The 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*

[2] *Instituto de Física 'Gleb Wataghin', Universidade Estadual de Campinas, Unicamp 13083-970, Campinas, SP, Brasil*

Abstract

In genetics the Moran model describes the neutral evolution of a bi-allelic 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 new 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 to 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 take place: increasing diversity promoted by enhanced effective mutations versus decreasing diversity promoted by similarity between mates. Stabilization occurs with speciation via pattern formation.

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 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

$$p_{ij} = \left(\begin{array}{c} 2N \\ j \end{array} \right) (i/2N)^j [1-(i/2N)^{2N-j}].$$

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]. This provides

a simple and elegant representation of the complete set of eigenvectors of the problem. The connection with the network dynamics gives, to our knowledge, the first complete solution of the Moran model.

Networks that are not fully connected map into non-random 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 Sewall Wright [9]. It has been recently shown [10] that, not only this process can lead to speciation, but also that the patterns of diversity that arise are fully compatible with the universal characteristics observed in nature [11] . Although no exact solution exists for the Moran model for structured populations, approximate solutions do exist for the equivalent network problem [8]. In this paper we explore this connection to discuss the mechanisms underlying the so called *topopatric* speciation [10].

The paper is organized as follows: in sections II and III we define the network dynamical system associated to the Moran process and write down its master equation and transition probabilities. In section IV we show how the Moran model can be mapped into this network problem. In section V we summarize the Moran-Network main properties: the distribution of allele frequencies at equilibrium, with its mean value and variance, and the limit of large populations. In section VI we discuss approximations for other network topologies and, in section VII, 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 [12]. 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 abstract; the connection with population genetics will be established in section IV, although the analogy 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 which 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, $\mathcal{A}_{ij} = 0$ if they are not and $\mathcal{A}_{ii} = 0$. We refer to the free nodes connected to i as their neighbors. The degree $k_i = \sum_j \mathcal{A}_{ij}$ is the number of neighbors of node i.

The dynamics on 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 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:

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

(b) A sexually reproducing population of N haploid individuals where 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 will 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, that we call σ_k, $k = 0, 1, ..., N$. 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 above, the dynamic of the probabilities is described by the following equation:

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

$$P_t(m-1) \frac{(1-p)}{N(N+N_0+N_1-1)} (m+N_1-1)(N-m+1) +$$

$$P_t(m+1) \frac{(1-p)}{N(N+N_0+N_1-1)} (m+1)(N+N_0-m-1) \; .$$

The term inside the first 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 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

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

where the *evolution matrix* U, and also the auxiliary matrix A, is tri-diagonal. The non-zero 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 analogue of the Wright-Fisher transition probabilities described in the Introduction for the network dynamics.

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 the 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 r-th 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 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 $P(L,t;M,0)$ die out at large times.

The eigenvalues of A are given by [8]

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

which indeed implies that $0 \leq p \leq \lambda_r \leq 1$. Therefore, if and only if $N_0 = N_1 = 0$ there are two asymptotic (absorbing) states, corresponding to $r = 0$ and $r = 1$, given by σ_0 (all node 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 [8]. We do not write them down here because we are only interested in equilibrium properties.

IV. MAPPING THE MORAN MODEL

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 value 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 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:

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

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

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

where the non-zero 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

$$
\begin{aligned}
N_1 \; &\equiv \frac{2\mu_-(N-1)}{1-2\bar{\mu}} \\
N_0 \; &\equiv \frac{2\mu_+(N-1)}{1-2\bar{\mu}}
\end{aligned} \tag{3}
$$

and

$$\bar{\mu} = \frac{\mu_+ + \mu_-}{2}. \tag{4}$$

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 relation between the mutation rates μ_- and μ_+ and the frozen nodes N_0 and N_1. These solutions are described in the next section.

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, depend on the initial distribution of alleles in the population.

