doi:10.1111/j.1558-5646.2011.01528.x

TWO MEASURES OF EFFECTIVE POPULATION SIZE FOR GRAPHS

Mark Broom[1,2] **and Bernhard Voelkl**[3]

[1] *Centre for Mathematical Science, City University, London, EC1V 0HB, United Kingdom*

[2] *E-mail: Mark.Broom.1@city.ac.uk*

[3] *Institute for Theoretical Biology, Humboldt University at Berlin, D-10115 Berlin, Germany*

Received February 24, 2011
Accepted November 16, 2011

Effective population size is a key parameter in population ecology because it allows prediction of the dynamics of genetic variation and the rate of genetic drift and inbreeding. It is important for the definition of “nearly neutral” mutations and, hence, has consequences for the fixation or extinction probabilities of advantageous and deleterious mutations. As graph-based population models become increasingly popular for studying evolution in spatially or socially structured populations, a neutral theory for evolution on graphs is called for. Here, we derive formulae for two alternative measures of effective population size, the variance effective and inbreeding effective size of general unweighted and undirected graphs. We show how these two quantities relate to each other and we derive effective sizes for the complete graph the cycle and bipartite graphs. For one-dimensional lattices and small-world graphs, we estimate the inbreeding effective size using simulations. The presented method is suitable for any structured population of haploid individuals with overlapping generations.

KEY WORDS: Coalescent theory, genetic drift, inbreeding, population structure.

In population genetics, early models of evolutionary processes were based on simplifying assumptions of idealized well-mixed populations. To study the dispersal process in finite populations, Fisher (1930) and Wright (1931) introduced a simple model population that was later termed the “Wright–Fisher idealized population.” This population model assumes random mating among individuals, nonoverlapping generations, and a constant population size of N diploid individuals. Reproduction is a binomial sampling process from an infinitely large pool of gametes produced by the previous generation. Studying the dynamics of this population was a highly educative venture and indispensable for our understanding of evolutionary processes. Yet, as natural populations do not comply with these assumptions, a next generation of models, like the finite island model (Wright 1940) or the stepping stone model (Kimura 1953), incorporated aspects of population structure. Several extensions of these models were suggested that relaxed some of the original assumptions (e.g., Maruyama 1970a; 1970b; Slatkin 1977; Maruyama and Kimura 1980; Ewens 1989; Barton 1993; Chesser et al. 1993; Whitlock and Barton 1997; Whigham et al. 2008). To study the evolutionary process in spatially distributed populations, reaction–diffusion models were developed where space is represented explicitly as a continuous area (Nagylaki 1978). More recently, population ecologists have turned to a different approach and model populations as graphs, which allows one to study effectively any possible population structure within a single framework (Whigham et al. 2008).

Combadao et al. (2007) used simulations to study the speed of Mueller’s ratchet—the accumulation of deleterious mutations in nonrecombining genomes—for regular graphs, small-world graphs, and random graphs. Whigham and colleagues (2008) used a similar approach to study the loss of neutral variation in structured populations of different ploidy levels. As in these studies, vertices did not represent single individuals but demes, they were conceptually still similar to the original island and stepping stone models. In a different class of graph-based models, individuals were modeled explicitly as vertices. Early models of that type were simple gridlike models (Nowak and May 1992; Killingback and Doebeli 1996; Nakamaru et al. 1997; Hauert and Doebeli 2004) where the grid was thought to represent a spatially structured population. Lieberman et al. (2005) then argued that grids or lattices are only special cases of graphs and that one can gain much more insights by studying evolution on a more general class of graphs. This approach was primarily applied for studying the evolution of cooperation in structured populations (Ohtsuki and Nowak 2006; Santos et al. 2006; Grafen 2007; Lehmann et al. 2007; Taylor et al. 2007; Voelkl and Kasper 2009) and investigating important properties such as the fixation probabilities of advantageous mutations or their fixation time (Broom and Rychtar 2008; Broom et al. 2009). However, although literature in this area is accumulating rapidly, a neutral theory of evolution on graphs is still missing.

In finite populations, allele frequencies fluctuate from generation to generation due to random sampling from a finite pool of gametes. These changes drive the dispersal process, or genetic drift, which is suggested to be the dominating factor governing molecular evolution (Kimura 1983). The ultimate consequence of genetic drift in the absence of migration, mutation, and natural selection is the loss of genetic variation and, finally, the fixation of a single allele. The expected time for this to happen depends on the size of the population but also on the structure of the population. Sewall Wright (1931) introduced a quantity that he termed "effective population number," which he defined operationally as the size of an idealized Wright–Fisher population that would undergo the same amount of genetic change as the population in question. Depending on the genetic property one is interested in, the effective population size has to be calculated in different ways and these calculations can yield different results (Ewens 1979; Caballero 1994), though it was shown that for sufficiently large populations, these different effective sizes converge in the long-term limit (Whitlock and Barton 1997). We will restrict our discussion to two measures—the inbreeding effective size (Wright 1931) and the variance effective size (Crow 1954) for three reasons. First, these two measures have the longest history in classical population genetics (Kimura 1983). Second, they are still frequently used and have proven to be very useful—for example, the inbreeding effective size allowed the development of statistical methods for making inferences about evolutionary processes (Charlesworth et al. 2003). And finally, both measures are based on intuitive concepts—the estimated time to common ancestry and the sampling variance from one generation to the next.

The Wright–Fisher process assumes discrete, nonoverlapping generations, where all individuals reproduce at the same time and contribute to a common pool of gametes from which a sample is drawn as the next generation's individuals. Although such a process might be an appropriate model for annual reproducing plants or certain invertebrates, many other organisms do not exhibit synchronized reproduction. Moran (1962) therefore suggested a completely different model, where at each time-point, only a single individual is allowed to reproduce. This process is equivalent to the process in the voter model (Holley and Liggett 1975) in statistical mechanics (see also Blythe and McKane 2007; Sood et al. 2008; Pugliese and Castellano 2009).

In the following sections, we will derive the effective population size for the Moran reproduction scheme for the class of undirected and unweighted graphs. That is, the resulting formulae will not be restricted to well-mixed populations but will allow us to evaluate the effective size for any structured haploid population of constant size with overlapping generations.

