Moran model as a dynamical process on networks and its implications for neutral speciation

Marcus A. M. de Aguiar[1,2] and Yaneer Bar-Yam[1]
[1]*New England Complex Systems Institute, Cambridge, Massachusetts 02142, USA*
[2]*Instituto de Física "Gleb Wataghin," Universidade Estadual de Campinas, Unicamp 13083-859, Campinas, SP, Brazil*
(Received 15 April 2011; revised manuscript received 18 July 2011; published 1 September 2011)

In population genetics, the Moran model describes the neutral evolution of a biallelic gene in a population of haploid individuals subjected to mutations. We show in this paper that this model can be mapped into an influence dynamical process on networks subjected to external influences. The panmictic case considered by Moran corresponds to fully connected networks and can be completely solved in terms of hypergeometric functions. Other types of networks correspond to structured populations, for which approximate solutions are also available. This approach to the classic Moran model leads to a relation between regular networks based on spatial grids and the mechanism of isolation by distance. We discuss the consequences of this connection for topopatric speciation and the theory of neutral speciation and biodiversity. We show that the effect of mutations in structured populations, where individuals can mate only with neighbors, is greatly enhanced with respect to the panmictic case. If mating is further constrained by genetic proximity between individuals, a balance of opposing tendencies takes place: increasing diversity promoted by enhanced effective mutations versus decreasing diversity promoted by similarity between mates. Resolution of large enough opposing tendencies occurs through speciation via pattern formation. We derive an explicit expression that indicates when speciation is possible involving the parameters characterizing the population. We also show that the time to speciation is greatly reduced in comparison with the panmictic case.

DOI: 10.1103/PhysRevE.84.031901 PACS number(s): 87.23.Cc, 87.10.Hk, 02.50.Ga, 02.50.Ey

I. INTRODUCTION

A basic problem in population genetics is to predict how allele frequencies change in a population according to the underlying rules governing reproduction. For very large populations, the Hardy-Weinberg law applies and no change is expected between consecutive generations. However, for finite populations this is not necessarily true, and drift due to random variation can play an important role.

One of the first models to describe genetic drift in a finite population is the Wright-Fisher model [1]. It considers a population of N diploid individuals and a single gene with two alleles A_0 and A_1, so that there are a total of $2N$ genes. Given that the number of alleles A_1 in the population at time t is i, one can easily compute the probability to have j alleles A_1 at time $t+1$. Assuming that reproduction occurs by randomly picking $2N$ genes among the previous population with replacement and that there is no mutation, this probability is given by the binomial distribution

$$p_{ij} = \binom{2N}{j} (i/2N)^j [1-(i/2N)]^{2N-j}.$$

These *transition probabilities* form a matrix whose eigenvalues and eigenvectors contain all the information about the evolution of the system. Although the Wright-Fisher matrix is rather complicated, several analytical results can be extracted from it and even mutations can be included [1].

Other models were developed later that allowed for simpler mathematical treatment than the Wright-Fisher model or its generalization by Cannings [2]. Of particular importance is the Moran model [1,3,4], which considers haploid individuals and overlapping generations. Here a single hermaphroditic individual reproduces at each time step, with the offspring replacing the expiring parent. The transition probabilities can also be written down explicitly and all its eigenvalues and eigenvectors can be calculated for the case of zero mutations [5,6]. When mutations are included, the eigenvalues of the transition matrix and the stationary probability distribution, corresponding to the first eigenvector, can still be calculated [2,7].

Here we show that the Moran model can be mapped into a dynamical problem on networks, putting this classic model of population genetics in a broader and modern perspective. The mapping takes a panmictic population into a fully connected network, where the dynamical problem can be completely solved in terms of generating functions [8,9]. This provides a simple and elegant representation of the complete set of eigenvectors of the problem. The connection with the network dynamics yields, to our knowledge, the first complete solution of the Moran model.

Networks that are not fully connected map into nonrandom mating in structured populations. In particular, regular networks based on two-dimensional grids relate to spatially structured populations where mating is allowed only between neighbors. This, in turn, provides the basic mechanism of isolation by distance, as first proposed by Wright [10]. It has been recently shown [11] that this process can lead to speciation, termed topopatric speciation, and that the patterns of diversity that arise are fully compatible with the characteristics of biodiversity observed across many types of species in nature [12]. In this paper, we use the connection to the network dynamics to discuss analytically the mechanisms underlying topopatric speciation. We use approximate solutions for the network problem [9] that to our knowledge have not been obtained in the Moran model literature.

Speciation is the process by which multiple species are created from a single ancestral species. It can be triggered by geographic isolation, competition for resources, and genetic drift, among others [13,14]. The mechanism of genetic drift

 ©2011 American Physical Society

is termed "neutral" if the genes involved in speciation do not affect the fitness of the individuals.

The idea of neutral evolution, where the role of natural selection is not considered, has been disputed since its proposal by Kimura [15]. The debate was further fueled by the work of Hubbell [16], which demonstrated that realistic patterns of abundance distribution can be obtained within a neutral theory of biogeography in which species originate randomly [17–23]. Recent numerical simulations have shown that the same patterns of diversity emerge in explicitly genetic neutral models if reproduction is constrained by spatial and genetic proximity between individuals, and quantitative agreement of observed and simulated diversity was demonstrated [11]. Moreover, analysis suggests that if the number of genes involved in the process is very large, speciation may occur even without the spatial constraint [24,25]. Genetic proximity in mating may be imposed by a variety of mechanisms including sexual selection, which has been linked to speciation in the numerous species of cichlid fish in African lakes, that diverged from a common ancestral species only a few thousands years ago [26,27]. The theory developed in this paper sheds light on the way drift can lead to speciation and clarifies the role of the many parameters involved in these models, such as mutation rate, genome size, spatial and genetic restrictions, and population density, in the outcome of neutral evolution. The time to speciation is also estimated and shown to be greatly reduced in structured populations relative to the panmictic case, rebutting the key criticism of neutral processes as being very slow [28].

The paper is organized as follows: In Secs. II and III, we define the network dynamical system associated with the Moran process and write down its master equation and transition probabilities. In Sec. IV, we explicitly show how the Moran model can be mapped into this network problem. In Sec. V, we summarize the Moran-network properties: the distribution of allele frequencies at equilibrium, with its mean value and variance, and the limit of large populations. In Sec. VI, we discuss approximations for other network topologies, and in Sec. VII, we discuss their consequences for speciation.

