_{ij}^{\mathrm{S}} \ = \ N^2 \sum_{l=1}^{M} x_i^l x_j^l, \tag{79}$$

$$q_{ij}^{\mathrm{D}} \ = \ (1 - \phi) N^2 \sum_{l=1}^{M} \sum_{r=1, r \neq l}^{M} x_i^l x_j^r. \tag{80}$$

TABLE I: The critical preference for homophily, p_c, at limiting cases. An illustrative example ($\mu \to 0$, $\nu \to \infty$) can be found in Fig. S5.

μ	ν	p_c	a, b	Natural selection favors
$\mu \to 0$	$\nu \to 0$	$p_c = 1/\sqrt{3}$	$a > 0$	$p > 1/\sqrt{3}$
$\mu \to 0$	$\nu \to 0$	$p_c = 1/\sqrt{3}$	$a < 0$	$p < 1/\sqrt{3}$
$\mu \to 0$	$\nu \to \infty$	[a] $p_c = p_1$	$b > 0, a < (M-1)b/2$	$p < p_1$
$\mu \to 0$	$\nu \to \infty$	[a] $p_c = \{p_1, p_2\}$	$b > 0, (M-1)b/2 < a < 2(M-1)b$	$p < p_1$ and $p > p_2$
$\mu \to 0$	$\nu \to \infty$	[a] $p_c = p_2$	$b > 0, a > 2(M-1)b$	$p > p_2$
$\mu \to 0$	$\nu \to \infty$	[a] $p_c = p_1$	$b < 0, a < 2(M-1)b$	$p < p_1$
$\mu \to 0$	$\nu \to \infty$	[a] $p_c = \{p_1, p_2\}$	$b < 0, 2(M-1)b < a < (M-1)b/2$	$p_2 < p < p_1$
$\mu \to 0$	$\nu \to \infty$	[a] $p_c = p_2$	$b < 0, a > (M-1)b/2$	$p > p_2$
$\mu \to \infty$	$\nu > 0$	$p_c = 1/2$	$a > \frac{\nu(M-1)}{\nu+M} b$	$p > 1/2$
$\mu \to \infty$	$\nu > 0$	$p_c = 1/2$	$a < \frac{\nu(M-1)}{\nu+M} b$	$p < 1/2$
$\mu \to 0$	$\nu > 0$	$p_c = 1/\sqrt{3}$	$a > 0, b = 0$	$p > 1/\sqrt{3}$
$\mu \to 0$	$\nu > 0$	$p_c = 1/\sqrt{3}$	$a < 0, b = 0$	$p < 1/\sqrt{3}$
$\mu \to \infty$	$\nu > 0$	$p_c = 1/2$	$a > 0, b = 0$	$p > 1/2$
$\mu \to \infty$	$\nu > 0$	$p_c = 1/2$	$a < 0, b = 0$	$p < 1/2$
$\mu \to 0$	$\nu > 0$	$p_c = 1 - 1/\sqrt{3}$	$a = 0, b > 0$	$p < 1 - 1/\sqrt{3}$
$\mu \to 0$	$\nu > 0$	$p_c = 1 - 1/\sqrt{3}$	$a = 0, b < 0$	$p > 1 - 1/\sqrt{3}$
$\mu \to \infty$	$\nu > 0$	$p_c = 1/2$	$a = 0, b > 0$	$p < 1/2$
$\mu \to \infty$	$\nu > 0$	$p_c = 1/2$	$a = 0, b < 0$	$p > 1/2$

[a] $p_1 = \frac{3b(M-1) - \sqrt{3}\sqrt{a^2+b^2(M-1)^2+ab(1-M)}}{3(a+b(M-1))}$, $p_2 = \frac{3b(M-1) + \sqrt{3}\sqrt{a^2+b^2(M-1)^2+ab(1-M)}}{3(a+b(M-1))}$.

Using the same algebra as we did for the simple version of the model, the condition for preference k to be selected becomes

$$\left[\lambda_1^{\mathrm{S}}(a_{kk}^{\mathrm{S}} - \overline{a_{**}^{\mathrm{S}}}) + \lambda_2^{\mathrm{S}}(\overline{a_{k*}^{\mathrm{S}}} - \overline{a_{*k}^{\mathrm{S}}}) + \lambda_3^{\mathrm{S}}(\overline{a_{k*}^{\mathrm{S}}} - \overline{a^{\mathrm{S}}})\right] \tag{81}$$
$$+(1-\phi)\left[\lambda_1^{\mathrm{D}}(a_{kk}^{\mathrm{D}} - \overline{a_{**}^{\mathrm{D}}}) + \lambda_2^{\mathrm{D}}(\overline{a_{k*}^{\mathrm{D}}} - \overline{a_{*k}^{\mathrm{D}}}) + \lambda_3^{\mathrm{D}}(\overline{a_{k*}^{\mathrm{D}}} - \overline{a^{\mathrm{D}}})\right] > 0,$$

where the structural coefficients λ_i^{S}'s and λ_i^{D}'s are exactly the same as before. In effect, the introduction of matching bias discounts the payoffs obtained from heterophilic interactions. As a

result, the condition for the population to evolve homophilic preferences is

$$a > (1 - \phi)Kb, \tag{82}$$

where K is the same as above. Notice that the right hand side is decreasing in ϕ. This means that there is an even greater range of payoffs to coordination that allow homophily to evolve.

III. LOCAL MUTATION

So far we have assumed a global mutation mechanism, in which a mutant offspring adopts a preference randomly and uniformly on the interval $[0,\ 1]$. An alternative is to consider a local mutation mechanism, in which a mutant offspring's preference is drawn from a Gaussian distribution, with the parental preference as the mean and with a small standard deviation, σ. We study this alternative model using adaptive dynamics to assess the most likely evolutionary path of adaptation. For this purpose, we only consider the limit of low preference mutation ($u \to 0$).

For rare preference mutations, the fate of a mutant preference is determined—either it takes over the population or goes extinct before the next mutant arises. Therefore, the evolutionary competition of multiple preferences reduces to pairwise invasion dynamics between only two preferences. While standard adaptive dynamics consider the invasion fitness of a mutant against the resident population [6, 7], here we generalize it to account for the stochasticity in our model [8].

Whether a mutant, y, fares better than the resident preference, x, is described by the selection condition (22). For $\mu \to 0$, the structural coefficient associated with the term $(\overline{a_{k*}} - \overline{a})$ approaches 0. The sum of the remaining terms, denoted by $\mathcal{D}(y,\ x)$, are equivalent to the comparison of fixation probabilities. The mutant preference x is favored over the resident preference y by natural selection if $\mathcal{D}(y,\ x)$ is positive, and otherwise disfavored by natural selection if $\mathcal{D}(y,\ x)$ is negative. If $\mathcal{D}(y,\ x) = 0$, the two preferences are neutral.

Based on the previous calculations, we can see

$$\mathcal{D}(y, x) = (\nu + 1)(\nu + 3)(y - x)\left[(a + bK_0)(x + y) - 2bK_0\right], \tag{83}$$

where K_0 is $\lim_{\mu\to 0} K = \nu(\nu + 2)(M - 1)/\left[\nu(\nu + 2) + (2\nu + 3)M\right]$, and K is as given in Eq. (60).