Evolutionary Dynamics

To define the effective population size for graphs, we assume the following population dynamics introduced by Lieberman et al. (2005), called the Invasion Process. The population is represented as a finite, undirected, and connected graph $G = (V, E)$, where V is the set of vertices and E is the set of edges (see Table 1 for further definitions). Each vertex represents one individual of the population. Each individual can be either of type **A** or **a**, where these two types can be seen as two alleles of a single genetic locus. We treat the population dynamics as a discrete time Markov chain. In each unit of time, one individual is chosen at random to produce exactly one offspring of identical type (carrying the same allele). Immediately afterwards, one randomly chosen neighbor of the selected individual dies and is replaced by the offspring. This birth–death update process ensures that the population size remains constant. If we denote the set of all individuals of type **A** as set $C \subset V$, then in the next step type **A** individuals will either inhabit all vertices in

(i) a set $C \cup j$, $j \notin C$, if an individual $i \in C$ was chosen for reproduction and an individual j in the neighborhood of i was replaced by the offspring of i; or
(ii) a set $C \setminus i$, $i \in C$, if an individual $j \notin C$ was chosen for reproduction and an individual i in the neighborhood of j was replaced by the offspring of j; or
(iii) a set C, if an individual $i \in C$ was chosen for reproduction and the offspring of i replaces an individual $j \in C$ in the neighborhood of i, or if an individual $i \notin C$ was chosen for reproduction and the offspring of i replaces an individual $j \notin C$ in the neighborhood of i.

The states $C = V$ and $C = \emptyset$ are the absorbing states of the dynamics.

Table 1. Graph theoretic terms used in this article.

Term	Definition
Graph	A graph is an abstract representation of a set V of objects—called vertices—where pairs of objects can be connected by links—called edges.
Complete graph	A complete graph is a graph where every vertex is connected to any other vertex of the graph.
Density	The density of a graph is the ratio of the number of existing edges over the number of edges for a complete graph of the same size.
Degree	The degree k of a vertex is the number of edges incident to that vertex.
Regular graph	A regular graph is a graph where all vertices have the same degree.
Bipartite graph	A bipartite graph is a graph with two disjoint subsets of vertices A and B such that every edge connects a vertex in A to a vertex in B. A complete bipartite graph has edges between every vertex in A and every vertex in B.
Neighborhood	A vertex which is connected to a vertex i is a neighbor of i. The set of all vertices connected to vertex i is the neighborhood Γ_i of i.
Path	A path in a graph is a sequence of vertices such that there is an edge from each vertex to the next vertex in the sequence.
Cycle	A cycle is a graph that consists of a single path such that the initial vertex and the final vertex are the same. The number of vertices equals the number of edges and every vertex is of degree two.
Characteristic path length	The characteristic path length L is the average number of edges in the shortest path between all pairs of vertices in a graph.
Clustering coefficient	The clustering coefficient γ_i of vertex i measures the extent to which neighbors of i are connected between themselves. It is given by $\gamma_i = \lvert E(\Gamma_i)\rvert / \binom{k_i}{2}$, where $E(\Gamma_i)$ is the set of edges in the neighborhood of i. The clustering coefficient γ of a graph is γ_i averaged over all $i \in V$.
Small world	A small-world graph is a graph in which most vertices are not neighbors of one another but can be reached from every other vertex by a small number of steps. It is usually defined as a graph where the shortest path between two randomly chosen vertices grows proportionally to the logarithm of the number of vertices of the graph.
Rewiring	Rewiring is the process of replacing an edge from vertex i to vertex j by an edge from vertex i to any other vertex $k \in V \setminus \{i, j\}$ of the graph.

Measures of Effective Population Size

INBREEDING EFFECTIVE POPULATION SIZE

Wright (1931) defined the inbreeding effective population size as the size of an idealized Wright–Fisher population which would give rise to the same change in the rate of inbreeding as the actual population under consideration. The coefficient of inbreeding of diploid individuals, with respect to a specific locus, is the probability that the two alleles carried by an individual are identical by descent (Crow and Kimura 1970). Although it is questionable whether the term "inbreeding" still makes sense for haploid individuals, it is possible to calculate an inbreeding coefficient for a population of haploid individuals by evaluating the probability that two randomly chosen individuals are identical by descent. In the Moran model, the probability that two randomly chosen individuals are not identical by descent declines at a rate of $2/N^2$ per unit of time (Moran 1962). Because in this model a generation comprises N time units, the probability of nonidentity declines at a rate of $2/N$ per generation and, consequently, the inbreeding effective size for a well-mixed Moran population is $N/2$ with respect to the haploid Wright–Fisher population (Felsenstein 1971).

In the following, we shall denote: the time for the individuals at two randomly chosen nodes in a complete graph to be copies of a single original node (coalescence of two random lineages) as T_c; the time for complete coalescence in the population as $T_{c,pop}$; the time for coalescence to occur given that all individuals are a copy of original individual j as $T_{c,j}$; the probability that all individuals are a copy of individual j as P_j; the time of a randomly placed mutant to fix (given that it does) as T_f; and the time to fixation of individual j (given that it does) as $T_{f,j}$.

It should be noted that for most graphs (although not the complete graph) $T_{c,j}$ will depend upon j. Further, the terms $T_{c,j}$ and $T_{f,j}$ are really two alternative ways of writing the same thing. We can see that in general

$$E[T_{c,pop}] = \sum_{j=1}^{N} E[T_{c,j}]P_j = \sum_{j=1}^{N} E[T_{f,j}]P_j. \qquad (1)$$

On any regular graph, and in particular the complete graph, $P_j = 1/N$ for all j, so that

$$E[T_{c,pop}] = \frac{1}{N}\sum_{j=1}^{N} E[T_{f,j}] = E[T_f]. \tag{2}$$

Thus, T_f has the same expectation as the time to complete coalescence of the population for neutral mutants on a complete graph. The time to complete coalescence on the complete graph can be expressed in terms of the sum of the times to successively remove alleles from the population (Kingman 1982) giving

$$\begin{aligned} T_{c,pop} &= \sum_{i=2}^{N} \frac{2}{i(i-1)} T_c \\ &= 2T_c \left(\frac{1}{N-1} - \frac{1}{N} + \frac{1}{N-2} - \frac{1}{N-1} + \cdots \right. \\ &\quad \left. + \frac{1}{2} - \frac{1}{3} + 1 - \frac{1}{2} \right) \\ &= 2T_c \frac{N-1}{N}. \end{aligned}$$

Thus, we have

$$E[T_f] = E[T_{c,pop}] = 2E[T_c]\frac{N-1}{N}. \tag{3}$$

The inbreeding effective population size is defined as $N_{e,f}$, where $E[T_c] = NN_{e,f}$ (Felsenstein 1971). This corresponds to $N_{e,f}$ complete generations, because in the Moran process, every time step constitutes a single change only, so for a complete generational updating we need N steps. Thus, replacing $E[T_c]$ by its equivalent expression for complete graphs, we can find a general expression for $N_{e,f}$ for any graph:

$$N_{e,f} = \frac{E[T_c]}{N} = \frac{N}{2(N-1)} E[T_f] \frac{1}{N} = \frac{E[T_f]}{2(N-1)}. \tag{4}$$

In general, T_f can be found using the approach presented in Broom et al. (2010) or, for more complicated graphs, by simulation.