II. THE NETWORK DYNAMICAL SYSTEM

Networks are mathematical structures composed of nodes and links between the nodes. The nodes often represent parts of a system and the links the interaction between the parts. Networks can model a wide range of systems in biology, engineering, and the social sciences [29]. In this work, we will associate nodes to a particular gene carried by individuals in a population, and links will be established between individuals that can mate with each other. In this section, networks will be treated as mathematical abstractions with a particular dynamics of network states; the connection with population genetics will be established in Sec. IV, although the correspondence with the Moran process is going to become evident as we proceed.

Consider a network with $N + N_0 + N_1$ nodes. To each node i we assign an internal state x_i that can take only the values 0 or 1. The nodes are divided into three categories: N nodes are free to change their internal state (according to the rule stated below); N_1 nodes are frozen in the state $x_i = 1$; and N_0 nodes are frozen in $x_i = 0$. The frozen nodes are assumed to be connected to all free nodes, and we consider them as perturbations to the "free" network, composed of the free nodes only. The information about the free network topology is contained in its adjacency matrix $\mathcal{A}$ defined as $\mathcal{A}_{ij} = 1$ if nodes i and j are connected, and $\mathcal{A}_{ij} = 0$ if they are not.

Our treatment of the model considers the most natural network case in which there is no self-connection, $\mathcal{A}_{ii} = 0$. This is also appropriate for the biological case of sexual reproduction without any probability of asexual replication. Changing the model to include the possibility of self-connection, $\mathcal{A}_{ii} = 1$, only slightly affects the mathematical treatment of the model, changing the number of neighbors of a node from $N + N_0 + N_1 - 1$ to $N + N_0 + N_1$. Formal expressions are changed accordingly, and analytic results, including time scales of the process, change only minimally since cloning is the same as not updating the system.

We refer to the free nodes connected to node i as its neighbors. The degree $k_i = \sum_j \mathcal{A}_{ij}$ is the number of neighbors of node i. The dynamics of the free nodes is defined as follows: at each time step, a node is selected at random to be updated. With probability p, the state of the node does not change, and with probability $1 - p$ it copies the state of one of its connected nodes, selected randomly among the k_i free neighbors or $N_0 + N_1$ frozen nodes. If the node to be updated is node i, then

$$x_i^{t+1} = \begin{cases} x_i^t & \text{with probability } p, \\ x_j^t & \text{with probability } \frac{1-p}{k_i+N_0+N_1}, \end{cases}$$

where j is connected to i.

We call this process an *influence dynamics*, since the state of a node changes according to the state of its neighbors. This system can model a number of interesting situations, such as, for example, the following:

(a) An election with two candidates in which some of the voters have a fixed opinion while the others change their intention according to the opinion of others.

(b) A sexually reproducing population of N haploid individuals in which the internal state represents two alleles of a gene. Taking $p = 1/2$, the update of a node mimics the mating of the focal individual with one of its neighbors. The focal individual is replaced by the offspring, which can take the allele of each parent with 50% probability. Since the free node can also copy the state of a frozen node, the values of N_0 and N_1 can be associated with mutation rates, as we will show later.

(c) A ferromagnetic material composed of atoms with magnetic moment $\pm 1/2$ interacting with an external magnetic field.

Although the influence process is very simple, its analysis can be quite complicated for networks of arbitrary topology. We first consider the simpler case of fully connected networks, where $\mathcal{A}_{ij} = 1$ if $i \neq j$, $\mathcal{A}_{ii} = 0$, and $k_i = N - 1$. Later we will discuss the consequences of other topologies and provide approximate results for these cases using the fully connected case as a basis.

III. MASTER EQUATION AND TRANSITION PROBABILITIES

For fully connected networks, the nodes are indistinguishable and there are only $N+1$ global states, which we call σ_k, $k = 0,1,\ldots,N$. The state σ_k has k free nodes in the state 1 and $N-k$ free nodes in the state 0. There is no need to count the frozen nodes, since they never change. If $P_t(m)$ is the probability of finding the network in the state σ_m at the time t, then $P_{t+1}(m)$ can depend only on $P_t(m)$, $P_t(m+1)$, and $P_t(m-1)$, since only one node is updated per time step. According to the updating rule given above, the dynamics of the probabilities is described by

$$\begin{aligned} P_{t+1}(m) = P_t(m)\Bigg\{ & p + \frac{(1-p)}{N(N+N_0+N_1-1)} \\ & \times [m(m+N_1-1)+(N-m)(N+N_0-m-1)]\Bigg\} \\ & + P_t(m-1)\frac{(1-p)}{N(N+N_0+N_1-1)}(m+N_1-1) \\ & \times (N-m+1) + P_t(m+1)\frac{(1-p)}{N(N+N_0+N_1-1)} \\ & \times (m+1)(N+N_0-m-1)\ . \end{aligned}$$

The term inside the first set of brackets gives the probability that the state σ_m does not change in that time step and is divided into two contributions: the probability p that the node does not change plus the probability $1-p$ that the node does change. In the latter case, the state of the node is $x_i = 1$ with probability m/N, and it may copy a different node in the same state, $x_j = 1$, with probability $(m-1+N_1)/(N+N_0+N_1-1)$. Also, if $x_i = 0$, which has probability $(N-m)/N$, it may copy another node $x_j = 0$ with probability $(N-m-1+N_0)/(N+N_0+N_1-1)$. The other terms are obtained similarly.

The probabilities $P_t(m)$ define a P_t vector of $N+1$ components. In terms of P_t, the above master equation can be written in matrix form as

$$P_{t+1} = U P_t \equiv \left[1 - \frac{(1-p)}{N(N+N_0+N_1-1)}A\right] P_t,$$

where the *evolution matrix* U, and also the auxiliary matrix A, is tridiagonal. The nonzero elements of A are independent of p and are given by

$$\begin{aligned} A_{m,m} &= 2m(N-m) + N_1(N-m) + N_0 m, \\ A_{m,m+1} &= -(m+1)(N+N_0-m-1), \\ A_{m,m-1} &= -(N-m+1)(N_1+m-1). \end{aligned}$$

These transition elements are the network dynamics analog of the Wright-Fisher transition probabilities described in the Introduction.

