J. theor. Biol. (2000) **203,** 303–315
doi:10.1006/jtbi.2000.1095, available online at http://www.idealibrary.com on IDEAL®

On the Genealogy of a Population of Biparental Individuals

BERNARD DERRIDA*, SUSANNA C. MANRUBIA† AND DAMIÁN H. ZANETTE‡

**Laboratoire de Physique Statistique de l'École Normale Supérieure*, 24 *rue Lhomond*, *F*-75231 *Paris* 05 *Cedex*, *France*, †*Fritz-Haber-Institut der Max-Planck-Gesellschaft*, *Faradayweg* 4-6, 14195 *Berlin*, *Germany and* ‡*Consejo Nacional de Investigaciones Científicas y Técnicas*, *Centro Atómico Bariloche e Instituto Balseiro*, 8400 *S.C. de Bariloche*, *Río Negro*, *Argentina*

(*Received on* 28 *July* 1999, *Accepted in revised form on* 18 *January* 2000)

If one goes backward in time, the number of ancestors of an individual doubles at each generation. This exponential growth very quickly exceeds the population size, when this size is finite. As a consequence, the ancestors of a given individual cannot be all different and most remote ancestors are repeated many times in any genealogical tree. The statistical properties of these repetitions in genealogical trees of individuals for a panmictic closed population of constant size N can be calculated. We show that the distribution of the repetitions of ancestors reaches a stationary shape after a small number $G_c \propto \log N$ of generations in the past, that only about 80% of the ancestral population belongs to the tree (due to coalescence of branches), and that two trees for individuals in the same population become identical after G_c generations have elapsed. Our analysis is easy to extend to the case of exponentially growing population.

1. Introduction

In the case of sexual reproduction, the ancestry of an individual is formed by two parents, four grandparents two generations ago, and in general 2^G individuals G generations back into the past. The explosive growth of the number of ancestors belonging to the genealogical tree of, say, a present human should stop at some point due, at least, to the finite size of previous populations. For instance, only $G \simeq 33$ generations ago (spanning a period of less than 1000 years), the number of potential ancestors in the tree of any of us is about 8.5×10^9, more than the present population of the Earth, and of course much larger than the population living about the year 1000. The answer to this apparent paradox is simple: the branches of a typical genealogical tree often coalesce, indicating that many of the ancestors were in fact relatives and appear repeatedly in the tree (Ohno, 1996; Derrida *et al.*, 1999; Gouyon, 1999). It might be difficult to test the statistical properties of such repetitions for an actual large, randomly mating population. Nevertheless, some exceptions can be found in royal genealogy. Since nobles usually married within their own castes, the presence of repeated ancestors in royal genealogical trees is far from rare. The example of the English king Edward III, where some ancestors appear up to six times, has been analysed in our previous work (Derrida *et al.*, 1999).§

Much attention has been paid in the past to a related problem, namely the statistical properties of branching processes (Harris, 1963) and its applications to the characteristics of the successive descendants of a single ancestor

---

§We used the tree of Edward III which can be found at ⟨http://uts.cc.utexas.edu/~churchh/edw3chrt.html⟩.

(Kingman, 1993). Actually, first applications of the branching processes technique go back to the 1920s. Haldane (1927) calculated the probability that a mutant allele be fixed in a population through a method developed previously by Fisher (1922). There, the relevant quantity was the survival probability of the descendants of the first individual carrying the mutation. All these studies apply to the vertical transmission of names, to the inheritance of characters coming only from one of the parents, like mitochondrial DNA or the Y chromosome, or to the fate of a mutant gene, for example, and correspond to an effective monoparental population. The heart of our problem is to take into consideration the fact that reproduction is biparental. The distribution of repetitions of ancestors described below does however satisfy an equation similar to those which appear in branching processes (Harris, 1963).

Our problem of repetitions of ancestors in genealogical trees is much closer to the counting of the descendants of an individual in a sexual population. For example, in the case of a population of constant size, the average number of offspring is *two* per couple. Therefore, after G generations each individual has, on average, 2^G descendants. What prevents the number of descendants from growing exponentially with G and to exceed the population size is interbreeding: when 2^G becomes comparable to the population size, interbreeding happens between the descendants and different lines of descent coalesce. The problem of the statistical properties of these coalescences is very similar to our present study of genealogical trees. None of them has, to our knowledge, yet been analysed.