Using Eq. (83), we show the pairwise invasion plots in Fig. S8. Mutants are favored over the resident in the red regions, that is, $\mathcal{D}(y, x) > 0$, and disfavored in the blue regions where $\mathcal{D}(y, x) < 0$. The (x, y)-plane is divided into four parts, by the two lines $y = x$ and $x + y = 2bK_0/(a + bK_0)$.

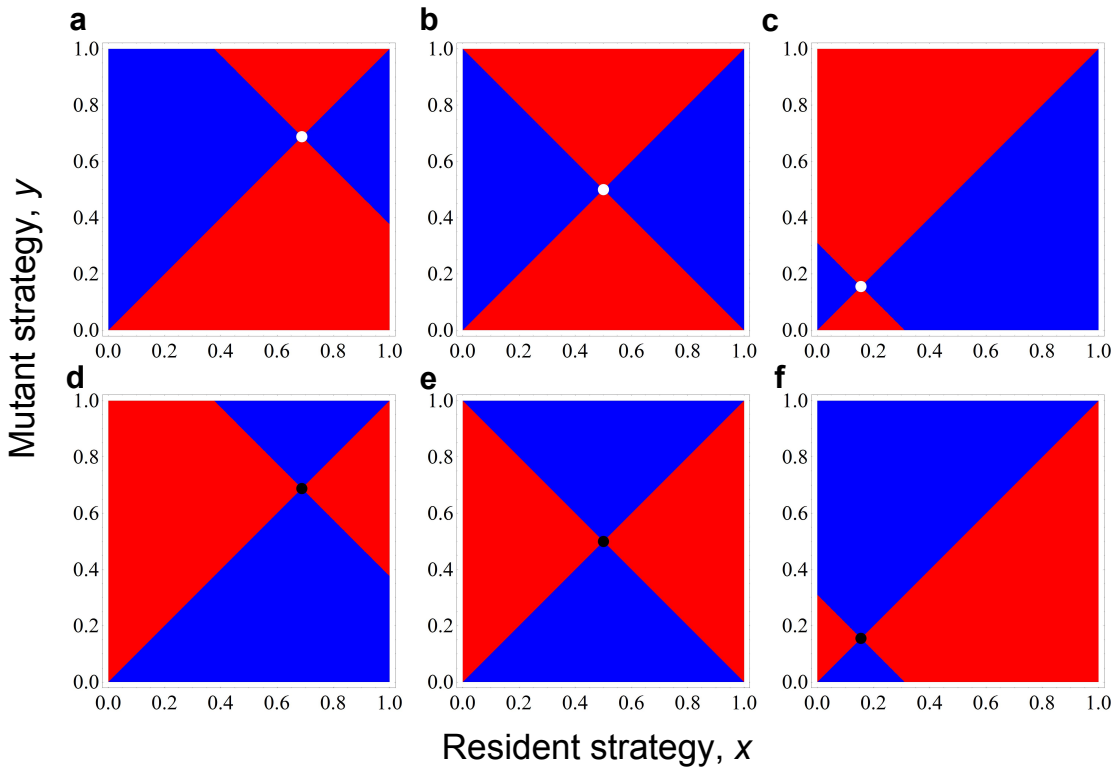

FIG. S8: Pairwise invasibility plots. Red denotes that mutant preference wins while blue denotes that resident preference wins. The empty circles represent unstable interior equilibrium ('repeller'), and the filled circles denote stable interior equilibrium ('attractor'). Parameters: $M = 3$, $\nu = 2$, (**a**–**c**) $b = 1$, $a = 1/4, 16/29, 3$, (**d**–**f**) $b = -1$, $a = -1/4, -16/29, -3$.

Their intersection point, $p^* = bK_0/(a + bK_0)$, gives the possible interior equilibrium

$$p^* = \frac{bK_0}{a + bK_0}. \tag{84}$$

Let us further check the existence of $p^* \in (0,\ 1)$ and its stability. To this end, let us assume that the mutant preference, y, is drawn from the infinitesimal neighborhood of the resident preference, x. In this limit, the gradient of the function $\mathcal{D}(y,\ x)$ with respect to y, evaluated at x ($\frac{\partial \mathcal{D}(y,\ x)}{\partial y}\,|_{y=x}$), gives the most likely direction of evolutionary adaptation. Drawing on classic adaptive dynamics, we can use the following deterministic differential equation to describe the evolution of the homophilic trait value

$$\begin{aligned} \dot{x} &= \frac{\partial \mathcal{D}(y,\ x)}{\partial y}\,|_{y=x}\,, \\ &= 2(\nu + 1)(\nu + 3)[ax + bK_0(x - 1)]. \end{aligned} \tag{85}$$

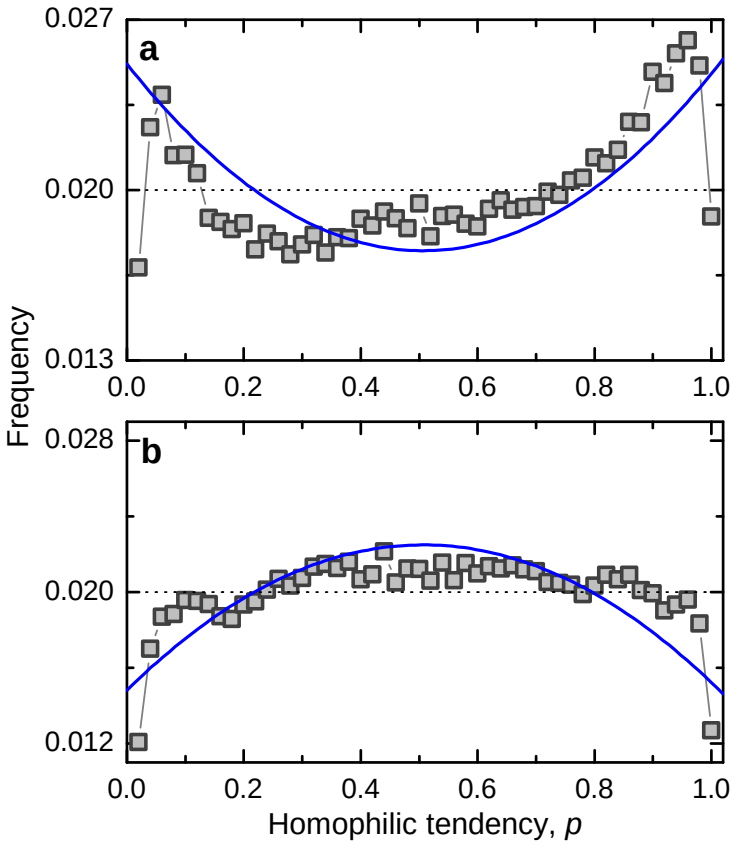

FIG. S9: Simulation results validate the adaptive dynamics analysis. Shown are the distributions of preferences under local mutation. For positive payoff values, adaptive dynamics analysis predicts bistability (unstable interior equilibrium), and the resulting distribution is U-shaped. In contrast, negative payoff values a, b, lead to coexistence (stable interior equilibrium) and the resulting distribution is bell-shaped. Note that due to boundary effects, the frequency distribution precipitously drops around the two extremes: $p = 0$ and $p = 1$. The solid blue lines are the theoretical distributions obtained in the limit of rare mutation ($\mu \to 0$). Parameters: $N = 30$, $M = 3$, $\beta = 0.005$, $u = 0.005$, $\sigma^2 = 0.0005$, $v = 0.06$, (**a**) $a = 0.5$, $b = 1$, (**b**) $a = -0.5$, $b = -1$. Results are averaged over $T = 2 \times 10^9$ time steps.