Let $\vec{a}_r$ and $\vec{b}_r$ be the right and left eigenvectors of U (and therefore of A) and λ_r the corresponding eigenvalues, so that $U\vec{a}_r = \lambda_r \vec{a}_r$ and $U^T \vec{b}_r = \lambda_r \vec{b}_r$. The transition probability between two states σ_M and σ_L after a time t can be written as

$$P(L,t;M,0) = \sum_{r=0}^{N} b_{rM} a_{rL} \lambda_r^t, \tag{1}$$

where a_{rL} and b_{rM} are the components of the right and left rth eigenvectors. The eigenvalues of U are given by

$$\lambda_r = 1 - \frac{(1-p)}{N(N+N_0+N_1-1)}\mu_r,$$

where μ_r are the eigenvalues of A. Equation (1) indicates that the λ_r have to be smaller than or equal to 1, otherwise $P(L,t;M,0)$ would eventually become larger than 1. Moreover, the eigenvectors corresponding to $\lambda = 1$ completely determine the asymptotic behavior of the system, since the contributions of all the others to $P(L,t;M,0)$ disappear at large times.

The eigenvalues of A are given by [9]

$$\mu_r = r(r-1+N_0+N_1),$$

which indeed implies that $0 \leqslant p \leqslant \lambda_r \leqslant 1$. Therefore, if and only if $N_0 = N_1 = 0$, there are two asymptotic (absorbing) states, corresponding to $r = 0$ and 1, given by σ_0 (all nodes in state 0) and σ_N (all nodes in state 1). Otherwise, there is only one possible asymptotic state, corresponding to $r = 0$. All other eigenvectors, related to the transient dynamics, can be calculated explicitly in terms of hypergeometric generating functions [9]. We do not write all of the eigenvalues and eigenvectors here because we are mostly interested in the equilibrium properties. However, the time to equilibration can be estimated from the value of the second largest eigenvalue, λ_1. Defining one unit of time (or one generation) as N updates of individual nodes, we find $\lambda_1^t \approx \exp(-t/\tau)$, where

$$\tau = \frac{(N+N_0+N_1-1)}{(1-p)(N_0+N_1)}. \tag{2}$$

IV. MAPPING THE MORAN MODEL ONTO NETWORK DYNAMICS

In order to map the evolution of a panmictic population of N hermaphroditic individuals into the fully connected network problem described above, we use the following notation: we associate x_i to the allele of the haploid individual i, which is either 0 for allele A_0 or 1 for allele A_1. At each time step, a random individual i is chosen to reproduce, and a random mate j is selected among the remaining $N-1$ individuals. The focal individual i is then replaced by the offspring.

Reproduction is carried out in two steps. The first step is the sexual reproduction itself: with probability 1/2 the allele x_i is passed to the offspring, and with probability 1/2 it takes the allele x_j. The second step takes mutation into account: after having taken the allele of the focal individual or its mate, the allele might change from 0 to 1 with probability μ_- or from 1 to 0 with probability μ_+. This corresponds to the Moran model with asymmetric mutations and is very similar to the influence process previously described for networks. In the framework of networks, the update of the node by keeping its own state or copying the state of a free neighbor corresponds to sexual reproduction. Copying the state of a frozen node represents mutation and its rate depends on N_0 and N_1.

However, the two processes are not quite the same: in the network dynamics, the frozen nodes play a role only if the node "decides" to copy a neighbor (probability $1-p$).

Here mutation acts even if the allele is passed from the focal individual i to the offspring. The master equation that includes mutation is therefore slightly different. Using $p = 1/2$, which is appropriate for unbiased reproduction, we have

$$\begin{aligned} P_{t+1}(m) = P_t(m) &\left\{ \frac{1}{2}\left(\frac{m}{N}\right)(1-\mu_+) + \frac{1}{2}\left(\frac{N-m}{N}\right)(1-\mu_-) + \frac{1}{2}\left(\frac{m}{N}\right)\left[\left(\frac{m-1}{N-1}\right)(1-\mu_+) + \left(\frac{N-m}{N-1}\right)\mu_-\right]\right. \\ &\left. + \frac{1}{2}\left(\frac{N-m}{N}\right)\left[\left(\frac{N-m-1}{N-1}\right)(1-\mu_-) + \left(\frac{m}{N-1}\right)\mu_+\right]\right\} + P_t(m-1)\left(\frac{N-m+1}{N}\right) \\ &\times \left[\frac{\mu_-}{2} + \frac{1}{2}\left(\frac{m-1}{N-1}\right)(1-\mu_+) + \frac{1}{2}\left(\frac{N-m}{N-1}\right)\mu_-\right] + P_t(m+1)\left(\frac{m+1}{N}\right) \\ &\times \left[\frac{\mu_+}{2} + \frac{1}{2}\left(\frac{N-m-1}{N-1}\right)(1-\mu_-) + \frac{1}{2}\left(\frac{m}{N-1}\right)\mu_+\right]. \end{aligned}$$

The first terms can be understood as follows: if the population has m individuals with allele A_1 at time t, it can remain that way in the next time step in several ways. First, if $x_i = 1$ (probability m/N), the offspring can keep the allele A_1 if it gets it from individual i (probability $1/2$) and it does not mutate after reproduction (probability $1-\mu_+$). Similarly, if $x_i = 0$ [probability $(N-m)/N$], the offspring can keep the allele A_0 if it gets it from individual i (probability $1/2$) and does not mutate after reproduction (probability $1-\mu_-$). The other terms have similar interpretations.

This equation is greatly simplified when written in matrix form. We obtain

$$P_{t+1} = U P_t \equiv \left[1 - \frac{(1+2\bar{\mu})}{2N(N-1)}A\right] P_t, \tag{3}$$

where the nonzero elements of A are given by

$$\begin{aligned} A_{m,m} &= 2m(N-m) + N_1(N-m) + N_0 m, \\ A_{m,m+1} &= -(m+1)(N-m-1+N_0), \\ A_{m,m-1} &= -(N-m+1)(m-1+N_1) \end{aligned}$$

with

$$\begin{aligned} N_1 &\equiv \frac{2\mu_-(N-1)}{1-2\bar{\mu}}, \\ N_0 &\equiv \frac{2\mu_+(N-1)}{1-2\bar{\mu}}, \end{aligned} \tag{4}$$

and

$$\bar{\mu} = \frac{\mu_+ + \mu_-}{2}. \tag{5}$$

With our mapping of the parameters, this is identical to the original matrix A of the network dynamics. Therefore, all the known solutions of the network problem can be directly transferred to the genetic problem via the above relationships between the mutation rates μ_- and μ_+ and the frozen nodes N_0 and N_1. These solutions are described in the next section. In particular, the time to equilibration, Eq. (2), maps onto

$$\tau_f = \frac{1}{2\bar{\mu}} \tag{6}$$

for large N and small mutation rates, where the subscript f denotes *fully mixed* populations.