In the present work, we study theoretically the problem of repetitions in the genealogical trees in the case of a closed, panmictic population. The study of the properties of a single tree with coalescent branches and the comparison of the genealogical trees of two contemporary individuals allows us to show that

1. There is a finite fraction (about 20% for a population of constant size N) of the initial population whose descendants become extinct after a number of generations $G_c \propto \log N$. All the rest of the initial population (about 80%) belongs to all genealogical trees.
2. The distribution of the repetitions of ancestors living more than G generations ago reaches a stationary shape after about G_c generations.
3. The genealogical trees of two individuals in the same population become identical after a small number of generations G_c back into the past.
4. The similarity between two genealogical trees changes from 1% (almost all ancestors in the two trees are different) to 99% (the repetitions of the ancestors in the two trees are almost identical) within 14 generations around G_c, independent of the population size N.

Our work can be generalized (see Section 4) to describe coalescent processes, understood as the study of the gene tree originated when looking for the ancestry of a random sample of sequences (Kingman, 1982; Hudson, 1991; Donnelly & Tavaré, 1995). In the absence of recombination, each sequence has a single ancestor. The topology of thus reconstructed trees is equivalent to that generated through branching processes. Next in complexity, one can consider a two-locus sequence and assume that recombination can occur only between the two loci and with a small probability (meaning correlated genealogies[‖] for the two loci). The statistical properties of such a process can be estimated until the most recent common ancestor (MRCA) is reached (Hudson, 1991). Instead, if one faces the study of a chromosome (Wiuf & Hein, 1997; Derrida & Jung-Muller, 1999) or of the whole genome, the number of ancestors grows as one proceeds back in time, since each individual has two parents and, apart from coalescence, also recombination (meaning splitting of the branches in the tree) is frequent.

[‖] In this paper, we use the term *genealogy* to refer to the ancestry of a single gene or of a whole set of sequences. In all cases, the *genealogy* is the complete set of ancestors contributing to the present object, this object being an individual (as in Section 2), a group of individuals (as in Section 3), a sequence (Section 4), or a single locus (as quoted here). In this case, *correlated genealogies* simply means that the different sets of ancestors for the two loci are not independent.

If one considers a population or a sample of individuals within a population, there are relevant differences between the genealogy of a single gene and the genealogy of a chromosome or of the whole genome (which we study here). While in the first case, in fact, there exists an MRCA for the sample (where the gene tree ends), the genealogical tree of a chromosome or of the genome with two parents proceeds backwards in time and never reduces to a single ancestor. The genealogical tree representing the pedigree of a diploid organism contains a large fraction of the ancestral population. In this case, one may then talk about the *most recent common set of ancestors*, and study the similarities among different individuals now within the same population.

2. Statistical Properties of an Individual Tree

Here we consider a simple neutral model of a closed population evolving under sexual reproduction and with non-overlapping generations.¶ If the population size is $N(G)$ at generation G in the past, we form couples at random (by randomly choosing $N(G)/2$ pairs of individuals) and assign each couple a random number k of descendants. The probability p_k of the number k of offspring is given and if the population size is N at present, its size $N(G)$ at generation G in the past is given by

$$N(G) = \left(\frac{2}{m}\right)^G N, \tag{1}$$

where the factor m is obtained from

$$m = \sum_k k\, p_k. \tag{2}$$

For $m = 2$, the population size remains constant in time, whereas for $m \neq 2$ the number of individuals in the next generation is multiplied by a factor $m/2$. After a number of generations the tree of each of the individuals in the youngest generation is reconstructed. To quantify the contribution of each of the ancestors to the genealogical tree of an individual, we define the *weight* $w_\gamma^{(\alpha)}(G)$ of an ancestor γ in the tree of individual α at generation G in the past as

$$w_\gamma^{(\alpha)}(G+1) = \frac{1}{2} \sum_{\gamma' \text{ children of } \gamma} w_{\gamma'}^{(\alpha)}(G). \tag{3}$$

We take $w_\gamma^{(\alpha)}(0) = \delta_{\alpha,\gamma}$, as this ensures that at generation $G = 0$ all the weight is carried by the individual α itself. The factor 1/2 in eqn (3) keeps the sum of the weights normalized $\sum_{\gamma=1}^{N(G)} w_\gamma^{(\alpha)}(G) = 1$, for any past generation G. The weight $w_\gamma^{(\alpha)}(G)$ can be thought of as the probability of reaching ancestor γ if one climbs up the reconstructed genealogical tree of individual α by choosing at each generation one of the two parents at random. The weights essentially measure the repetitions (see Fig. 1) in the genealogical tree. Without repetitions, $w_\gamma^{(\alpha)}(G)$ would simply be 2^{-G} for each ancestor γ in the tree.