In this way, we can also derive the interior equilibrium p^* as shown in Eq. (84). Its existence requires $ab > 0$, namely, the payoff values a, b must have the same sign. The stability of the interior equilibrium is determined by the sign of $\frac{\partial^2 \mathcal{D}(y,\, x)}{\partial y^2}|_{y=x} = a + K_0 b$. Hence, we have the following possible cases:

- $a = 0$ and $b = 0$. The evolutionary dynamics are completely neutral.
- $b = 0$. There are no interior equilibria. If $a > 0$, $p^* = 0$ is unstable and $p^* = 1$ is stable. If

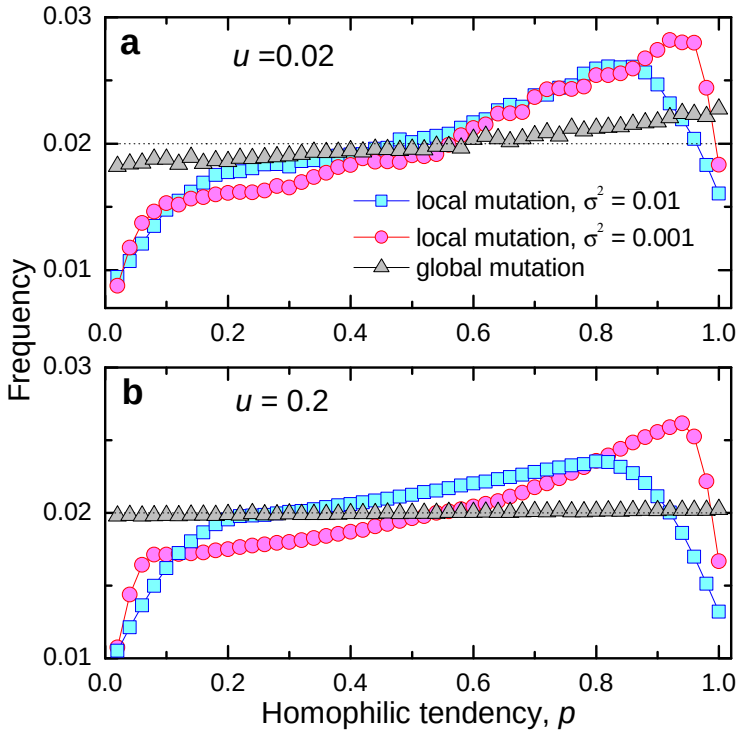

FIG. S10: Stationary distributions of preferences under local mutation. Parameters: $N = 100$, $M = 2$, $\beta = 0.002$, $v = 0.06$, $a = 0.1$, $b = 0$, (**a**) $u = 0.02$, (**b**) $u = 0.2$. Results are averaged over $T = 2 \times 10^9$ time steps.

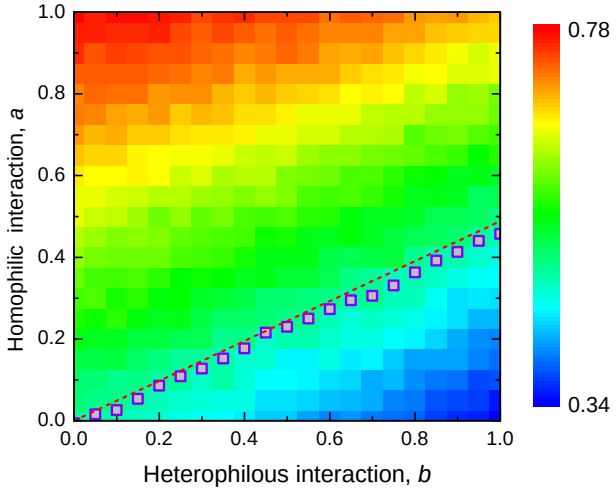

FIG. S11: Population average $\langle p \rangle$ as a function of (b, a) under local mutation. Parameters: $N = 50$, $\beta = 5 \times 10^{-3}$, $M = 2$, $u = 0.04$, $v = 0.06$, $\sigma^2 = 0.01$. Results are averaged over $T = 10^8$ time steps.

$a < 0$, $p^* = 0$ is stable and $p^* = 1$ is unstable.

- $a = 0$. There are no interior equilibria. If $b > 0$, $p^* = 0$ is stable and $p^* = 1$ is unstable. If $b < 0$, $p^* = 0$ is unstable and $p^* = 1$ is stable.

- $ab > 0$. There exists an interior equilibrium. If $a, b > 0$, the interior equilibrium $p^* = \frac{bK_0}{a+bK_0}$ is unstable and the two boundary points $p^* = 0, 1$ are both stable. If $a, b < 0$, the interior equilibrium $p^* = \frac{bK_0}{a+bK_0}$ is stable and the two boundary points $p^* = 0, 1$ are unstable.

- $ab < 0$. There are no interior equilibria. If $a > 0 > b$, $p^* = 0$ is unstable and $p^* = 1$ is stable. If $a < 0 < b$, $p^* = 0$ is stable and $p^* = 1$ is unstable.

Notably, the evolutionary adaptation of homophilic preferences shows bistability. Specifically, the interior equilibrium $p^* = bK_0/(a + bK_0)$ is a repeller, if both homophilic and heterophilic associations are beneficial ($a, b > 0$) (see Figs. S8a-c). The direction of evolution depends on the initial conditions. If the initial homophilic preference of the population $p < p^*$, the population eventually becomes completely heterophilic ($p = 0$). Otherwise, if the initial homophilic preference of the population $p > p^*$, the population becomes completely heterophilic ($p = 1$). In contrast, the interior equilibrium $p^* = bK_0/(a + bK_0)$ is an attractor (*i.e.*, coexistence) if both homophilic and heterophilous associations are antagonistic ($a, b < 0$) (Figs. S8d-f). In this situation, natural selection most favors the intermediate homophilic preference p^*.

There is an interesting correspondence between these results for local mutation and those obtained previously for global mutation. In the limit of low preference mutation ($\mu \to 0$) and for payoff values satisfying $ab > 0$, we can show that the axis of symmetry of the parabola $\mathcal{D}(p)$ is always located within the region $(0, 1)$, and that its position is exactly the same as given in Eq. (84), $p^* = bK_0/(a + bK_0)$. This result suggests that the maximum of the parabola (if $a, b > 0$) or its minimum (if $a, b < 0$) is reached at the interior point, p^*, which is the singular point shown in the adaptive dynamics analysis.

For $a, b > 0$, natural selection favors the evolution of homophily, if and only if $p = 1$ has a larger attraction basin than $p = 0$, that is, $p^* < 1/2$. For $a, b < 0$, natural selection favors the evolution of homophily, if and only if the attractor $p^* > 1/2$. Both conditions lead to $a > K_0 b$, which is identical to the evolutionary condition derived under global mutation. Further, it is not difficult to show that for $ab \leq 0$ the evolution of homophily under local mutation still requires $a > K_0 b$.