V. EQUILIBRIUM DISTRIBUTION

The cases $N_0 = 0$ or $N_1 = 0$, corresponding to $\mu_+ = 0$ or $\mu_- = 0$, are trivial since all individuals in the population will eventually become identical, with allele A_0 or A_1, respectively. If N_0 and N_1 are both zero, the individuals will also eventually become identical, but the probability of each outcome, all A_0 or all A_1, depends on the initial distribution of alleles in the population.

If N_0 and N_1 are both nonzero, the probability of finding m nodes in state 1, or m individuals with allele A_1, in equilibrium is given by [1,7,9]

$$\rho(k) = A(N, N_0, N_1)\frac{\Gamma(N_1+k)\Gamma(N+N_0-k)}{\Gamma(N-k+1)\Gamma(k+1)}, \tag{7}$$

where

$$A(N, N_0, N_1) = \frac{\Gamma(N+1)\Gamma(N_0+N_1)}{\Gamma(N+N_0+N_1)\Gamma(N_1)\Gamma(N_0)} \tag{8}$$

is a normalization constant and $\Gamma(x)$ is the Gamma function. This result is valid even if N_0 and N_1 are not integers. In an actual network, when N_0 and N_1 are integers, the Γ functions can be replaced by factorials.

Note that, because of the mutation rates (or frozen nodes), a particular realization of the dynamics never stabilizes in a particular state: the number of individuals with allele A_1 continues to change. The probability of finding the population with m alleles A_1, however, is independent of the time, and given by the expression above. One interesting feature of this solution is that for $N_0 = N_1 = 1$, we obtain $\rho(m) = 1/(N+1)$ for all values of m, meaning that all states are equally likely, independent of the population size.

The mean value $m_0 = \sum_m m\rho(m)$ and the variance $\sigma_2 = \sum_m m^2\rho(m) - \bar{m}^2$ can also be calculated explicitly. We obtain

$$m_0 = N\frac{N_1}{N_0+N_1} \tag{9}$$

and

$$\sigma_2 = \frac{N N_1 N_0 (N_1+N_0+N)}{(N_1+N_0)^2(1+N_1+N_0)}. \tag{10}$$

Higher-order correlations can also be calculated explicitly, but the results become progressively more complicated.

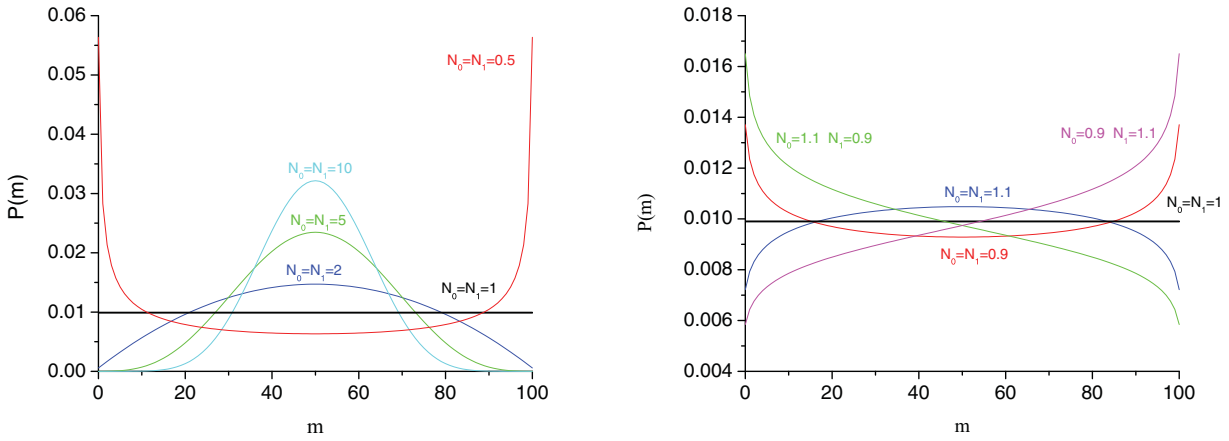

FIG. 1. (Color online) Asymptotic probability distribution for a network with $N = 100$ nodes and several values of N_0 and N_1.

Figure 1 shows a few examples of the distribution $\rho(m)$ for a network with $N = 100$ and various values of N_0 and N_1.

If N is very large, $\rho(m)$ peaks around m_0 and can be approximated by a Gaussian:

$$\rho(m) = \rho_0 \exp - \left[\frac{(m - m_0)^2}{2\Delta^2} \right]$$

with

$$\Delta = \left[\frac{N N_0 N_1 (N + N_0 + N_1)}{(N_0 + N_1)^3} \right]^{1/2}$$

and

$$\rho_0 = \frac{1}{\sqrt{2\pi}\,\Delta}.$$

In terms of the continuous variables $x = m/N$, $n_0 = N_0/N$, and $n_1 = N_1/N$, we can also write

$$\rho(x) = \rho_0 \exp - \left[\frac{(x - x_0)^2}{2\delta^2} \right]$$

with

$$\delta = \left[\frac{n_0 n_1 (1 + n_0 + n_1)}{N(n_0 + n_1)^3} \right]^{1/2},$$

$x_0 = m_0/N$, and $\rho_0 = 1/\sqrt{2\pi}\delta$, showing that the width of the distribution goes to zero as N goes to infinity, in agreement with the Hardy-Weinberg law.

VI. STRUCTURED NETWORKS

For networks that are not fully connected, the effect of the frozen nodes is amplified. To see this, we note that the probability that a free node copies a frozen node is $P_i = (N_0 + N_1)/(N_0 + N_1 + k_i)$, where k_i is the degree of the node. For fully connected networks, $k_i = N - 1$ and $P_i \equiv P_{\rm FC}$. For general networks, an average value $P_{\rm av}$ can be calculated by replacing k_i by the average degree $k_{\rm av}$. We can then define effective numbers of frozen nodes, $N_{0\rm ef}$ and $N_{1\rm ef}$, as being the values of N_0 and N_1 in $P_{\rm FC}$ for which $P_{\rm av} \equiv P_{\rm FC}$. This leads to

$$N_{0\rm ef} = f N_0, \qquad N_{1\rm ef} = f N_1, \tag{11}$$

where $f = (N - 1)/k_{\rm av}$. For well-behaved distributions, corrections involving higher moments can be obtained by integrating P_i times the degree distribution and expanding around $k_{\rm av}$.