We note that we have assumed a dynamics which does not allow an offspring to replace its own parent. If we allowed this in a well-mixed population, the effect would be to slow the evolution of the process as at every time step the kind of transition that we have described occurs with probability $1 - 1/N$, so that every transition time has expectation $N/(N-1)$ times the original one. This is perhaps a fairer comparison to the original Wright–Fisher model, as this allowed replacement of a parent. Thus, we define an adjusted inbreeding effective population size $N^*_{e,f}$ by

$$N^*_{e,f} = \frac{N}{N-1} N_{e,f} = \frac{NE[T_f]}{2(N-1)^2}. \tag{5}$$

VARIANCE EFFECTIVE POPULATION SIZE

Following Crow (1954), we define the variance effective size of a population as the number of individuals that an idealized Wright–Fisher population would need to show the same amount of sampling variance in allele frequency as the population in question. For the Moran process, the variance effective size equals the inbreeding effective size $N_{e,v} = N_{e,f} = N/2$. We consider a population with two alleles, **A** and **a**, occurring at frequencies $1 - p$ and p, respectively. If the variance in the change of allele frequency in the graph per time unit is var(Δp), then according to Felsenstein (1971)

$$\mathrm{var}[\Delta p] \approx \frac{1}{NN_{e,v}} p(1-p) \Rightarrow N_{e,v} \approx \frac{1}{2N} \frac{2p(1-p)}{\mathrm{var}[\Delta p]}.$$

Defining T_v by

$$T_v = \frac{2p(1-p)}{\mathrm{var}[\Delta p]}$$

(var[Δp] is a change per unit of time, so correspondingly T_v is a measure of time), we consider a population as it changes from the introduction of a rare mutant to its fixation. We consider the expectation of T_v, $E[T_v]$, over this process, conditional on fixation occurring. We condition on fixation because we are primarily interested in the effect of effective population size on a successful mutant; most mutants do not fix, and when they do not they are usually eliminated early having little effect on the population. We set

$$N_{e,v} = \frac{E[T_v]}{2N}. \tag{6}$$

For the same reason as for the inbreeding effective population size, we also define an adjusted variance effective population size $N^*_{e,v}$ by

$$N^*_{e,v} = \frac{N}{N-1} N_{e,v} = \frac{E[T_v]}{2(N-1)}. \tag{7}$$

When considering the fixation of allele **a** recall that we start from a single mutant **a** and follow through the process until fixation, conditional on fixation occurring. Thus to find var[Δp] for any number i of alleles **a** (so $p = \frac{i}{N}$), we must condition on the fixation of **a**. We shall denote F_i as the event that **a** fixes from a population where there are i **a** alleles and $M_{i,j}$ is the event that the number of **a** alleles changes from i to j in a single time step (thus $M_{i,j}$ has zero probability for $j \neq i-1, i, i+1$). For any regular graph (and thus the complete graph) and neutral mutation,

$P[F_i] = \frac{i}{N}$, $E[\Delta p] = 0$, and

$$\begin{aligned} \text{var}[\Delta p] &= \frac{1}{N^2}(P[M_{i,i+1}|F_i] + P[M_{i,i-1}|F_i]) \\ &= \frac{1}{N^2}\left(\frac{P[M_{i,i+1}\cap F_i]}{P[F_i]} + \frac{P[M_{i,i-1}\cap F_i]}{P[F_i]}\right) \\ &= \frac{1}{N^2}\left(\frac{P[M_{i,i+1}]P[F_{i+1}]}{P[F_i]} + \frac{P[M_{i,i-1}]P[F_{i-1}]}{P[F_i]}\right) \\ &= \frac{1}{N^2}\frac{\frac{i+1}{N}P[M_{i,i+1}] + \frac{i-1}{N}P[M_{i,i-1}]}{i/N} \\ &= \frac{1}{N^2}(1-P[M_{i,i}]) = \frac{1}{N^2 d_i}, \end{aligned}$$

because $P[M_{i,i+1}] = P[M_{i,i-1}] = \frac{1}{2}(1 - P[M_{i,i}])$, defining d_i as the expected duration of a single stay at state i. Thus,

$$\begin{aligned} E[T_v] &= E\left[\frac{2p(1-p)}{\text{var}[\Delta p]}\right] = E\left[\frac{2\frac{i}{N}\frac{N-i}{N}}{\frac{1}{N^2 d_i}}\right] = E[2i(N-i)d_i] \\ &= \sum_{i=1}^{N-1} 2i(N-i)\,d_i\,P[I=i], \end{aligned}$$

where $P[I = i]$ is the proportion of time spent in state i. Defining m_i as the expected number of visits to state i, it is easy to see that

$$P[I=i] = \frac{m_i d_i}{\sum_{j=1}^{N-1} m_j d_j}.$$

Thus, we obtain

$$\begin{aligned} E[T_v] &= \sum_{i=1}^{N-1} 2i(N-i)\,d_i \frac{m_i d_i}{\sum_{j=1}^{N-1} m_j d_j} \\ &= \frac{2}{\sum_{j=1}^{N-1} m_j d_j}\sum_{i=1}^{N-1} i(N-i) m_i d_i^2. \end{aligned} \tag{8}$$

In the Appendix, we show that

$$m_i = \frac{2i(N-i)}{N}. \tag{9}$$

Using this equation, we can get the following expression for $E[T_v]$:

$$E[T_v] = \frac{N}{\sum_{i=1}^{N-1} m_i d_i}\sum_{i=1}^{N-1} m_i^2 d_i^2 \Rightarrow$$

$$E[T_v] = \frac{N}{\sum_{i=1}^{N-1} E[T_i]}\sum_{i=1}^{N-1} E[T_i^2], \tag{10}$$

where T_i is the total time spent in state i.

We note that this formula holds for regular graphs only. For irregular graphs, the probability of the next change in the mutant population size being an increase is dependent on the precise set of vertices occupied by the mutants, and so every such set needs to be considered separately and so finding an equivalent formula for irregular graphs would be a lot more complex.