We performed some simulations to corroborate these theoretical insights. Adaptive dynamics analysis for local mutation shows that positive payoff values $a, b > 0$ lead to bistability (an unstable interior equilibrium), and thus the resulting frequency distribution of preferences is U-shaped (Fig. S9a). Negative payoff values $a, b < 0$ result in coexistence (a stable interior equilibrium) and therefore the resulting frequency distribution of preferences is bell-shaped (Fig. S9b). We note that in the limit of low mutation, the theoretical distribution obtained for global mutation is still a fairly good approximation for local mutation (Fig. S9).

Compared with Fig. S7, local mutation increases the disparity in frequency between preferences, leading to more skewed distributions (Fig. S10). We can see that decreasing the standard deviation in the mutation kernel leads to even more skewed distributions. The frequencies of preferences in the vicinity of $p = 0$ and $p = 1$ dramatically plunge due to boundary effects. Figure S11 shows the population average $\langle p \rangle$ as a function of the payoff values (b, a) under local mutation. Our theoretical analysis above suggests that the evolution of homophily under local mutation requires the same critical condition $a > Kb$ as obtained for the global mutation case. As shown in Fig. S11, the condition $a > Kb$ still works relatively well for local mutation.

IV. STRONG SELECTION

Our analytical theory works very well for weak selection, as evidenced by the close agreement between theoretical predictions and agent-based simulation results. Increasing the strength of natural selection, β, magnifies the fitness difference between traits and, as a result, the evolutionary dynamics become increasingly deterministic.

Figure S12 shows how increasing selection strength acts on the resulting stationary distributions of homophilic preferences. Here we simulate the simplest special case $a > b = 0$ as in Fig. S7. We observe that increasing the selection pressure makes the frequency distributions more skewed and also shifts the critical preference p_c towards the boundary. Increasing the mutation rate, u, also widens the region of trait values favored by natural selection (cf. Figs. S7 and S12). Under strong selection, only the extreme value trait, *i.e.*, either $p = 0$ and $p = 1$, is selected, depending on the model parameters specified. These simulations confirm that the above results for weak selection qualitatively hold for non-weak selection.

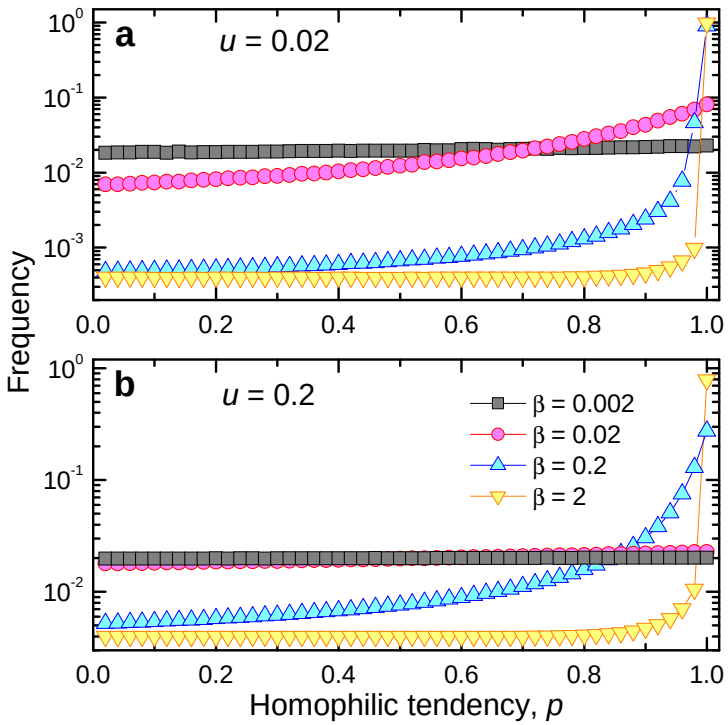

FIG. S12: Effects of selection strength on the evolution of homophily. Parameters: $N = 100$, $M = 2$, $v = 0.06$, $a = 0.1$, $b = 0$, (**a**) $u = 0.02$, (**b**) $u = 0.2$. Results are averaged over $T = 2 \times 10^9$ time steps.

V. EXTENSION TO FULL STRATEGY SPACE

In the basic model introduced above, an individual's homophilic and heterophilic preferences, denoted by p and q respectively, are in exact tradeoff (*i.e.*, $p + q = 1$); increasing the homophilic preference lowers the heterophilic preference and vice versa. In this way, the level of sociality (namely, the degree of unilateral willingness of engaging in all kinds of potential social interactions, which is well quantified by the unweighted sum $p + q$) is kept the same for every individual. For any similar or dissimilar encounter, individuals have to make their own unilateral choices whether or not to interact, depending on their homophilic preferences.

This constraint can be relaxed by considering the full strategy space, that is, the set of all $(p,\ q)$ points in the unit square. To be concrete, each individual's preference now is described by a pair of variables, $(p,\ q)$, within the unit square, $[0,\ 1]^2$. The homophilic preference p (the heterophilic preference q, resp.) denotes the probability that an individual agrees to establish a pairwise social relationship (or interaction) when meeting another individual bearing the same phenotype (carrying a different phenotype, resp.). The rest of notations are the same as in the

basic model. The number of phenotypic variations is M. Each individual receives a payoff a (b, resp.) for every successfully established assortative (disassortative, resp.) relationship based on bilateral agreement.

Specifically, let us first consider the four binary preferences located in the corners. That is, $(0,\ 0)$, 'solitary' which has no social interactions at all; $(1,\ 0)$, 'coordination' which forms assortative social interactions exclusively with these carrying the same phenotype; $(0,\ 1)$, 'anti-coordination' which forms disassortative social interactions exclusively with these carrying different phenotypes; $(1,\ 1)$, 'unbiased' which unselectively forms social interactions with everyone else, irrespective of the phenotypic similarity. In what follows, we discuss the ranking of the relative abundance density of these four binary preferences under different conditions, and also derive conditions for the evolution of homophily.

We adopt the same evolutionary updating rule as used in the basic model. An individual reproduces proportional to its fitness, $f_i = \exp(\beta\pi_i)$, where β is the intensity of selection and π_i is the accumulated payoff from social interactions. Reproduction is subject to mutation. The strategy mutation occurs with probability u, in which the offspring chooses a strategy which is randomly and uniformly drawn from the unit square. The phenotypic mutation happens with probability v, in which the offspring randomly chooses one out of the total M phenotypes.

Using the same analytical method presented above, we can obtain the relative abundance density of the preference $(p,\ q)$ in the mutation-selection equilibrium as follows:

$$\begin{aligned}
\mathcal{D}(p,\ q) \ =\ & b(M-1)\left(-\frac{1}{3}+q^2\right)\nu(1+\nu)(3+\mu+\nu)(4+3\mu+2\nu) \\
+\ & a\left(p^2-\frac{1}{3}\right)(1+\nu)(3+\mu+\nu)(M(2+\mu)(3+3\mu+2\nu)+\nu(4+3\mu+2\nu)) \\
+\ & \frac{1}{4}b(M-1)(-1+2q)\mu\nu\left(34+3\mu^3+40\nu+2\nu^2(8+\nu)+\mu(3+\nu)(16+7\nu)+\mu^2(21+8\nu)\right) \\
+\ & \frac{1}{4}a(2p-1)\mu\left(M(2+\mu)\left(9+3\mu(4+\mu)+7\nu+5\mu\nu+2\nu^2\right)+\right. \\
& \left.+\nu\left(34+3\mu^3+40\nu+2\nu^2(8+\nu)+\mu(3+\nu)(16+7\nu)+\mu^2(21+8\nu)\right)\right). \qquad (86)
\end{aligned}$$

Nature selection favors $(p,\ q)$ if and only if $\mathcal{D}(p,\ q) > 0$.