Figure 2 shows examples of the equilibrium distribution for four different networks with $N = 100$ and $N_0 = N_1 = 5$. Panel (a) shows the result for a random network constructed by connecting any pair of nodes with probability 0.3. In this case, $k_{\rm av} = 29.7$ and $f = 3.3$. The theoretical result was obtained with Eq. (7) with $N_{0\rm ef} = N_{1\rm ef} = 17$. For a scale-free network [panel (b)] grown from an initial cluster of six nodes adding nodes with three connections each following the preferential attachment rule [29], $f = 99/6$ and the effective values of N_0 and N_1 are approximately 82. Panel (c) shows the probability distribution for a finite two-dimensional (2D) regular lattice with 10×10 nodes connected to nearest neighbors for which $k_{\rm av} = 3.6$ (the nodes near the boundaries have fewer than four links), $f = 99/3.6 \approx 28$. Finally, panel (d) shows a small world version of the regular lattice [29], where 30 connections were randomly reallocated, creating shortcuts between otherwise distant nodes. These results show that the approximate rescaling of frozen nodes (or, equivalently, the mutation rates) is accurate for many network topologies. Still, extreme cases such as a star network do present different distributions, and this is confirmed by simulations.

The time to equilibration in structured networks changes to

$$\tau = \frac{2(k_{\rm av} + N_0 + N_1)}{N_0 + N_1}, \tag{12}$$

which can be considerably smaller than the case of fully connected networks.

VII. SPECIATION AND BIODIVERSITY

In the previous sections, we derived two important theoretical results: (a) the relationship between the process of influence dynamics on networks and the Moran model; (b) the approximate equilibrium distribution for structured networks, obtained by rescaling the number of frozen nodes. We will show now that these two results allow us to infer important properties about the genetic evolution of spatially extended populations.

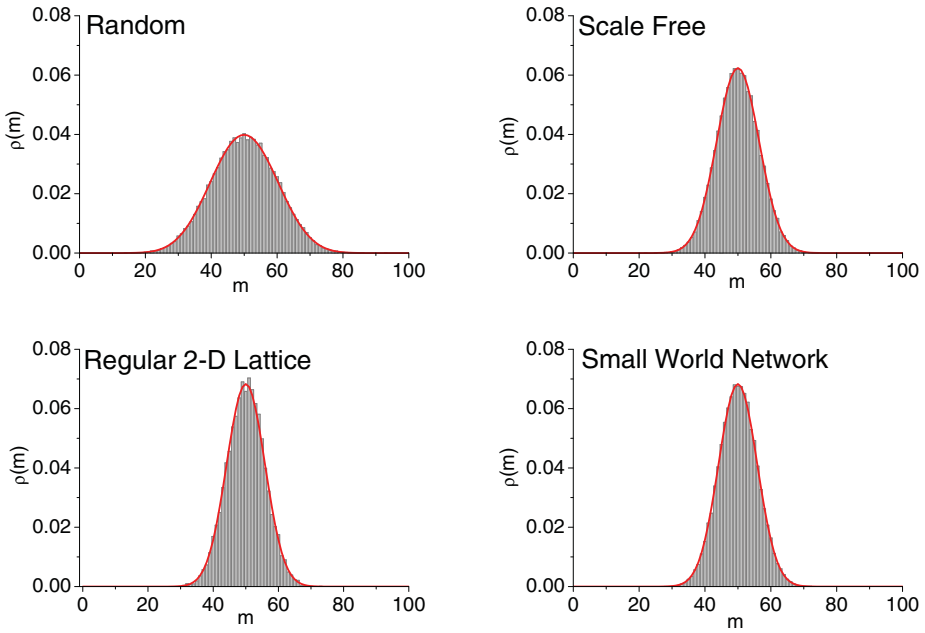

FIG. 2. (Color online) Equilibrium probability distribution for networks with different topologies. In all cases, $N = 100$, $N_0 = N_1 = 5$, $t = 10\,000$, and the number of simulations is 50 000 (histograms). The theoretical curve (solid red) is drawn with effective numbers of frozen nodes $N_{0\mathrm{ef}} = fN_0$ and $N_{1\mathrm{ef}} = fN_1$: (a) random network $N_{0\mathrm{ef}} = N_{1\mathrm{ef}} = 17$; (b) scale-free $N_{0\mathrm{ef}} = N_{1\mathrm{ef}} = 82$; (c) regular 2D lattice $N_{0\mathrm{ef}} = N_{1\mathrm{ef}} = 140$; (d) small world network $N_{0\mathrm{ef}} = N_{1\mathrm{ef}} = 140$.

It has been recently shown [11,30] that when mating is constrained by both spatial and genetic proximity between individuals, neutral evolution by drift alone can lead to speciation, i.e., to the spontaneous breakup of the population into reproductively isolated groups. Moreover, the patterns of abundance distributions generated by this mechanism are compatible with those observed in nature [11]. In what follows, we discuss the process of neutral speciation promoted by spatial and genetic constraints, termed *topopatric speciation*, in light of the theory developed above.

To make the analysis simpler, we restrict ourselves to the case of symmetric mutation rates, $\mu_- = \mu_+ \equiv \mu$, or, equivalently, an equal number of frozen nodes $N_0 = N_1 \equiv N_z$. In this case, the connection between mutations and frozen nodes simplifies to

$$N_z = \frac{2\mu(N-1)}{1-2\mu}. \tag{13}$$

Let P_{id} be the probability that two individuals picked at random in the population have identical genes at equilibrium. This is given by the sum of the probabilities that their alleles are both A_1 or both A_0:

$$P_{\mathrm{id}} = \sum_{m=0}^{N} \rho(m) \left[\frac{m}{N}\frac{m-1}{N-1} + \frac{N-m}{N}\frac{N-m-1}{N-1} \right]$$
$$= 1 + \frac{2}{N(N-1)}[\sigma^2 + \langle m\rangle^2 - N\langle m\rangle].$$

Using Eqs. (9), (10), and (13), we obtain

$$P_{\mathrm{id}} = \frac{1+N_z}{1+2N_z} = \frac{1+2\mu(N-2)}{1+2\mu(2N-3)}. \tag{14}$$

The probability that the two individuals are different, which is the heterozygosity, is

$$P_{\mathrm{ht}} = 1 - P_{\mathrm{id}} = \frac{2\mu(N-1)}{1+2\mu(2N-3)} \approx \frac{2\mu N}{1+4\mu N}, \tag{15}$$

where the approximation holds for $N \gg 1$.