A COMPARISON BETWEEN INBREEDING EFFECTIVE SIZE AND VARIANCE EFFECTIVE SIZE

From above, we can obtain the following alternative expression for the fixation time T_f,

$$T_f = \sum_{i=1}^{N-1} T_i. \tag{11}$$

Combining equations (10) and (11) gives

$$\begin{aligned} \frac{E[T_v]}{E[T_f]} &= \frac{N\sum_{i=1}^{N-1} E[T_i^2]}{\left(\sum_{i=1}^{N-1} E[T_i]\right)^2} = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} E[T_i^2]}{\left(\frac{1}{N-1}\sum_{i=1}^{N-1} E[T_i]\right)^2}\frac{N}{N-1} \\ &= \frac{NE[U^2]}{(N-1)(E[U])^2}, \end{aligned} \tag{12}$$

where U is a discrete uniform distribution on the values $E[T_1], \ldots, E[T_{N-1}]$, that is, a random variable with

$$P(U = E[T_i]) = 1/(N-1) \quad i = 1, \ldots, N-1.$$

Thus,

$$\frac{N_{e,v}}{N_{e,f}} = \frac{E[T_v]}{2N}\frac{2(N-1)}{E[T_f]} = \frac{N-1}{N}\frac{N}{N-1}\frac{E[U^2]}{(E[U])^2} = \frac{E[U^2]}{(E[U])^2}$$

and so

$$\frac{N_{e,v}}{N_{e,f}} - 1 = \frac{E[U^2] - (E[U])^2}{(E[U])^2} = \frac{\text{var}[U]}{(E[U])^2}. \tag{13}$$

Since clearly $\text{var}[U] \geq 0$, $N_{e,v} \geq N_{e,f}$ always for regular graphs. The fraction $\text{var}[U]/(E[U])^2$ can be rewritten as $\text{var}[U/E[U]]$, that is, the variance of U scaled by its mean $E[U]$ to make it dimensionless. The larger this quantity, the more the variance effective population size proportionately exceeds the inbreeding effective population size. The more variable $U/E[U]$ is, the more likely the population is to be observed at the larger values of U, that is, in the states where the population spends more time, and so slower transitions are more likely to be observed, with a correspondingly larger effective population size. This corresponds to sampling from a distribution either (1) completely at random or (2) proportional to the length of a random variable. The difference between these two methods is referred to as "length biased" sampling (Rao 1977), when the second is used in error as an estimate of the first. The more variable the distribution, the bigger the difference will be.

Effective Population Size for Some Example Graphs

COMPLETE GRAPHS

Complete graphs represent the case of a well-mixed population. By calculating the effective population size for a complete graph we should, therefore, retrieve the effective population size of a Moran population. To evaluate $E[T_i]$, we first have to evaluate the expected number of visits of each state, m_i and the expected duration of a visit for each state, d_i. For the latter, we first calculate the probability of leaving state i, $1 - P[M_{i,i}]$.

$$1 - P[M_{i,i}] = P[M_{i,i+1}] + P[M_{i,i-1}] = \frac{i(N-i)}{N(N-1)} + \frac{i(N-i)}{N(N-1)} = \frac{2i(N-i)}{N(N-1)} \Rightarrow$$

$$d_i = \frac{N(N-1)}{2i(N-i)} \Rightarrow$$

$$E[T_i] = m_i d_i = \frac{2i(N-i)}{N}\frac{N(N-1)}{2i(N-i)} = N-1.$$

Thus, $\mathrm{var}[U] = 0$ and so

$$N_{e,v} = N_{e,f} = \frac{E[T_f]}{2(N-1)} = \frac{1}{2(N-1)}\sum_{i=1}^{N-1} T_i = \frac{1}{2(N-1)}(N-1)(N-1) = \frac{N-1}{2}$$

and $N^*_{e,v} = N^*_{e,f} = \frac{N}{N-1}\frac{N-1}{2} = \frac{N}{2}$, which is equal to the result given by Felsenstein (1971).

CYCLES

Although for the population ecologist, the cycle might seem a highly arbitrary structure that has no correspondence in any natural population, it is an interesting graph for two reasons. First, it represents an extreme case, where the effects of spatial structure on evolutionary dynamics are strongest. Second, due to its simplicity it is possible to find analytic solutions for certain evolutionary scenarios. As for the complete graph, we first have to calculate $1 - P[M_{i,i}]$ to get $E[T_i]$.

$$1 - P[M_{i,i}] = P[M_{i,i+1}] + P[M_{i,i-1}] = \frac{1}{N} + \frac{1}{N} = \frac{2}{N} \Rightarrow$$

$$d_i = \frac{N}{2} \Rightarrow$$

$$E[T_i] = \frac{2i(N-i)}{N}\frac{N}{2} = i(N-i).$$

Thus, we find the inbreeding effective population size using

$$\begin{aligned} E[T_f] &= \sum_{i=1}^{N-1} E[T_i] = \sum_{i=1}^{N-1} i(N-i) = N\sum_{i=1}^{N-1} i - \sum_{i=1}^{N-1} i^2 \\ &= \frac{(N-1)N^2}{2} - \left(\frac{N}{6} - \frac{N^2}{2} + \frac{N^3}{3}\right) = \frac{N^3 - N}{6} \\ &= \frac{N(N^2-1)}{6}. \end{aligned}$$

This gives

$$N_{e,f} = \frac{1}{2(N-1)}\frac{N(N^2-1)}{6} = \frac{N(N+1)}{12}. \qquad (14)$$

To find the variance effective population size, we use equation (10) to give

$$E[T_v] = \frac{N}{\left(\frac{N(N^2-1)}{6}\right)}\sum_{i=1}^{N-1} i^2(N-i)^2 = \frac{6}{N^2-1}\sum_{i=1}^{N-1}(i(N-i))^2.$$

Alternatively

$$\begin{aligned} E[U] &= \sum_{i=1}^{N-1} E[T_i]P[U = T_i] = \sum_{i=1}^{N-1}\frac{1}{N-1}i(N-i) \\ &= \frac{1}{N-1}\left(N\sum_{i=1}^{N-1} i - \sum_{i=1}^{N-1} i^2\right) \\ &= \frac{1}{N-1}\left(N\frac{(N-1)N}{2} - \frac{(N-1)N(2N-1)}{6}\right) \\ &= \frac{N(N+1)}{6}. \end{aligned}$$

$$\begin{aligned} E[U^2] &= \sum_{i=1}^{N-1}\frac{1}{N-1}i^2(N-i)^2 \\ &= \frac{1}{N-1}\left(N^2\sum_{i=1}^{N-1} i^2 - 2N\sum_{i=1}^{N-1} i^3 + \sum_{i=1}^{N-1} i^4\right) \\ &= \left(\frac{N^3(2N-1)}{6} - \frac{N^3(N-1)}{2} + \frac{1}{30}(6N^4 - 9N^2 + N + 1)\right) \\ &= \frac{N}{30}(N^3 + N^2 + N + 1). \end{aligned}$$

Thus,

$$\begin{aligned} \mathrm{var}[U] &= \frac{N}{30}(N^3 + N^2 + N + 1) - \left(\frac{N(N+1)}{6}\right)^2 \\ &= \frac{(N+1)N(N-2)(N-3)}{180} \approx \frac{N^4}{180}. \end{aligned}$$

and

$$\frac{\mathrm{var}[U]}{(E[U])^2} \approx \frac{N^4/180}{N^4/36} = \frac{1}{5}.$$

Thus,

$$\frac{N_{e,v}}{N_{e,f}} \approx 1 + \frac{1}{5} = \frac{6}{5} \Rightarrow$$

$$N_{e,v} \approx \frac{6}{5}\frac{N(N+1)}{12} \approx \frac{N^2}{10}. \qquad (15)$$

We can see that the effective population size for a large cycle is large, being of the order of N^2. This is as we would expect as evolution is slow on the cycle compared to the well-mixed population of the complete graph, so happening at a rate that corresponds to a much larger well-mixed population.