Up to a same constant factor, the population average $\langle p\rangle$ and $\langle q\rangle$ is given by, respectively,

$$\langle p\rangle \ \propto\ \frac{1}{24}a(2+\mu)(1+\mu+\nu)(6+3\mu+2\nu)(\nu(2+\mu+\nu)+M(3+\mu+2\nu)), \qquad (87)$$

$$\langle q\rangle \ \propto\ \frac{1}{24}b(M-1)(2+\mu)\nu(1+\mu+\nu)(2+\mu+\nu)(6+3\mu+2\nu). \qquad (88)$$

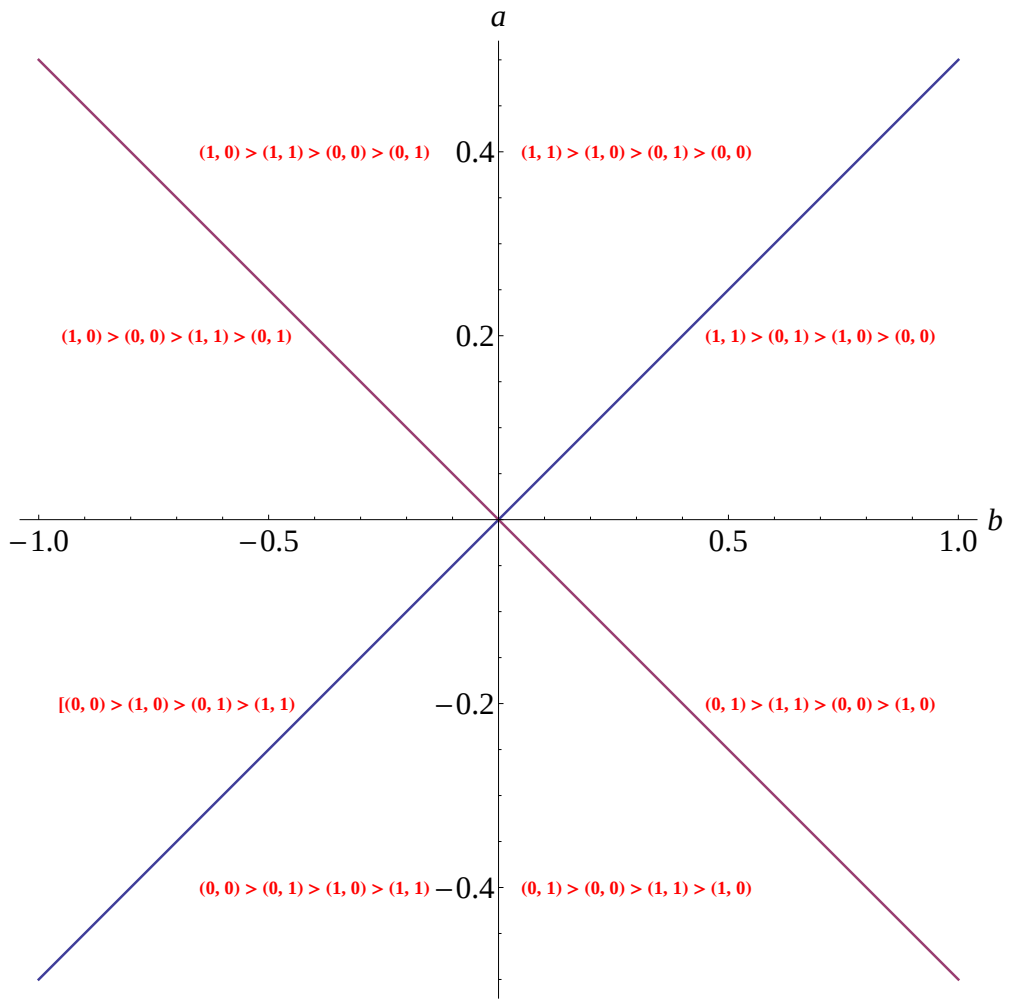

FIG. S13: Abundance ranking of the four binary preferences under the mutation-selection equilibrium. The b-axis, the a-axis and the two lines $a = Kb$ and $a = -Kb$ divide the $(b,\ a)$-plane into eight regions. The abundance ranking order of the four preferences, $(1,\ 1)$, $(1,\ 0)$, $(0,\ 1)$, $(0,\ 0)$, is given for each region. Parameters: $M = 2$, $\mu = 2$, and $\nu = 1$.

The condition for that the average homophilic preference $\langle p \rangle$ is greater than the average heterophilic preference $\langle q \rangle$ is:

$$a > Kb, \tag{89}$$

where the term K is the *same* as given in the basic model,

$$K = \frac{\nu(\mu + \nu + 2)(M - 1)}{\nu(\mu + \nu + 2) + (\mu + 2\nu + 3)M}. \tag{90}$$

Apart from population averages, we further ask the following question: what is the probability to find a homophilic individual, who is randomly drawn from the population equilibrium distri-

bution? The relative probability, $\Pr\{p > q\}$, is given by integrating the relative abundance density function $\mathcal{D}(p,\ q)$ over the region $0 \le q < p \le 1$:

$$\Pr\{p > q\} = \int_{p=0}^{p=1} \int_{q=0}^{q=p} \mathcal{D}(p,\ q) dq dp. \tag{91}$$

We obtain, up to a positive factor,

$$\Pr\{p > q\} \propto a\,(\nu(2 + \mu + \nu) + M(3 + \mu + 2\nu)) - b(M - 1)\nu(2 + \mu + \nu). \tag{92}$$

Therefore, it is more likely to find a homophilic individual than a heterophilic individual if $\Pr\{p > q\} > 0$. This leads to exactly the *same* condition $a > Kb$ as derived before.

Furthermore, we find that for fixed q, $\mathcal{D}(p,\ q)$ is an increasing function of p if and only if $a > 0$. For fixed p, $\mathcal{D}(p,\ q)$ is an increasing function of q if and only if $b > 0$. We now can rank the equilibrium densities of the four binary preferences mentioned above. Note that there exist two critical lines, given by $a = Kb$ and $a = -Kb$, which divide the four quadrants into eight regions. For (b, a) values within each region, the corresponding abundance ranking order of these binary preferences is given in Fig. S13.

VI. MULTIPLE SETS OF PHENOTYPES

So far, we have considered only one dimension of phenotypes in the analysis. Our theoretical framework presented here can be readily extended to study multiple sets of phenotypes. For simplicity, let us specifically consider two sets of phenotypes—cases with more than two sets of phenotypes can be analyzed analogously.

Let us assume the number of different phenotypes in set 1 is M_1 and in set 2 is M_2. Payoff values may vary for different sets of phenotypes, as each set serves a different function in social interactions. Let us assume that for interactions in set 1, the payoff to homophilic interaction is a and the payoff to heterophilic interaction is b, whereas in set 2 the payoff to homophilic interaction is c and the payoff to heterophilic interaction is d. Individuals may have different homophilic preferences across different sets of phenotypes. Let us assume an individual's homophilic preference toward the phenotypic set 1 is p, and that toward phenotypic set 2 it is q.