Consider now a population in equilibrium where the N individuals have B independent genes [11,24,30–33]. The average genetic distance between two individuals is

$$\langle d\rangle = B P_{\mathrm{ht}} \approx \frac{B}{2}\left(\frac{4\mu N}{1+4\mu N}\right). \tag{16}$$

This expression provides a connection between the size of the population and the average genetic distance between individuals, which is a measure of diversity within the population. Two interesting relations can be derived from this equation: first, for given B and μ we can calculate the size N_G that corresponds to a particular average genetic distance $\langle d\rangle = G$:

$$N_G = \frac{G}{2\mu(B-2G)}. \tag{17}$$

Second, for given N and B, we calculate the mutation rate μ_G that corresponds to $\langle d\rangle = G$:

$$\mu_G = \frac{G}{2N(B-2G)}. \tag{18}$$

Notice that $N_G\mu = N\mu_G$.

When mating in panmictic populations is constrained by genetic proximity between individuals, so that pairs whose genetic distance is larger than G are incompatible, the distribution of genetic distances stays very close to $\langle d\rangle = G$, as if the genome had an effective size $B_{\mathrm{ef}} = 2G$. On the other hand, if mating is constrained by spatial proximity, the effective mutation rate tends to increase. Indeed, a spatial restriction in mating corresponds to influence processes on networks constructed over regular lattices, which amplifies the effect of frozen nodes and, therefore, of mutations.

Consider a spatial area with L^2 lattice sites and periodic boundary conditions. A population resides in this area, where each individual is a node in the resulting network of potential mates. Due to spatial mating restrictions, a node is connected to neighbors that are within a distance S from itself (measured in units of lattice spacing). Let N be the number of individuals

TABLE I. List of parameters in the speciation model.

Parameter	Description	Value used for figures
B	number of biallelic genes	125
G	maximum genetic difference for mating	20
μ	mutation rate per gene	0.001
L	linear size of spatial environment	128
S	spatial radius of mating neighborhood	6
N	population size, held fixed	2000
$k_{\rm av}$	average number of individuals in spatial neighborhood S	14
P	minimum number of potential mates in mating neighborhood	8
$S_{\rm min}$	spatial radius of mating neighborhood containing P individuals	4.6
S_c	critical value of S above which no speciation occurs	Eq. (24)
G_c	critical value of G above which no speciation occurs	Eq. (25)

in the population, so that the density is $\rho = N/L^2$. The area where an individual can look for a mate, its "mating neighborhood," is approximately πS^2. The average degree of the network is the density times the area $k_{\rm av} = \pi N S^2/L^2$. Table I displays a list of the parameters and variables involved in the process.

According to our discussion in Sec. VI, the resulting population distribution can be modeled using a fully connected network with an effective number of frozen nodes,

$$N_{\rm ef} = f N_z = \frac{N-1}{k_{\rm av}} N_z \approx \frac{L^2}{\pi S^2} N_z. \tag{19}$$

The corresponding effective mutation rate is obtained from Eq. (13),

$$N_{\rm ef} = \frac{2\mu_{\rm ef}(N-1)}{1-2\mu_{\rm ef}},$$

which gives

$$\mu_{\rm ef} = \frac{f}{1+2\mu(f-1)}\mu \approx \frac{\mu f}{1+2\mu f}. \tag{20}$$

Note that $\mu_{\rm ef} \to 1/2$ for $\mu f \gg 1$.

When mating between individuals is constrained only by their spatial distance, as measured by the parameter S, the effective mutation rate Eq. (20) can be dramatically enhanced with respect to a panmictic population. This, in turn, increases the average genetic distance between individuals, which approaches $B/2$ for large populations and fixed $k_{\rm av}$ (corresponding to large values of N_z). The distribution of genetic distances approaches a broad symmetric distribution.

On the other hand, if mating is constrained only by the genetic distance between individuals, the distribution of genetic distances shrinks to about G. This corresponds to an effective reduction in genome size from B to $2G$.

When both spatial and genetic restrictions are present, as in Ref. [11], the population feels a large effective mutation rate, tending to spread out the genome distribution. On the other hand, the individuals are compelled by the mating condition to stay genetically close to each other. The only stable outcome of these opposing forces is the formation of local groups where $\langle d \rangle \leqslant G$ within the group but $\langle d \rangle > G$ among groups. This characterizes the groups as reproductively isolated from each other and, therefore, as separate species.

The average number of individuals in each group is given approximately by N_G, Eq. (17), which is usually much smaller than N. This also implies that the individuals within groups are highly connected to each other, so that $f \approx 1$ and $\mu_{\rm ef} \approx \mu$, restoring the equilibrium of the system.

The conditions for speciation can be estimated as follows. When S is very large, the effect of the genetic mating restriction is to reduce the effective size of the genome, $B_{\rm ef}$, from B to $2G$, so that, from Eq. (16), $\langle d \rangle$ is at most G. As S is reduced, the effective mutation rate increases and additional genes are incorporated into the effective (variable) genome, increasing the average genetic distance between individuals. When $\langle d \rangle$ becomes larger than about $2G$, the population can no longer hold itself together and splits. This has been confirmed by numerical simulations as illustrated in Fig. 3.

We write

$$B_{\rm ef} = 2G + (B-2G)\mathcal{P}, \tag{21}$$

where $\mathcal{P}$ is the probability that a new gene is fixed into the effective genome.

$\mathcal{P}$ goes to 0 for large values of S and reaches 1 for small S. It must depend only on the mutation rate μ, genome length B, and the size of the local mating population $\pi S^2 \rho = \pi S^2 N/L^2 = k_{\rm av}$. This local mating population has to be at least 2, otherwise mating is not possible. More generally, if the minimum number of potential mates for reproduction is P, we can define the minimum S by $\pi S_{\rm min}^2 \rho = P$, or

$$S_{\rm min} = L\sqrt{P/\pi N}. \tag{22}$$

The probability $\mathcal{P}$ must be small if the local mating population is large. On the other hand, it must increase with the mutation rate and size of the genome. We may therefore write the ansatz

$$\mathcal{P} = \exp\left\{-c\left[\frac{\pi(S-S_{\rm min})^2 N/L^2)}{B\mu}\right]^2\right\}$$

or

$$\mathcal{P} = \exp\left\{-\frac{\pi^2(S-S_{\rm min})^4 N^2}{\gamma^4 L^4 B^2 \mu^2}\right\}, \tag{23}$$

where the constant of proportionality c is rewritten as γ^{-4} for convenience. The exponential dependence of $\mathcal{P}$ on the square of $k_{\rm av}/B\mu$ is suggested by numerical simulations.