As an illustration of the previous quantities, we represent in Fig. 1 the result of random matings inside a small closed population of constant size $N = 14$ (thus $m = 2$) during 7 generations. The lines link progenitors with their offspring. The grey scale gives the weight $w_\gamma(G)$ of each of the individuals in the tree. The numbers on the left, all of them of the form $r/2^G$, give the weight of the leftmost individual in each generation. The denominators simply indicate the potential maximum number of ancestors at each generation. As counted by the numerator, each of them would appear repeated r times in this tree if all the branches were explicitly shown.

We further assume that the probability p_k of having k children per couple follows a Poisson distribution, $p_k = m^k e^{-m}/k!$ (most of what follows could be easily extended to other choices of p_k). We represent in Fig. 2 the probability for an English couple to have k marrying sons during the period 1350–1986 (Dewdney, 1986). The solid line corresponds to a Poisson distribution with average 1.15 (i.e. the average number of offspring *per individual* in that period, which corresponds to $m = 2.3$ in our analysis), and implies that the total population is growing. These data spanning

---

¶The Wright–Fisher model for allele frequencies works in the same set of hypotheses (Wright, 1931; Fisher, 1930). More recently, Serva & Peliti (1991) obtained a number of statistical results for the genetic distance between individuals in a sexual population evolving in the absence of natural selection.

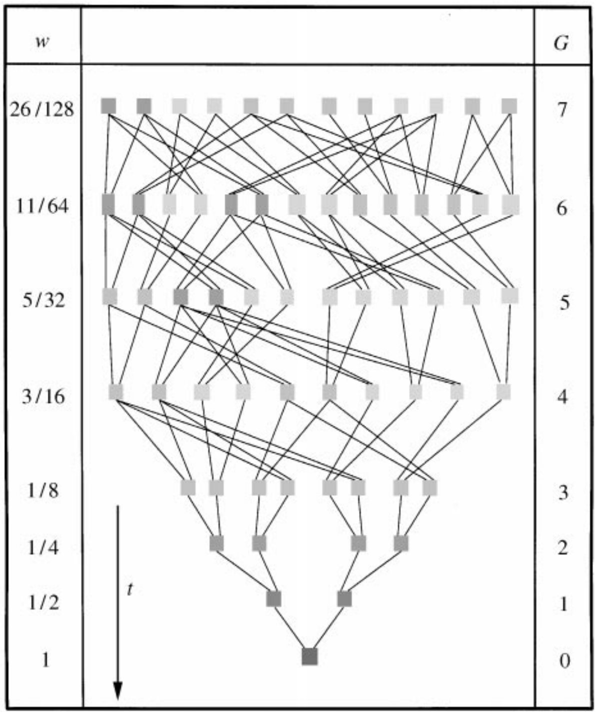

FIG. 1. Coalescence of branches in a genealogical tree. We display the reconstructed ancestry of a present individual in a small population of constant size $N = 14$. Numbers on the left-hand side stand for the *weight* w of the leftmost individual at each generation. The grey scale changes from light grey (small w) to dark grey (large w) proportionally to the logarithm of the weight. The exact values are calculated according to eqn (3). The weight is a measure proportional to the number of times that an ancestor appears in a tree, or, equivalently, to the number of branches which have coalesced up to that point.

six centuries and taken over an homogeneous population support the hypothesis that the number of offspring is indeed Poisson distributed.**

If we define $S^{(\alpha)}(G)$, the fraction of the population (at a generation G in the past) which does not belong to the genealogical tree of individual α (i.e. such that $w_\gamma^{(\alpha)}(G) = 0$) one can show (see the appendix) that

$$S^{(\alpha)}(G+1) = \exp[-m + mS^{(\alpha)}(G)]. \tag{4}$$

This recursion, together with the initial condition $S^{(\alpha)}(0) = 1 - 1/N$, determines this quantity for any G (Derrida *et al.*, 1999).