BIPARTITE GRAPHS

In this section, we consider a type of irregular graph, the (complete) bipartite graph. In such a graph, the N vertices are split into two sets, A and B, with number of vertices n_a and n_b, respectively, where $N = n_a + n_b$. Every pair of vertices which includes one member from each set is connected by an edge, whereas no two members of the same set are connected. Although the bipartite graph has no direct biological correspondent, it is an interesting example of a heterogeneous graph that has received considerable attention in previous work (e.g., Sood et al. 2008; Blythe 2010). It is difficult to find an exact mathematical solution for the effective population size for this type of graph in general, but we can make good progress for the inbreeding effective population size using an approximation. Sood et al. (2008) show that for the voter model dynamics, provided that both n_a and n_b are sufficiently large, there is a two-stage process of evolution, where the fraction of mutants equalizes between the two sets quickly, and then evolution to fixation or elimination is a much slower process. By similar reasoning, this is also true for the dynamics that we consider (the Invasion Process, which was described in the Introduction).

Let us now assume that $n_b >> n_a$. The probability that the next individual selected to reproduce will be in set B (and so the next individual to be changed will be in set A) is n_b/N. Thus, change will happen much faster in A, but the fraction of mutants will tend to oscillate about the fraction of mutants in the slower evolving set B. Thus, time to fixation will be governed by the time for this to occur in B.

We can see from the Complete Graph section that the fixation time for the complete graph is $E[T_f] = (N-1)^2 \approx N^2$. Because the fraction of mutants is approximately equivalent in A and B, a good approximation for the number of steps which involve a B individual being selected for replacement prior to fixation is n_b^2. The proportion of all steps which are of this type is n_a/N, so that provided that n_a and n_b are sufficiently large, and that n_b is sufficiently larger than n_a, a good approximation for $E[T_f]$ is

$$E[T_f] \approx \frac{n_b^2}{n_a/N} = \frac{n_b^2 N}{N - n_b}$$

which yields

$$N_{e,f} \approx \frac{n_b^2}{2(N - n_b)}.$$

How large do n_a and n_b have to be, and how much larger must n_b be than n_a, for this approximation to be valid? We carried out simulations, and the approximation was good for both relatively small n_b and n_a, and n_b/n_a. For simulations involving 100 vertices, $n_a = 30$, $n_b = 70$ and $n_a = 10$, $n_b = 90$ yielded good agreement with the formula (although as expected for $n_a = n_b = 50$ the formula failed).

LATTICE GRAPHS

Complete graphs and cycles are both extreme cases with respect to the graph's density. Although the complete graph is obviously the graph with the highest density possible, the cycle has the lowest density for the class of regular graphs. As we have shown, these two graphs also differ markedly in their effective population sizes. We would therefore like to know the effective population size for regular graphs with intermediate density values. Here, we restrict our investigation to one-dimensional (1D) lattice graphs and to the inbreeding effective size. (Results for variance effective size are not shown for brevity but can be obtained equivalently.) We start with a 1D lattice graph, where each vertex has degree $k = 2$, which is a cycle. We then gradually increase the degree k of all vertices by connecting vertex i to all vertices $i+1, i+2, \ldots, i+k/2$ and $i-1, i-2, \ldots, i-k/2$ on the cycle. The density for these graphs is given as $D = k/(N-1)$. To evaluate the inbreeding effective population size for these graphs, we simulate neutral evolution to retrieve an estimate for the expected time T_f of a single mutant to fixate (given that it does). Using equation (3), we can then calculate an estimate for $N_{e,f}$. In addition to density, we want to relate the effective population size to the characteristic path length. Fig. 1 shows the results for graphs of $N = 100$. Although for the cycle (with $k = 2$ and a density of 0.02), $N_{e,f}^*$ is approximately 840, it rapidly declines with increasing density, approaching the effective size for the well-mixed population (i.e., the complete graph) already at a density level of about 0.3. The relation between inbreeding effective size and characteristic path length can be seen in Fig. 1B. As we would expect, the graphs with the largest characteristic path length have a higher effective population size, as long path lengths slow down the evolutionary process.

SMALL-WORLD GRAPHS

By altering the degree of the lattice, we are not only changing the density of the graph but also the characteristic path length and the degree of clustering. Thus, to investigate whether the path length or the degree of clustering affect the effective population size in the absence of density differences, we followed an approach proposed by Watts and Strogatz (1998). We start with a 1D lattice of $N = 100$ and $k = 4$. We then pseudo-randomly rewire (Table 1) a certain proportion r of the edges (the rewiring algorithm excludes self-loops, multiedges, and resulting disconnected graphs, Fig. 2A). For $r = 0$, we have a regular 1D lattice which is characterized by a large characteristic path length and a large clustering degree. The characteristic path length increases linearly with population size, whereas the clustering coefficient is independent of population size. By increasing r, we are introducing long-range links in the network which leads to a rapid decrease in the characteristic path length whereas the clustering coefficient remains initially nearly unchanged. Networks with such properties have been labeled "small-world networks" (Watts and Strogatz 1998).

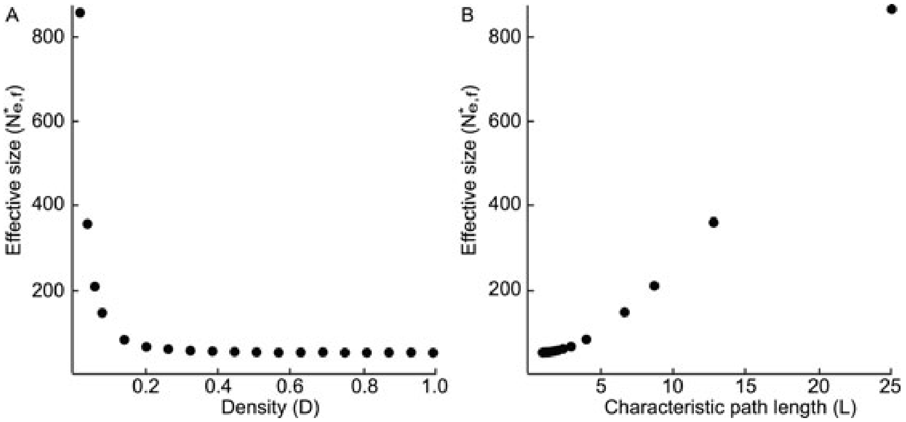

Figure 1. Inbreeding effective size for one-dimensional lattice graphs ($N = 100$, $k = 2, 4, 6, 8, 14, 20, \ldots, 98$) in relation to (A) density and (B) characteristic path length. Depicted values are mean estimates based on 10^6 simulations for each degree k.