As each set of phenotypes is different in functionality in social interactions, individuals meet and interact with partners across different phenotypic sets. Specifically, two individuals, i and j, choose to interact with probability $p_i p_j$ (with probability $q_i q_j$) if they are similar in their phenotypes in set 1 (set 2, respectively). They choose to interact with probability $(1 - p_i)(1 - p_j)$ (with

probability $(1-q_i)(1-q_j)$) if they are different in their phenotypes in set 1 (set 2, respectively). This mutual partner selection process leads to a two-layer network of social interactions. The interaction structure in each layer is tuned by individuals' homophilic preferences p and q, respectively. However, these two layers of networks are interdependent since payoffs accrued from each layer jointly contribute to an individual's overall fitness.

When two randomly chosen individuals, i and j, are matched to form social connections, four possibilities arise, depending on their phenotypic similarity: (1) they have the same phenotypes in both set 1 and set 2; (2) they share the same phenotype in set 1 but carry different phenotypes in set 2; (3) they share the same phenotype in set 2 but carry different phenotypes in set 1; and (4) they carry different phenotypes in both set 1 and set 2. The expected payoff to individual i is $\pi_{ij}^1 = ap_ip_j+cq_iq_j$ in case (1), $\pi_{ij}^2 = ap_ip_j+d(1-q_i)(1-q_j)$ in case (2), $\pi_{ij}^3 = b(1-p_i)(1-p_j)+cq_iq_j$ in case (3), $\pi_{ij}^4 = b(1-p_i)(1-p_j)+d(1-q_i)(1-q_j)$ in case (4).

Here we consider an unbiased matching process as before. Denote by x_i^{lr} the fraction of individuals in the population with preference (p_i, q_i) and phenotypes (l, r). The total number of encounters between individuals with (p_i, q_i) and with (p_j, q_j) can be defined for four possible cases as above; that is, $q_{ij}^1 = N^2\sum_{l=1}^{M_1}\sum_{r=1}^{M_2}x_i^{lr}x_j^{lr}$ in case (1), $q_{ij}^2 = N^2\sum_{l=1}^{M_1}\sum_{r=1}^{M_2}\sum_{h=1,h\neq r}^{M_2}x_i^{lr}x_j^{lh}$ in case (2), $q_{ij}^3 = N^2\sum_{l=1}^{M_2}\sum_{r=1}^{M_1}\sum_{h=1,h\neq r}^{M_1}x_i^{rl}x_j^{hl}$ in case (3), and $q_{ij}^4 = N^2\sum_{t=1}^{M_1}\sum_{l=1,l\neq t}^{M_1}\sum_{r=1}^{M_2}\sum_{h=1,h\neq r}^{M_2}x_i^{tr}x_j^{lh}$ in case (4). Then the expected total payoff to individual i with preference (p_i, q_i) is $\pi_i = \sum_m\sum_j\pi_{ij}^m q_{ij}^m/(N\sum_l\sum_r x_i^{lr})$. Using our analytical method introduced before, we can obtain the following condition for natural selection to favor the preference (p_k, q_k):

$$\sum_{m=1}^{4}\lambda_1^m(\pi_{kk}^m-\overline{\pi_{**}^m})+\lambda_2^m(\overline{\pi_{k*}^m}-\overline{\pi_{*k}^m})+\lambda_3^m(\overline{\pi_{k*}^m}-\overline{\pi^m})>0, \tag{93}$$

where, up to a same positive common factor, the structural coefficients λ_i^m are given below:

$$\lambda_1^1 \propto (1+\nu)(3+\mu+\nu)(M_1M_2(2+\mu)(3+3\mu+2\nu)+\nu(4+3\mu+2\nu)), \tag{94}$$

$$\lambda_2^1 \propto M_1M_2(2+\mu)\left(9+3\mu(4+\mu)+6\nu+5\mu\nu+\nu^2\right) + \nu\left(3\mu^3+2(2+\nu)(3+\nu)^2+\mu^2(21+8\nu)+\mu(49+\nu(38+7\nu))\right), \tag{95}$$

$$\lambda_3^1 \propto \mu\Big[M_1M_2(2+\mu)\left(9+3\mu(4+\mu)+7\nu+5\mu\nu+2\nu^2\right) + \nu\left(34+3\mu^3+40\nu+2\nu^2(8+\nu)+\mu(3+\nu)(16+7\nu)+\mu^2(21+8\nu)\right)\Big], \tag{96}$$

$$\lambda_1^2 \propto (M_2-1)\nu(1+\nu)(3+\mu+\nu)(4+3\mu+2\nu), \tag{97}$$

$$\lambda_2^2 \propto (M_2-1)\nu\left(3\mu^3+2(2+\nu)(3+\nu)^2+\mu^2(21+8\nu)+\mu(49+\nu(38+7\nu))\right), \tag{98}$$

$$\lambda_3^2 \propto (M_2-1)\mu\nu\left(34+3\mu^3+40\nu+2\nu^2(8+\nu)+\mu(3+\nu)(16+7\nu)+\mu^2(21+8\nu)\right), \tag{99}$$

$$\lambda_1^3 \propto (M_1-1)\nu(1+\nu)(3+\mu+\nu)(4+3\mu+2\nu), \tag{100}$$

$$\lambda_2^3 \propto (M_1-1)\nu\left(3\mu^3+2(2+\nu)(3+\nu)^2+\mu^2(21+8\nu)+\mu(49+\nu(38+7\nu))\right), \tag{101}$$

$$\lambda_3^3 \propto (M_1-1)\mu\nu\left(34+3\mu^3+40\nu+2\nu^2(8+\nu)+\mu(3+\nu)(16+7\nu)+\mu^2(21+8\nu)\right) \tag{102}$$

$$\lambda_1^4 \propto (M_1-1)(M_2-1)\nu(1+\nu)(3+\mu+\nu)(4+3\mu+2\nu), \tag{103}$$

$$\lambda_2^4 \propto (M_1-1)(M_2-1)\nu\left(3\mu^3+2(2+\nu)(3+\nu)^2+\mu^2(21+8\nu)+\mu(49+\nu(38+7\nu))\right), \tag{104}$$

$$\lambda_3^4 \propto (M_1-1)(M_2-1)\mu\nu\left(34+3\mu^3+40\nu+2\nu^2(8+\nu)+\mu(3+\nu)(16+7\nu)+ \mu^2(21+8\nu)\right). \tag{105}$$

The selection condition for a continuum of preferences in the unit $[0,\ 1]^2$ can be obtained by replacing the sums with integrals in the condition (93). We find that the population tends to evolve homophilic preferences in phenotypic set 1, *i.e.*, $\langle p\rangle > 1/2$, if the following condition holds:

$$a > K'b, \tag{106}$$

where the term K' is given by

$$K' = \frac{\nu(\mu+\nu+2)(M_1-1)}{\nu(\mu+\nu+2)+(\mu+2\nu+3)M_1}. \tag{107}$$

Meanwhile, the population tends to evolve homophilic preference in phenotypic set 2, *i.e.*, $\langle q\rangle > 1/2$, if the following condition holds:

$$c > K''d, \tag{108}$$

where the term K'' is given by

$$K'' = \frac{\nu(\mu+\nu+2)(M_2-1)}{\nu(\mu+\nu+2)+(\mu+2\nu+3)M_2}. \tag{109}$$

We note that these critical conditions for multiple sets of phenotypes are formally identical with the condition derived for only one phenotypic dimension. As shown in Fig. S14a, homophily is favored in both phenotypic sets if conditions (106) and (108) are both fulfilled. Otherwise, homophily is favored in either of the two phenotypic sets if only one of the two conditions holds (Fig. S14b). If neither of the two conditions is satisfied, natural selection favors heterophily in both sets (Fig. S14c).