If N_0 and N_1 are both non-zero, the probability of finding m nodes in state 1, or m individuals with allele A_1, in equilibrium is given by [1, 7, 8]

$$\rho(k) = A(N, N_0, N_1)\ \frac{\Gamma(N_1 + k)\ \Gamma(N + N_0 - k)}{\Gamma(N - k + 1)\ \Gamma(k + 1)}. \tag{5}$$

where

$$A(N, N_0, N_1) = \frac{\Gamma(N + 1)\ \Gamma(N_0 + N_1)}{\Gamma(N + N_0 + N_1)\ \Gamma(N_1)\ \Gamma(N_0)}. \tag{6}$$

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 a real network system, when N_0 and N_1 are integer numbers, the Gamma functions can be replaced by factorials.

Notice that, because of the mutation rates (or frozen nodes), a particular realization of the dynamics will never stabilize in any state: the number of individuals with allele A_1 will always 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, no matter how large is the population.

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

$$m_0 = N\frac{N_1}{N_0 + N_1}. \tag{7}$$

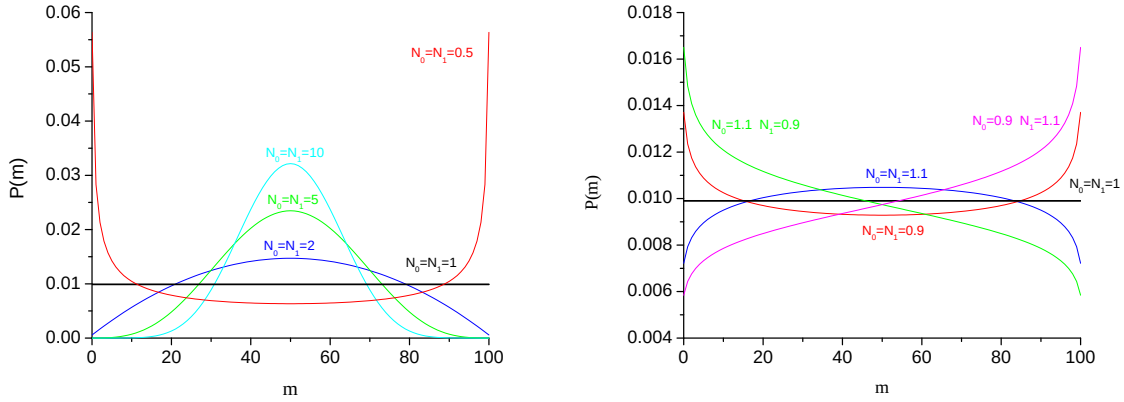

FIG. 1. Asymptotic probability distribution for a network with $N = 100$ nodes and several values of N_0 and N_1.

and

$$\sigma_2 = \frac{NN_1N_0(N_1 + N_0 + N)}{(N_1 + N_0)^2(1 + N_1 + N_0)} \tag{8}$$

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

Figures 1 and 2 show 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

$$\Delta = \left[\frac{NN_0N_1(N + N_0 + N_1)}{(N_0 + N_1)^3}\right]^{1/2}$$

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

$$\delta = \left[\frac{n_0n_1(1 + n_0 + n_1)}{N(n_0 + n_1)^3}\right]^{1/2}$$

$x_0 = m_0/N$ 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 $P_i \equiv P_{FC}$. For general networks an average value P_{av} can be calculated by replacing k_i by the average degree k_{av}. We can then define effective numbers of frozen nodes, N_{0ef} and N_{1ef}, as being the values of N_0 and N_1 in P_{FC} for which $P_{av} \equiv P_{FC}$. This leads to

$$N_{0ef} = fN_0, \qquad N_{1ef} = fN_1 \tag{9}$$

where $f = (N-1)/k_{av}$. Corrections involving higher moments can be obtained by integrating P_i with the degree distribution and expanding around k_{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 $k_{av} = 29.7$ and $f = 3.3$. The theoretical result was obtained with Eq. (5) with $N_{0ef} = N_{1ef} = 17$. For a scale-free network (panel (b)) grown from an initial cluster of 6 nodes adding nodes with 3 connections each following the preferential attachment rule [12], $f = 99/6$ and the effective values of N_0 and N_1 are approximately 82. Panel (c) shows the probability distribution for a 2-D regular lattice with 10×10 nodes connected to nearest neighbors for which $k_av = 3.6$ (the nodes near the border have less than 4 links) $f = 99/3.6 \approx 28$. Finally, panel (d) shows a small world version of the regular lattice [12], where 30 connections were randomly re-allocated, creating shortcuts between otherwise distant nodes. These results show that the approximate re-scaling 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.