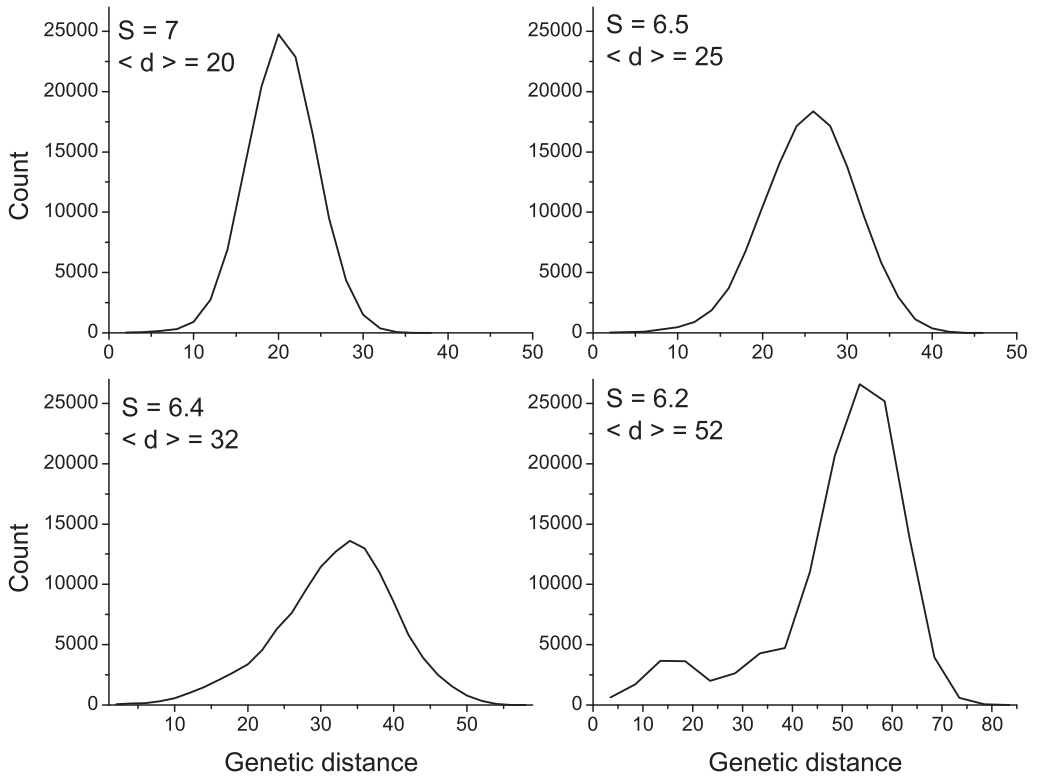

FIG. 3. Distribution of genetic distances between pairs of individuals for different vales of S. For large S, the average genetic distance $\langle d \rangle$ is of the order of G. As S is decreased, $\langle d \rangle$ increases and speciation occurs when $\langle d \rangle$ is about $2G$, as indicated by the breakup of the unimodal distribution in the lower right panel into a proximate part (within species distances) and a remote part (interspecies distances). In this example, $N = 2000$, $G = 20$, $\mu = 0.001$, $B = 125$, $L = 128$, and $P = 8$.

The condition for speciation is

$$\langle d \rangle = \frac{B_{\rm ef}}{2}\left(\frac{4\mu_{\rm ef}N}{1+4\mu_{\rm ef}N}\right) \gtrsim 2G.$$

Since the μN is usually of order 1 in most simulations, and $\mu_{\rm ef} \gg \mu$, the factor $4\mu_{\rm ef}N/(1+4\mu_{\rm ef}N)$ can be safely approximated by 1. Using Eqs. (21) and (23), we obtain

$$\frac{\pi^2(S-S_{\rm min})^4N^2}{\gamma^4L^4\mu^2B^2} \lesssim \log\left(\frac{B-2G}{2G}\right)$$

or

$$S \lesssim S_{\rm min} + \gamma L\sqrt{\frac{B\mu}{N\pi}}\left[\log\left(\frac{B-2G}{2G}\right)\right]^{1/4} \equiv S_c(G). \quad (24)$$

Inverting this equation, we obtain

$$G \lesssim \frac{B/2}{1+\exp\left(\frac{\pi^2N^2(S-S_{\rm min})^4}{\gamma^4\mu^2B^2L^4}\right)} \equiv G_c(S), \quad (25)$$

which gives the minimum value of G for a given S.

Equation (24) gives the maximum size of the mating neighborhood for which speciation is possible. This analytical result describes the dependence of speciation on six model parameters: B, G, μ, P, L, and N. It provides a very good quantitative estimate for the parameter region where speciation is possible, as illustrated in Fig. 4. The result also incorporates cutoffs at $G = B/4$ and at $S_{\rm min}$, which are in agreement with numerical simulations [11]. Furthermore, it also gives the scaling dependence of S_c on these various parameters. In particular, it predicts speciation at large values of S if B is sufficiently large. This corroborates the results in [24,25] but

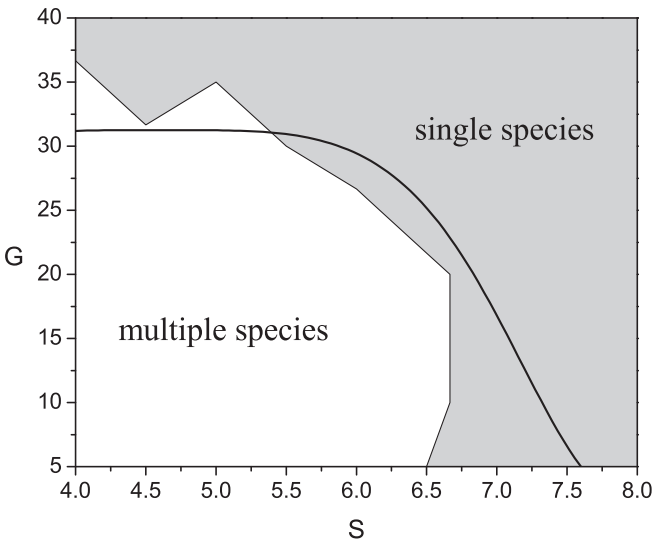

FIG. 4. Parameter region in the S-G plane where speciation is possible according to Eq. (25) (thick line) and numerical simulations (solid line and shading [11]). The other parameters are the same as in Fig. 3 and Table I with $\gamma = 4.2$.

shows that such space-independent speciation occurs only for very large values of B, since S_c increases with $B^{1/2}$.

Our analytical results constitute an important addition to the simulations presented in [11] and contribute to the understanding of the significant role of drift in speciation [11,13,16,20,21,24,30]. Equation (24) identifies the combination of parameters that makes this possible. For example, low mutation rates, which hinder speciation, can be compensated by a large number of participating genes or by low population density.