Decreases in the number of phenotypes in both sets (M_1 and M_2), the preference mutation rate, u, and the phenotypic mutation rate, v, always make it easier for homophily to evolve, as shown in Fig. S15. We find good agreement between simulations and our analytical predictions (Figs. S14 and S15). These results are consistent with our above conclusions of single phenotypic dimension, suggesting that the rule for the evolution of homophily is robust to model variations.

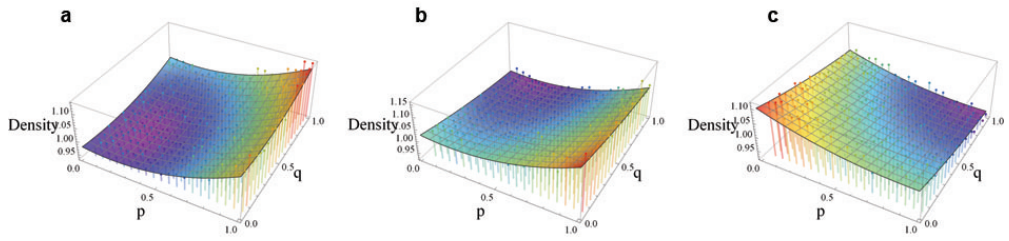

FIG. S14: Natural selection of homophily in two sets of phenotypes. Shown are the population equilibria, namely, probability density distribution of preferences $(p,\ q)$ over the unit $[0,\ 1]^2$, corresponding to (**a**) $a > K'b$ and $c > K''d$, (**b**) $a > K'b$ and $c < K''d$, and (**c**) $a < K'b$ and $c < K''d$. The scattered points are simulation results, which agree with our analytical theory. Parameters: $N = 50$, $\beta = 0.002$, $M_1 = 4$, $M_2 = 8$, $u = 0.06$, $v = 0.02$, (**a**) $a = 14/19$, $b = 1$, $c = 2/5$, $d = 1/2$, (**b**) $a = 14/19$, $b = 1$, $c = 1/10$, $d = 1/2$, (**c**) $a = 5/19$, $b = 1$, $c = 1/10$, $d = 1/2$. Results are averaged over $T = 10^9$ time steps.

VII. SUMMARY OF MODELS

We have analytically and by means of simulations studied the evolutionary origins of homophily. We derive a simple rule for the evolution of homophily, $a > Kb$. This rule is robust to model variations, as summarized in Table II. The mathematical framework presented here works well for weak selection.

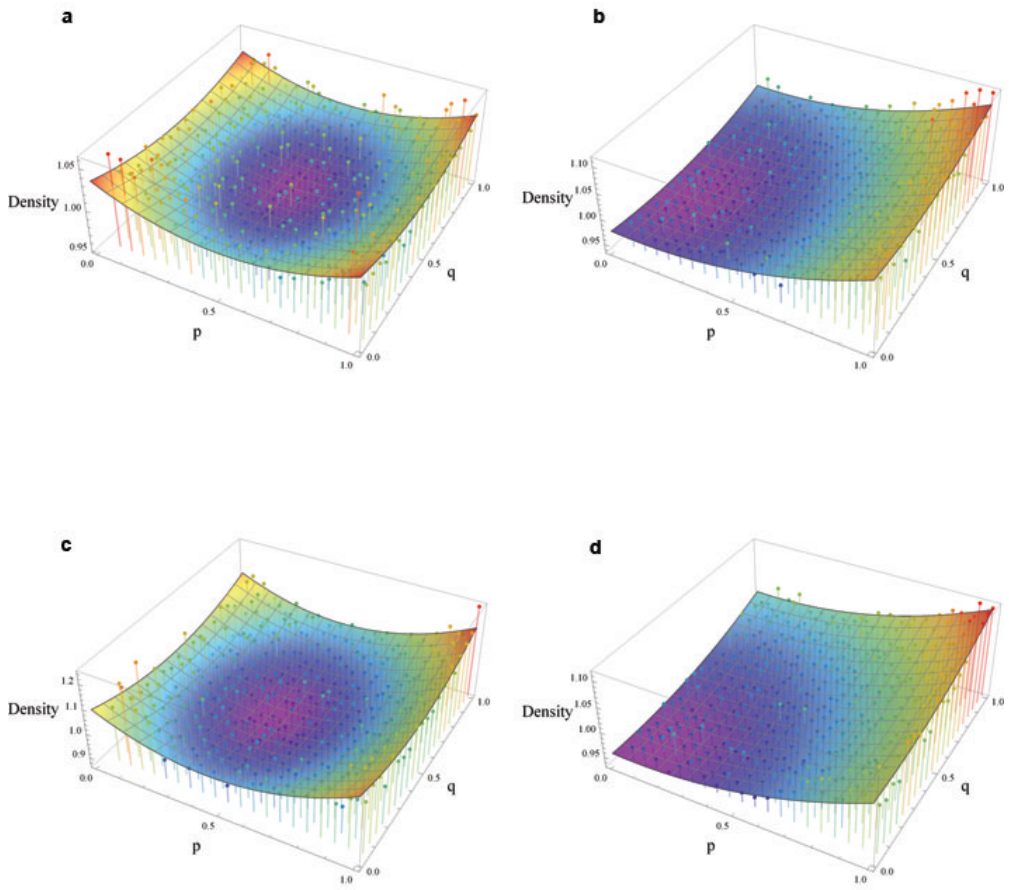

FIG. S15: Evolutionary determinants of homophily in two sets of phenotypes. The conditions $a = K'b$ and $c = K''d$ are fulfilled in (**a**). For other parameters fixed, decreases in (**b**) the number of phenotypes in both sets (M_1 and M_2), (**c**) the preference mutation rate, u, and (**d**) the phenotypic mutation rate, v, always make it easier for homophily to evolve. The scattered points are simulation results, which agree with our analytical theory. Parameters: $N = 50, \beta = 0.002, a = 9/19, b = 1, c = 3/10, d = 1/2$, (**a**) $M_1 = 4, M_2 = 8, u = 0.06, v = 0.02$, (**b**) $M_1 = 2, M_2 = 4, u = 0.06, v = 0.02$, (**c**) $M_1 = 4, M_2 = 8, u = 0.02, v = 0.02$, (**d**) $M_1 = 4, M_2 = 8, u = 0.06, v = 0.01$. Results are averaged over $T = 10^9$ time steps.

VIII. EMPIRICAL ESTIMATES OF HOMOPHILY

To estimate the average $\langle p \rangle$ for a population using real world data, we assume the social network under consideration is partitioned into M groups, each of size N_i ($i = 1, \cdots, M$) and $\sum_i N_i = N$. The expected total number n_{ij} of links of group i that are with group j can be calculated according

TABLE II: The simple rule for the evolution of homophily, $a > Kb$, is robust to model variations. See the text for more details.