VII. SPECIATION AND BIODIVERSITY

The connection between the influence dynamics on networks and the Moran model established in section IV and the approximate equilibrium distribution for structured networks obtained by re-scaling the number of frozen nodes described in section VI allow us to infer

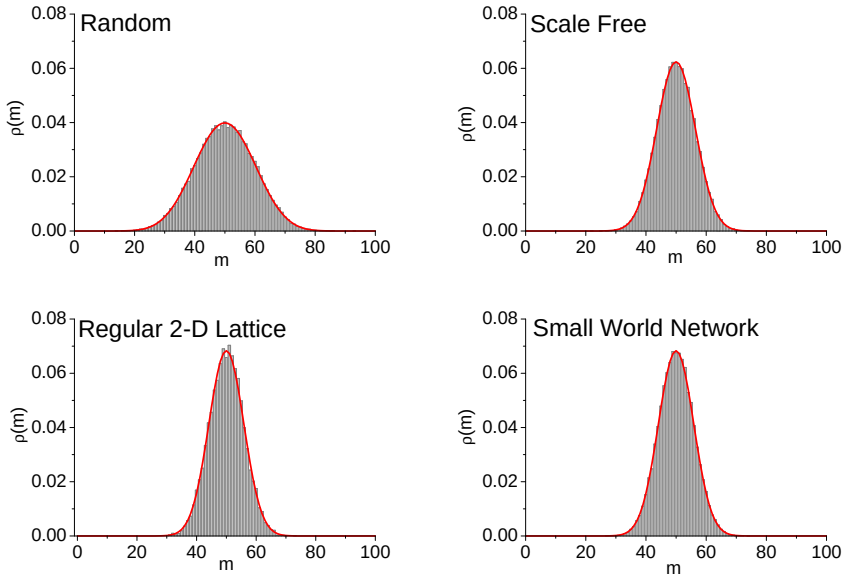

FIG. 2. Equilibrium probability distribution for networks with different topologies. In all cases $N = 100$, $N_0 = N_1 = 5$, $t = 10,000$, and the number of simulations is $50,000$. The theoretical (red) curve is drawn with effective numbers of frozen nodes $N_{0ef} = fN_0$ and $N_{1ef} = fN_1$: (a)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

important properties about the genetic evolution of spatially extended populations.

It has been recently shown [10, 13] that when mating is constrained by both spatial and genetic proximity between individuals, neutral evolution by drift alone might lead to speciation, i.e., to the spontaneous break up of the population into reproductively isolated clusters. Moreover, the patterns of abundance distributions generated by this mechanism are compatible with those observed in nature [10]. Neutral theories of biodiversity have become rather sophisticated [14], heating the neutralist-selectionist debate [15–19]. In this section we discuss the process of neutral speciation promoted by spatial and genetic constraints, termed *topopatric speciation*, in the light of the theory developed above.

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

$$N_z = \frac{2\mu(N-1)}{1-2\mu}. \tag{10}$$

The probability that two individuals picked at random are identical at equilibrium, P_{id},

is given by the sum of the probabilities that their alleles are are both A_1 or both A_0:

$$P_{id} = \sum_{m=0}^{N} \rho(m) \left[\frac{m}{N} \frac{m-1}{N-1} + \frac{N-m}{N} \frac{N-m-1}{N-1} \right]$$

$$= 1 + \frac{2}{N(N-1)} \left[\sigma^2 + \langle m \rangle^2 - N \langle m \rangle \right].$$