Finally, we address the frequent criticism that the time scale for speciation by drift is too long [13,28] and, therefore, should be very rare. The time to speciation should be proportional to the equilibration time of the dynamics, measured in number of generations. For a well-mixed population we found that $\tau_f = 2/\mu$, which can indeed be very large in realistic cases in which μ is very small. However, for structured populations, Eq. (12) maps into

$$\tau_s = \frac{2(k_{\rm av} + 4\mu N)}{4\mu N}, \tag{26}$$

where $k_{\rm av} = \pi S^2 \rho$ is the local population size within a mating neighborhood S. If $k_{\rm av} \gg 4\mu N$, we obtain $\tau \approx k_{\rm av}/2\mu N$, which is smaller than the fully mixed time by a factor $k_{\rm av}/N$ and can speed up speciation by several orders of magnitude. In our example, we find $\tau_s/\tau_f \approx 0.01$. Equivalently, this can be considered to be a result of the enhancement in the effective mutation rate in structured populations, as discussed above.

VIII. CONCLUSIONS

The process of speciation underlies the creation of the tree of life. Fossil records and molecular analysis allow the construction of detailed phylogenetic trees linking species to their ancestors, identifying the branching points of speciation. The way speciation occurred in each case, however, is rarely known with certainty and several mechanisms have been considered. A recently proposed mechanism of speciation [11] demonstrated that a spatially extended population can break up spontaneously into species when subjected to mutations and to spatial and genetic mating restrictions, even in the absence of natural selection. Numerical simulations have shown that this mechanism, termed topopatric speciation, occurs for a restricted range of parameters, which include population size N, mutation rate μ, and the parameters S and G controlling the spatial and genetic mating restrictions.

In this paper, we have introduced a mapping of genetic dynamics in an evolving population onto the dynamics of influence on a network, and we used this mapping to analytically study the process of topopatric speciation. This mapping gives, to our knowledge, the first complete solution of the Moran model, providing an elegant representation of the complete set of eigenvectors of the problem.

We have shown that, while fully connected networks correspond to panmictic populations, certain structured networks can be mapped into dynamic spatially extended populations. Moreover, the mapping shows that limiting mating to a fraction of the total population by network connections increases the effective mutation rate as compared to the panmictic case, and increases the genetic diversity of the population. By extending the model from one to multiple independent biallelic genes, we have shown that a genetic restriction on mating decreases the effective size of the genome, decreasing diversity. These opposing forces are resolved not by compromise but by pattern formation, breaking up the population into multiple species. This process, and its dependence on the most relevant characteristics of the population, is accurately described by Eq. (25). This equation provides a new and important tool to understand neutral speciation, revealing explicitly the relationships among the parameters involved in the process and the interplay of genetic processes whose opposition leads to spontaneous speciation.

ACKNOWLEDGMENTS

We thank Elizabeth M. Baptestini for helpful comments. M.A.M.A. acknowledges financial support from CNPq and FAPESP.

[1] W. J. Ewens, *Mathematical Population Genetics I. Theoretical Introduction Series: Biomathematics*, Vol. 9. (Springer-Verlag, New York, 1979).
[2] C. Cannings, Adv. Appl. Prob. **6**, 260 (1974).
[3] P. A. P. Moran, Proc. Cambridge Philos. Soc. **54**, 60 (1958).
[4] J. Wakeley, *Coalescent Theory* (Roberts & Company, Greenwood Village, Colorado, 2009).
[5] G. A. Watterson, Ann. Math. Statist. **32**, 716 (1961).
[6] K. Gladstien, Siam J. Appl. Math **34**, 630 (1978).
[7] J. H. Gillespie, *Population Genetics: A Concise Guide* (Johns Hopkins University Press, Baltimore, MD, 2004).
[8] M. A. M. de Aguiar, I. R. Epstein, and Y. Bar-Yam, Phys. Rev. E **72**, 067102 (2005).
[9] D. D. Chinellato, M. A. M. de Aguiar, I. R. Epstein, D. Braha, and Y. Bar-Yam, e-print arXiv:0705.4607v2 [nlin.SI].
[10] S. Wright, Genetics **28**, 114 (1943).
[11] M. A. M. de Aguiar, M. Baranger, E. M. Baptestini, L. Kaufman, and Y. Bar-Yam, Nature (London) **460**, 384 (2009).
[12] M. L. Rosenzweig, *Species Diversity in Space and Time* (Cambridge University Press, Cambridge, 1995).
[13] J. A. Coyne and H. A. Orr, *Speciation* (Sinauer Associates, Sunderland, MA, 2004).
[14] C. Pinho and J. Hey, Annu. Rev. Ecol. Evol. Syst. **41**, 215 (2010).
[15] M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, UK, 1883).
[16] S. P. Hubbell, *The Unified Neutral Theory of Biodiversity and Biogeography* (Princeton University Press, Princeton, NJ, 2001).
[17] S. Gavrilets, H. Li, and M. D. Vose, Evolution **54**, 1126 (2000).
[18] S. Nee and G. Stone, Trends Ecol. Evol. **18**, 433 (2003).
[19] J. R. Banavar and A. Maritan, Nature (London) **460**, 334 (2009).

[20] M. Kopp, BioEssays **32**, 564 (2010).
[21] H. Ter Steege, Biotropica **42**, 631 (2010).
[22] R. S. Etienne and B. Haegeman, Theor. Ecol. **4**, 87 (2011).
[23] J. Rosindell, S. P. Hubbell, and R. S. Etienne, Trends in Ecology & Evolution **26**, 340 (2011).
[24] P. G. Higgs and B. Derrida, J. Phys. A **24**, L985 (1991).
[25] P. G. Higgs and B. Derrida, J. Mol. Evol. **35**, 454 (1992).
[26] O. Seehausen *et al*., Nature (London) **455**, 620 (2008).
[27] M. Kirkpatrick and T. Price, Nature (London) **455**, 601 (2008).
[28] M. Nei, T. Maruyama, and C.-I. Wu, Genetics **103**, 557 (1983).
[29] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[30] G. A. Hoelzer, R. Drewes, J. Meier, and R. Doursat, PLoS Comput. Biol. **4**, e1000126 (2008).
[31] Y.-C. Zhang, Phys. Rev. E **55**, R3817 (1997).
[32] M. Hall, K. Christensen, S. A. di Collobiano, and H. J. Jensen, Phys. Rev. E **66**, 011904 (2002).
[33] K. Jain, Phys. Rev. E **76**, 031922 (2007).