Models	Condition for natural selection to favor homophily over heterophily
Basic model I (with unbiased matching)	$\langle p\rangle > 1/2 \Leftrightarrow a > Kb$
Extended model II (with biased matching)	[1] $\langle p\rangle > 1/2 \Leftrightarrow a > (1-\phi)Kb$
Extended model III (with local mutation, $\mu \to 0$)	[2] $\begin{cases} p^* < 1/2, \ a, b > 0; \\ p^* > 1/2, \ a, b < 0. \end{cases} \Leftrightarrow a > K_0 b$
Extended model IV (with full strategy set)	[3] $\langle p\rangle > \langle q\rangle \Leftrightarrow a > Kb$
Extended model V (with two sets of phenotypes)	[4] $\begin{cases} \langle p\rangle > 1/2 \Leftrightarrow a > K'b \\ \langle q\rangle > 1/2 \Leftrightarrow c > K''d \end{cases}$

[1] The parameter ϕ denotes the probability that individuals are assortatively matched to form bilateral social connections.

[2] $K_0 = \lim_{\mu\to 0} K$, and $p^* = bK_0/(a + bK_0)$.

[3] $\langle p\rangle$ and $\langle q\rangle$ denote, respectively, the average homophilic preference and the average heterophilic preference in equilibrium. Here the values of $(p, \ q)$ are constrained in the unit square, $[0, \ 1]^2$. Note that $\langle p\rangle > \langle q\rangle$ requires exactly the same condition as in the basic model where in effect $p + q = 1$ holds.

[4] The evolution of homophily in each set of phenotypes requires the same condition as in the basic model. Here, K' (K'') depends on the number of phenotypes in phenotypic set 1 (set 2).

to our homophily model: $n_{ij} = ppN_iN_j$ if $i = j$ (note that intra-group links are counted twice); otherwise $n_{ij} = (1-p)(1-p)N_iN_j$ if $i \neq j$. We can obtain the normalized mixing matrix $\mathbf{e} = \{e_{ij}\}$ given by our homophily model, where $e_{ij} = n_{ij}/\sum_{ij} n_{ij}$ is the fraction of links in the network that connect group i with group j. Let W_{ij} denote the observed number of links of individuals in group i that are with group j. Then the predicted number of links of individuals in group i that are with group j, $E_{ij} = e_{ij}\sum_i \sum_j W_{ij}$.

We use ordinary least squares to fit the model and estimate average homophilic preferences for each available population and phenotype. Specifically, empirical estimates of homophilic preferences are given by minimizing the sum $\chi^2(p) = \sum_{i=1}^{M} \sum_{j=1}^{M} (W_{ij} - E_{ij})^2$:

$$\hat{p} = \underset{0 \leq p \leq 1}{\operatorname{argmin}} \sum_{i=1}^{M} \sum_{j=1}^{M} (W_{ij} - E_{ij})^2. \tag{110}$$

The standard error of the estimate (the standard deviation of the residuals) is obtained by

$$\hat{\sigma}_e = \sqrt{\frac{\sum_{i,j} (W_{ij} - \hat{E}_{ij})^2}{z - k - 1}}, \tag{111}$$

where z is the total number of observations and k is the number of estimated parameters, and $\hat{E}_{ij}$ is evaluated at the estimated $\hat{p}$.

We identified six data sets that contained network information with published distributions of the age, sex, race, and/or caste of both individuals in each social tie. These data sets included three for humans, the Framingham Heart Study [9, 10], the National Longitudinal Study of Adolescent Health [11], and a study of 75 villages in rural Karnataka in Southern India [12, 13]. For all three of these studies we observed the phenotypes age and sex. We treated age as a discrete phenotype, truncating to the nearest decade for Framingham and India, and the nearest year for Add Health. Additionally, we measured the phenotype race in Add Health (black, Asian American, Native American, white, other) and caste in India (Scheduled Caste, Scheduled Tribe, OBC, General).

For animals, we measured homophily on sex for dolphins [14], *Colobus* monkeys [15], and zebras [16].

[1] Antal, T., Ohtsuki, H., Wakeley, J., Taylor, P. D. & Nowak, M.A. Evolution of cooperation by phenotypic similarity. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 8597–8600 (2009).

[2] Antal, T., Traulsen, A., Ohtsuki, H., Tarnita, C. E. & Nowak, M. A. Mutation-selection equilibrium in games with multiple strategies. *J. Theor. Biol.* **258,** 614–622 (2009).

[3] Tarnita, C. E., Wage, N. & Nowak, M. A. Multiple strategies in structured populations. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 2334–2337(2011).

[4] Tarnita, C. E., Antal, T. & Nowak, M. A. Mutation-selection equilibrium in games with mixed strategies. *J. Theor. Biol.* **261,** 50–57 (2009).

[5] Wakeley, J. *Coalescent Theory: An Introduction* (Roberts & Company Publishers, 2008).

[6] Nowak, M. & Sigmund, K. The evolution of stochastic strategies in the prisoner's dilemma. *Acta. Appl. Math.* **20,** 247–265 (1990).

[7] Metz, J. A. J., Geritz, S. A. H., Meszena, G., Jacobs, F. J. A. & van Heerwarden, J. S. Adaptive dynamics,a geometrical study of the consequences of nearly faithful reproduction. In: vanStrien, S.J., VerduynLunel, S.M.(Eds.), Stochastic and Spatial Structures of Dynamical Systems, K. Ned. Akad. VanWet. B, 45. North-Holland, Amsterdam, Holland, pp.183–231 (1996).

[8] Dieckmann, U. & Law, R. The dynamical theory of coevolution: A derivation from stochastic ecological processes. *J. Math. Biol.* **34,** 579–612 (1996).

[9] Dawber, T. R. *The Framingham study: The epidemiology of atherosclerotic disease* (Harvard Univ. Press, 1980).

[10] Christakis, N. A. & Fowler, J. H. The spread of obesity in a large social network over 32 years. *New Engl. J. Med.* **357,** 370–379 (2007).

[11] Harris, K. M., Bearman, P. S. & Udry, J. R. The national longitudinal study of adolescent health: Research design. Available at: http://www.cpc.unc.edu/projects/Add Health/design (2010).

[12] Banerjee, A., Chandrasekhar, A. G., Duflo, E., & Jackson, M. O. The diffusion of microfinance. Mimeo. Available at: http://econ-www. mit. edu/files/6910 (2011).

[13] Jackson, M. O., Rodriguez-Barraquer, T., & Tan, X. Social capital and social quilts: Network patterns of favor exchange. *Am. Econ. Rev.*, in press (2012).

[14] Lusseau, D. & Newman, M. E. J. Identifying the role that animals play in their social networks. *Proc. R. Soc. Lond. B* **271,** S477-S481 (2004).

[15] Kutsukake, N., Suetsugu, N. & Hasegawa, T. Pattern, distribution, and function of greeting behavior among black-and-white colobus. *Int. J. Primatol.* **27,** 1271–1291 (2006).

[16] Sundaresan, S. R., Fischhoff, I. R., Dushoff, J. & Rubenstein, D. I. Network metrics reveal differences in social organization between two fission-fusion species, Grevy's zebra and onager. *Oecologia* **151,** 140–149 (2007).