Finally, by further increasing r, we get graphs that correspond in their properties to random graphs with a low clustering coefficient and a characteristic path length that is proportional to the logarithm of the population size.

For a population size of $N = 100$ and average degree of $k = 4$, the effective population size first drops rapidly with increasing r and then approaches a value that is roughly twice the effective size for the complete graph (Fig. 2B). However, already

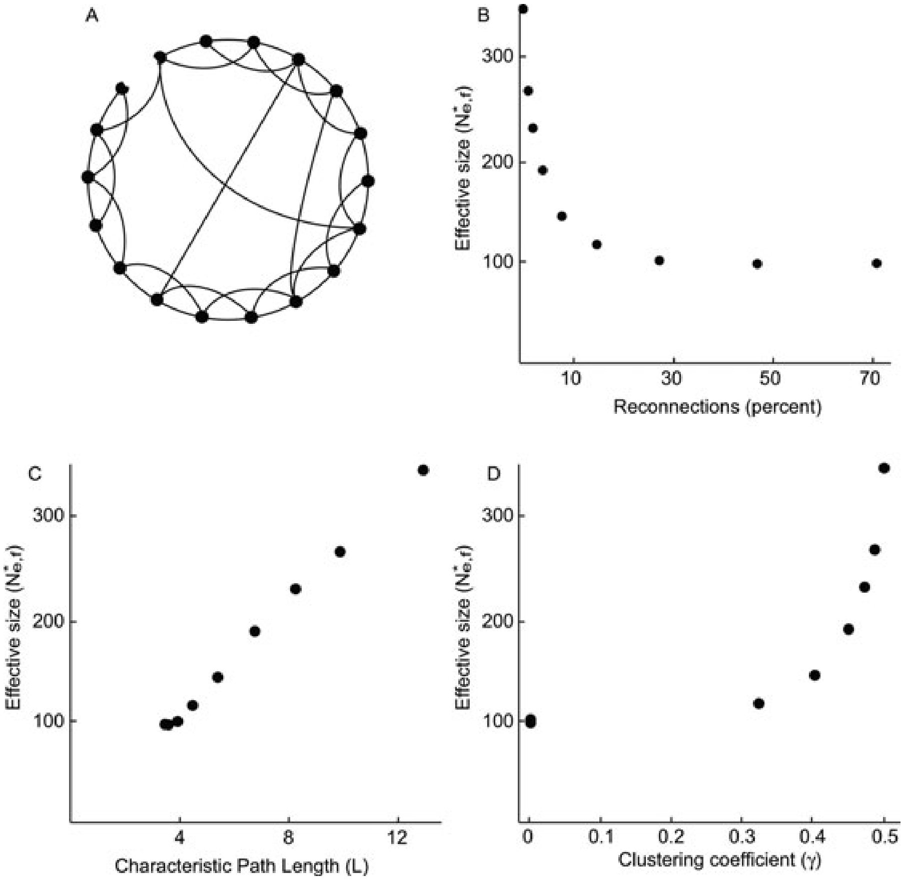

Figure 2. Inbreeding effective size for small-world graphs for different rates of re-wiring. (A) Example of a small-world graph based on a one-dimensional lattice ($N = 18$, $k = 4$) where three edges were pseudo-randomly reconnected. Mean estimates for the effective size for graphs with ($N = 100$, $k = 4$ and $0, 2, 4, 8, 16, \ldots, 256$) reconnections are given in relation to (B) percentage of reconnections, (C) characteristic path length, and (D) clustering coefficient.

for an average degree of $k = 6$, this asymptotic value is only 1.4 times the effective size for the complete graph. To illustrate the relation between the effective population size with the network properties of those graphs, we plot the inbreeding effective population size against characteristic path length (Fig. 2C) and clustering coefficient (Fig. 2D).

We again see that the effective population size increases with the characteristic path length, but it also increases with the value of the clustering coefficient. This is again as we would expect, as a high clustering coefficient means strong localized properties on the graph, which again slow evolution.

Conclusion

The importance of population structure for certain evolutionary processes was already recognized by Sewall Wright (1931) and intensively studied thereafter (for a review see Wang and Caballero 1999). In classical meta-population models, spatial structure was implemented by assuming that a population consists of many lines, local colonies, or demes, where reproduction within demes is random, whereas migration between those subunits is governed by a specific migration term. Coupled-map lattice models represent an extension where demes occupy the vertices of a graph (Lion and van Baalen 2008). For each deme, the local dynamics has a reaction term derived from a mean-field model and a coupling term that accounts for migration. For a further extension of the meta-population model, one can assume that each site can only be occupied by a discrete number of individuals (Metz and Gyllenberg 2001). Such models might add some more realism for cases with small deme size, but their main contribution is to bring in two important aspects: first, they add individuality to the model, and second, they represent a link to models derived from probabilistic cellular automata (Durrett and Levin 1994).

Here, we note that coupled map lattices and individual-based cellular automata are not fundamentally different, but the latter can be seen as a special case of the former with deme size fixed at one, while, at the same time, any population with larger deme size can—in principle—also be represented by an individual-based automaton. However, as long as the latter does not include some sort of dynamics mimicking migration, the population structure remains entirely static over all generations. This has important consequences, because in structured populations, genetic drift tends to produce random differences in allele frequencies between subpopulations, while allowing for small amounts of migration between them can counteract this effect. Mills and Allendorf (1996) emphasized that even "one migrant per generation" can be sufficient to prevent genetic differentiation between subpopulations. Killingback and Doebeli (1996) and Ifti et al. (2004) have studied the effects of dispersal in lattice models and came to similar conclusions. Thus, although excluding migration is a certain restriction which makes these models less suitable to represent real-life populations, individual-based models can still be a convenient approach to study some general principles (e.g., Lieberman et al. 2005; Ohtsuki et al. 2006; Grafen 2007; Lehmann et al. 2007; Taylor et al. 2007). Yet, a neutral theory for evolution on graphs is urgently needed, and this has been the focus of our paper.

In this article, we have developed a formula for the inbreeding effective population size for unweighted and undirected graphs in terms of their fixation time, which can be obtained explicitly in simple cases and by simulation for any graph. Thus, we have a ready means of finding this expression. The variance effective size is more complex. We have an approximation for this in terms of our expression T_v which can in general be simulated and we have an approximation formula for this for the case of regular graphs, which nevertheless requires simulation in the nonsimple cases. We see that variance effective population size is always at least as large as inbreeding population size for regular graphs. We have carried out some investigations, for example, graphs, including demonstrating consistency for the complete graph, and shown how effective population size depends upon the properties of other graphs (cycles, bipartite graphs, lattices, and small-world graphs). Regular graphs with a large number of edges in practice have very similar properties to the complete graph, and we can see this reflected in the effective population size of our examples. Similarly, small-world graphs with even moderate rewiring probability (roughly 10% or more) have substantially smaller effective population sizes than the original lattice structures. For bipartite graphs, the fixation time of a single mutant is primarily determined by the size of the larger subset of the graph.