Using equations (7) and (8) we obtain

$$P_{id} = \frac{1 + N_z}{1 + 2N_z}. \tag{11}$$

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

$$P_{ht} = 1 - P_{id} = \frac{N_z}{1 + 2N_z}. \tag{12}$$

Consider now a population in equilibrium where the N individuals have B independent genes. The average genetic distance between two individuals is

$$\langle d \rangle = B P_{ht} = \frac{B N_z}{1 + 2N_z} = \frac{2B\mu(N-1)}{1 - 2\mu + 4\mu(N-1)} \approx \frac{B}{2} \left(\frac{4\mu N}{1 + 4\mu N} \right). \tag{13}$$

where the approximation holds for $\mu << 1$ and $N >> 1$, which is usually the case.

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. For given B and μ we may also calculate the size N_G that corresponds to a particular average genetic distance $\langle d \rangle = G$. From (13) we obtain

$$N_G = \frac{G}{2\mu(B - 2G)}. \tag{14}$$

This indicates that groups of N_G individuals, if reproductively isolated from the rest of the population, would cluster in genetic space so as to have average genetic distance G between each other. In large populations such groups do not form spontaneously unless mating is restricted by spatial proximity between individuals [10]. Spatial restriction in mating corresponds to influence processes on networks constructed over regular lattices.

Consider a square lattice with $N = L^2$ nodes and periodic boundary conditions where each node is connect only to neighbors which are within a distance S from itself (measured in units of lattice spacing). The area where an individual can look for a mate, its 'mating neighborhood' is approximately πS^2, which is also the average degree k_{av} of the network.

According to our discussion in section VI, this can be modeled as fully connected network with effective number of frozen nodes

$$N_{ef} = fN_z = \frac{N-1}{k_{av}} N_z \approx \frac{NN_z}{\pi S^2}. \tag{15}$$

The corresponding effective mutation rate is obtained from (10)

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

which gives

$$\mu_{ef} = \frac{f}{1+2\mu(f-1)}\mu \approx \frac{\mu f}{1+2\mu f}. \tag{16}$$

Note that $\mu_{ef} \to 1/2$ if $\mu f >> 1$.

In order to understand the dynamics of neutral speciation, which occurs if mating is restricted by spatial and genetic constraints, we consider the action of each of these limitations separately. We assume that the individuals are genetically identical at time zero and that diversity is introduced by mutations and sexual reproduction.

When mating between individuals is constrained by their spatial distance, as measured by the parameter S, the effective mutation rate (16) 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_{av} (corresponding to large values of N_z). The allele distribution approaches a broad symmetric distribution.

On the other hand, if mating is constrained only by the genetic distance between individuals, so that individuals whose genetic distance is larger than G are incompatible irrespective of their spatial position, the distribution of genetic distances shrinks. Indeed, if individuals start with identical genomes and are allowed to mate only with others that are genetically similar, the net effect is to reduce the size of the genome, changing the maximum distance between individuals from $B/2$ to G. The average genetic distance in such a genetically constrained population is obtained from eq.(13) as $\langle d \rangle = 4\mu NG/(1+4\mu N)$.

When both spatial and genetic restrictions are present, as in [10], 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 $\langle d \rangle \approx G$ 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 (14), which is usually much smaller than N. This also implies that the individuals are highly connected to each other, so that $f \approx 1$ and $\mu_e \approx \mu$, restoring the equilibrium of the system.

The conditions for speciation can be estimated as follows. For $f \approx N/k_{av} = N/\pi S^2$ we get

$$\mu_{ef} \approx \frac{N\mu}{\pi S^2 + 2N\mu}.$$

The genetic distribution becomes unstable if the average genetic distance becomes larger than about $2G$. Substituting $\langle d \rangle \to 2G$ and $\mu \to \mu_{ef}$ in eq. (13) we obtain

$$\frac{B}{2}\left(\frac{4\mu_{ef}N}{1+4\mu_{ef}N}\right) = \frac{B}{2}\left(\frac{4\mu N^2}{\pi S^2 + 2\mu N + 4\mu N^2}\right) \gtrsim 2G.$$

Solving for S gives

$$S \lesssim \sqrt{\frac{\mu N^2}{\pi}}\sqrt{\frac{B-4G}{G}}. \tag{17}$$

This estimate holds for small μN and large B, where the fluctuations around the averages values used here are small. It is compatible with the numerical simulations in [10], which shows that the critical line for speciation in the S–G plane indeed behaves like eq.(17). It also indicates that speciation is possible only if G is smaller than the natural average distance $\langle d \rangle$ given by equation (13) and if G is smaller than about $B/4$. In the limit $B \to \infty$, which is of little biological interest, this equation shows that speciation may occur even without spatial restrictions in mating, a result that has been demonstrated by simulations [20, 21].

ACKNOWLEDGMENTS

MAMA acknowledges financial support from CNPq and FAPESP.

[1] W.J. Ewens *Mathematical Population Genetics I. Theoretical Introduction* Series: Biomathematics, Vol. 9 (New York: Springer Verlag, 1979).

[2] C. Cannings *The latent roots of certain Markov chains arising in genetics; a new approach. 1. Haploid models.* Adv. Appl. Prob. 6, 260 (1974).

[3] P.A.P. Moran. *Random processes in genetics.* Proc. Cam. Phil. Soc. 54, 60 (1958).

[4] J. Wakeley *Coalescent theory* (Roberts & Company Publishers, 2009).

[5] G.A. Watterson. *Markov Chains with Absorbing States: A genetic example.* The Annals of Mathematical Statistics 32, 716 (1961).

[6] K. Gladstien. *The Characteristic Values and Vectors for a Class of Stochastic Matrices Arising in Genetics.* Siam J. Appl. Math 34, 630 (1978).

[7] J.H. Gillespie. *Population Genetics: A concise guide* (The Johns Hopkins University Press, 2004).

[8] D.D. Chinellato, M.A.M. de Aguiar, I.R. Epstein, D. Braha and Y. Bar-Yam2. *Dynamical Response of Networks under External Perturbations: Exact Results* arXiv:0705.4607v2 [nlin.SI]

[9] S. Wright, *Isolation by distance*, Genetics, **28**, 114 (1943).

[10] M.A.M. de Aguiar, M. Baranger, E.M. Baptestini, L. Kaufman, and Y. Bar-Yam. *Global Patterns of Specation and Diversity* Nature 460 **384** (2009).

[11] M.L. Rosenzweig, *Species diversity in space and time*, (Cambridge University Press, 1995)

[12] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).

[13] G.A. Hoelzer, R. Drewes, J. Meier and R. Doursat, *Isolation-by-distance and outbreeding depression are sufficient to drive parapatric speciation in the absence of environmental influences*, PLoS Comput. Biol. **4** e1000126 (2008).

[14] S.P. Hubbell, *The Unified Neutral Theory of Biodiversity and Biogeography*, (Princeton University Press, New Jersey, 2001).

[15] M. Kopp, *Speciation and the neutral theory of biodiversity: Modes of speciation affect patterns of biodiversity in neutral communities*, BioEssays **32** 564 (2010).

[16] H. Ter Steege, *How neutral is ecology?*, Biotropica **42** 631 (2010)

[17] S. Gavrilets, Li Hai and M.D. Vose, *Patterns of Parapatric Speciation*, Evolution **54**(2000).

[18] R.S. Etienne and B. Haegeman, *The neutral theory of biodiversity with random fission speciation*, Theor. Ecol. DOI 10.1007/s12080-010-0076-y.

[19] J.R. Banavar and A. Maritan, *Towards a theory of biodiversity*, Nature **460** 334 (2009).

[20] P.G. Higgs and B.Derrida, it Stochastic models for species formation in evolving populations, J. Phys. A. **24**, L985 (1991).

[21] P.G. Higgs and B.Derrida, it Genetic distance and species formation in evolving populations, J. Mol. Evol., **35**, 454 (1992).