Following previous work in this area, we took the Wright–Fisher population as the reference point for defining the effective population size. As a consequence, the effective size for the complete graph under a sequential Moran updating rule does not equal N but $N/2$. Thus, researchers who would prefer to relate their results directly to those for the complete graph must multiply the observed effective size by two. Although this is a slight inconvenience, it has saved us from defining effective size anew and introducing new terms; something we have tried to avoid given the already existing plethora of concepts and definitions. Population ecologists have been studying evolutionary processes in structured populations for a long time whereas, more recently, mathematicians and physicists have become interested in evolutionary processes on graphs. Here, we have shown how the concept of effective size that is well established within population ecology can be applied to evolutionary scenarios on graphs, and by this we have tried to build a bridge between these two lines of scientific inquiry which have been partly running in parallel.

ACKNOWLEDGMENTS

We thank R. Noë for initiating this work by inviting MB to Strasbourg. Financial support from the Austrian Science Funds (FWF J-2933) and the INCORE program (EC grant 043318).

LITERATURE CITED

Barton, N. H. 1993. The probability of fixation of a favoured allele in a subdivided population. Genet. Res. 62:149–158.

Blythe, R. A. 2010. Ordering in voter models on networks: exact reduction to a single-coordinate diffusion. J. Phys. A-Math Theor. 43:385003.

Blythe, R. A., and A. J. McKane. 2007. Stochastic models of evolution in genetics, ecology and linguistics. J. Stat. Mech-Theor. E P. P07018.

Broom, M., C. Hadjichrysanthou, and J. Rychtar. 2010. Evolutionary games on graphs and the speed of the evolutionary process. Proc. R. Soc. A 466:1327–1346.

Broom, M., and J. Rychtar. 2008. An analysis of the fixation probability of a mutant on special classes of non-directed graphs. Proc. R. Soc. A 464:2609–2627.

Broom, M., J. Rychtar, and B. Stadler. 2009. Evolutionary dynamics on small-order graphs. J. Interdiscipl. Math. 12:129–140.

Caballero, A. 1994. Developments in the prediction of effective population size. Heredity 73:657–679.

Charlesworth, B., D. Charlesworth, and N. H. Barton. 2003. The effects of genetic and geographic structure on neutral variation. Ann. Rev. Ecol. Evol. Syst. 34:99–125.

Chesser, R. K., O. E. Rhodes, D. W. Sugg, and A. Schnabel. 1993. Effective sizes for subdivided populations. Genetics 135:1221–1232.

Combadao, J., P. R. A. Campos, F. Dionisio, and I. Gordo. 2007. Small-world networks decrease the speed of Muller's ratchet. Genet. Res. 89:7–18.

Crow, J. F. 1954. Breeding structure of populations. ii. effective population number. Pp. 543–556, *in* O. Kempthorne, T. Bancroft, J. W. Gowen, and J. L. Lush, eds. Statistics and mathematics in biology. Iowa State College Press, Ames, IA.

Crow, J. F., and M. Kimura. 1970. An introduction to population genetics theory. Harper and Row, New York.

Durrett, R., and S. A. Levin. 1994. The importance of being discrete (and spatial). Theor. Popul. Biol. 46:363–394.

Ewens, W. J. 1979. Mathematical population genetics. Springer, Berlin, Germany.

———. 1989. The effective population size in the presence of catastrophes. Pp. 9–25 *in* M. Feldman, ed. Mathematical evolutionary theory. Princeton Univ. Press, Princeton, NJ.

Felsenstein, J. 1971. Inbreeding and variance effective numbers in populations with overlapping generations. Genetics 68:58–597.

Fisher, R. A. 1930. The genetical theory of natural selection. Clarendon Press, Oxford.

Grafen, A. 2007. An inclusive fitness analysis of altruism on a cyclical network. J. Evol. Biol. 20:2278–2283.

Hauert, C., and M. Doebeli. 2004. Spatial structure often inhibits the evolution of cooperation in the snowdrift game. Nature 428:643–646.

Holley, R. A., and T. M. Liggett. 1975. Ergodic theorems for weakly interacting infinite systems and the voter model. Ann. Probab. 3:643–663.

Ifti, M., T. Killingback, and M. Doebeli. 2004. Effects of neighbourhood size and connectivity on the spatial continuous prisoner's dilemma. J. Theor. Biol. 231:97–106.

Killingback, T., and M. Doebeli. 1996. Spatial evolutionary game theory: hawks and doves revisited. Proc. R. Soc. Lond. B 263:1135–1144.

Kimura, M. 1953. 'Stepping-stone' model of population. Annu. Rep. Natl. Inst. Genet., Japan 3:62–63.

———. 1983. The neutral theory of molecular evolution. Cambridge Univ. Press, Cambridge.

Kingman, J. F. C. 1982. On the genealogy of large populations. J. Appl. Probab. 19:27–43.

Lehmann, L., L. Keller, and D. J. T. Sumpter. 2007. The evolution of helping and harming on graphs: the return of the inclusive fitness effect. J. Evol. Biol. 20:2284–2295.

Lieberman, E., C. Hauert, and M. A. Nowak. 2005. Evolutionary dynamics on graphs. Nature 433:312–316.

Lion, S., and M. van Baalen. 2008. Self-structuring in spatial evolutionary ecology. Ecol. Lett. 11:277–295.

Maruyama, T. 1970a. Effective number of alleles in a subdivided population. Theor. Popul. Biol. 1:273–306.

———. 1970b. On the rate of decrease of heterozygosity in circular stepping stone models of populations. Theor. Popul. Biol. 1:101–119.

Maruyama, T., and M. Kimura. 1980. Genetic variability and effective population size when local extinction and recolonization of subpopulations are frequent. Proc. Natl. Acad. Sci. USA 77:6710–6714.

Metz, J. A. J., and M. Gyllenberg. 2001. How should we define fitness in structured metapopulation models? Including an application to the calculation of evolutionary stable dispersal strategies. Proc. R. Soc. Lond. B 268:499–508.

Mills, L. S., and F. W. Allendorf. 1996. The one-migrant-per-generation rule in conservation and management. Cons. Biol. 10:1509–1518.

Moran, P. A. P. 1962. The statistical processes of natural selection. Clarendon Press, Oxford.

Nagylaki, T. 1978. Random genetic drift in a cline. Proc. Natl. Acad. Sci. USA 75:423–426.

Nakamaru, M., H. Matsuda, and Y. Iwasa. 1997. The evolution of cooperation in a lattice-structured population. J. Theor. Biol. 184:65–81.

Nowak, M. A., and R. M. May. 1992. Evolutionary games and spatial chaos. Nature 359:826–829.

Ohtsuki, H., C. Hauert, E. Lieberman, and M. A. Nowak. 2006. A simple rule for the evolution of cooperation on graphs and social networks. Nature 441:502–505.

Ohtsuki, H., and M. A. Nowak. 2006. Evolutionary games on cycles. Proc. Biol. Sci. 273:2249–2256.

Pugliese, E., and C. Castellano. 2009. Heterogeneous pair approximation for voter models on networks. Europhys. Lett. 88:58004.

Rao, C. R. 1977. A natural example of weighted distribution. Amer. Stat. 31:24–26.

Santos, F. C., J. M. Pacheco, and T. Lenaerts. 2006. Evolutionary dynamics of social dilemmas in structured heterogeneous populations. Proc. Natl. Acad. Sci. USA 103:3490–3494.

Slatkin, M. 1977. Gene flow and genetic drift in a species subject to frequent local extinctions. Theor. Popul. Biol. 12:253–262.

Sood, V., T. Antal, and S. Redner. 2008. Voter models on heterogeneous networks. Phys. Rev. E P. 041121.

Taylor, P. D., T. Day, and G. Wild. 2007. Evolution of cooperation in a finite homogeneous graph. Nature 447:469–472.

Voelkl, B., and C. Kasper. 2009. Social structure of primate interaction networks facilitates the emergence of cooperation. Biol. Lett. 5:462–464.

Wang, J., and A. Caballero. 1999. Developments in predicting the effective size of subdivided populations. Heredity 82:212–226.

Watts, D. J., and S. H. Strogatz. 1998. Collective dynamics of 'small-world' networks. Nature 393:440–442.

Whigham, P. A., G. C. Dick, and H. G. Spencer. 2008. Genetic drift on networks: ploidy and the time to fixation. Theor. Popul. Biol. 74:283–290.

Whitlock, M. C., and N. H. Barton. 1997. The effective size of a subdivided population. Genetics 146:427–441.

Wright, S. 1931. Evolution in mendelian populations. Genetics 16:97–259.

———. 1940. Breeding structure of populations in relation to speciation. Am. Nat. 74:232–248.

Associate Editor: Nick Barton

Appendix

Here, we show that the expected number of visits to the state i on a regular graph, m_i, is given by equation (9).

For neutral mutations and regular graphs, the next change in the mutant population size is as equally likely to be an increase as a decrease, whatever the precise set of vertices occupied by the mutants, and we effectively have a random walk with probability $p = 1/2$ of moving forward. An increase in the population size is just a move to the right in the random walk, and a decrease in the population size is a move to the left. Conditioning on reaching fixation, we must first reach i for the first time, thus $m_i = m_i^* + 1$, where m_i^* is the number of returns to state i. If we go to the left from $i - 1$, we must come back, if we go to the right, we may or may not come back. In general, the population either reaches 0 or N and then stops, as fixation or elimination of the mutant has occurred.

We get the following equation involving m_i^*

$$\begin{aligned} m_i^* &= P[A_L](m_i^*+1) + (1-P[A_L])[(m_i^*+1)P[E_R] \\ &\quad + 0(1-P[E_R])] \\ &= (m_i^*+1)(P[A_L] + (1-P[A_L])P[E_R]), \end{aligned} \tag{A1}$$

where A_L is the event that the next move is to the left given that the population subsequently reaches N, and E_R is the event that, starting from $i+1$, the population reaches i given that it reaches N.

We must consider the following events:

A_1: the next move is to the left and then the population reaches 0;

A_2: the next move is to the left and then the population reaches N;

A_3: the next move is to the right and then the population reaches 0;

A_4: the next move is to the right and then the population reaches N;

E_1: starting from $i+1$, the population reaches N before i;

E_2: starting from $i+1$, the population reaches i then N;

E_3: starting from $i+1$, the population reaches i then 0.

It is clear that $P[A_1 \cup A_2 \cup A_3 \cup A_4] = 1$, $P[A_i \cap A_j] = 0$, $P[A_1] = \frac{1}{2}(1 - P[F_{i-1}]) = \frac{1}{2}\frac{N-i+1}{N}$, $P[A_2] = \frac{1}{2}P[F_{i-1}] = \frac{1}{2}\frac{i-1}{N}$,

$P[A_3] = \frac{1}{2}(1 - P[F_{i+1}]) = \frac{1}{2}\frac{N-i-1}{N}$, $P[A_4] = \frac{1}{2}P[F_{i+1}] = \frac{1}{2}\frac{i+1}{N}$.

Similarly, it is clear that $P[E_1 \cup E_2 \cup E_3] = 1$, $P[E_i \cap E_j] = 0$, $P[E_1] = \frac{1}{N-i}$, $P[E_2] = \frac{N-i-1}{N-i}\frac{i}{N}$, and $P[E_3] = \frac{N-i-1}{N-i}\frac{N-i}{N}$.

$$\begin{aligned} P[E_R] &= P[E_2 \mid E_1 \cup E_2] = \frac{P[E_2]}{P[E_1] + P[E_2]} \\ &= \frac{\frac{N-i-1}{N-i}\frac{i}{N}}{\frac{1}{N-i} + \frac{N-i-1}{N-i}\frac{i}{N}} = \frac{(N-i-1)i}{(N-i-1)i + N}. \end{aligned} \tag{A2}$$

$$\begin{aligned} P[A_L] &= P[A_2 \mid A_2 \cup A_4] = \frac{P[A_2]}{P[A_2] + P[A_4]} \\ &= \frac{\frac{1}{2}\frac{i-1}{N}}{\frac{1}{2}\frac{i-1}{N} + \frac{1}{2}\frac{i+1}{N}} = \frac{i-1}{2i}. \end{aligned} \tag{A3}$$

Thus, $m_i^* = (m_i^* + 1)\left(\frac{i-1}{2i} + \frac{i+1}{2i}P[E_R]\right)$

$$\Rightarrow m_i^*(2i - (i-1) - (i+1)P[E_R]) = (i-1) + (i+1)P[E_R]$$

$$\begin{aligned} \Rightarrow m_i^* &= \frac{i - 1 + (i+1)P[E_R]}{(i+1)(1 - P[E_R])} = \frac{i - 1 + \frac{(i+1)(N-i-1)i}{(N-i-1)i+N}}{(i+1)\frac{N}{(N-i-1)i+N}} \\ &= \frac{(i-1)(N-i-1)i + N(i-1) + (i+1)i(N-i-1)}{(i+1)N}. \end{aligned}$$

Thus,

$$\begin{aligned} m_i = 1 + m_i^* &= \frac{2i}{i+1} + \frac{2i^2(N-i-1)}{(i+1)N} \\ &= \frac{2i}{i+1}\left(1 + \frac{i(N-i-1)}{N}\right) \\ &= \frac{2i}{i+1}\left(1 + i - \frac{i(i+1)}{N}\right) = \frac{2i(N-i)}{N}. \end{aligned} \tag{A4}$$

This is equation (9) as required.