For large G and for any individual α, this fraction $S^{(\alpha)}(G)$ converges to the fixed point $S(\infty)$ of eqn (4). This gives for $m = 2$ (i.e. for a population of constant size) a fraction $S(\infty) \simeq 0.2031878\ldots$ which becomes extinct, so that the remaining fraction $1 - S(\infty) \simeq 80\%$ of the population belongs to the genealogical tree *of any individual* α. A similar calculation shows that this 80% of the population which is not extinct after a large number of generations appears in the genealogical trees of all individuals: if $S^{(\alpha,\beta)}(G)$ is the fraction of the population which does not belong to any of the two trees of two distinct individuals α and β, $S^{(\alpha,\beta)}(G)$ satisfies the same recursion (4) as $S^{(\alpha)}(G)$, and converges to the same fixed value $S(\infty)$. Thus, within this neutral model, an individual either becomes extinct (with a probability of 20%) or becomes an ancestor of the whole population after a large number of generations (with a probability of 80%). For an exponentially growing population with $m = 2.3$ as in Fig. 2, the results are the same except for the precise value of $S(\infty)$ (for $m = 2.3$, one finds that $S(\infty) \simeq 14\%$).

When G is large enough, as shown in the appendix, the whole distribution $P(w)$ of the weights $w_\gamma^{(\alpha)}(G)$ reaches a stationary shape, the properties of which can be calculated (Derrida *et al.*, 1999). We show in Fig. 3 the distribution $P(w/\langle w\rangle)$ for different values of m. As can be seen, it has a power-law dependence, $P(w) \propto w^{\xi}$ for

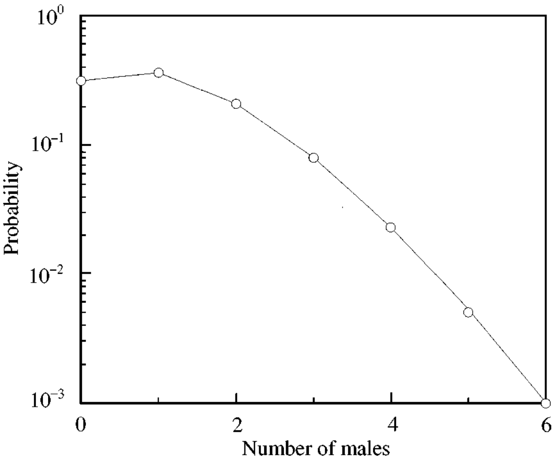

FIG. 2. Probability for an English couple to have k marrying sons during the period 1350–1986 (○). The (——) corresponds to a Poisson distribution of average 1.15 (data from Dewdney, 1986).

** Nonetheless, deviations from this distribution induced by a social transmission of the reproductive behaviour have been reported (Austerlitz & Heyer, 1998).

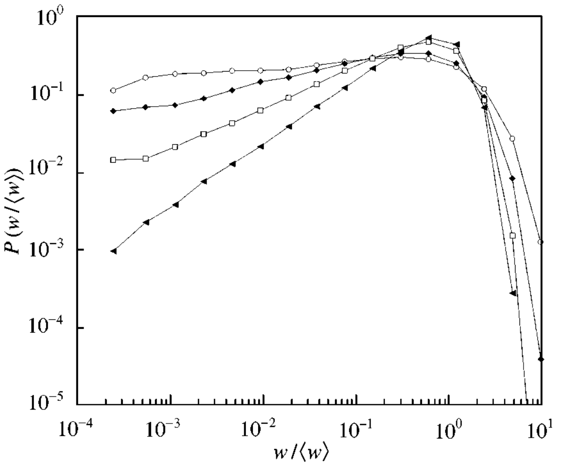

FIG. 3. Stationary shape of the distribution $P(w/\langle w\rangle)$ for different values of m. We compare the constant population case ($m = 2$) with shrinking ($m = 1.5$), and expanding ($m = 3$, 4) populations. Parameters are $N = 4096$, $G = 20$, and averages over 10^3 independent realizations have been performed: $m = 1.5$ (○——○); $m = 2$ (◆——◆); $m = 3$ (□——□); $m = 4$ (◀——◀).

small values of the ratio $w/\langle w\rangle$, with an exponent given by

$$\xi = -\frac{\log S(\infty)}{\log m} - 2, \tag{5}$$

and achieves a maximum value for $w/\langle w\rangle \simeq 1$.

3. Similarity between Two Trees

We would like to know how similar are the genealogical trees of two contemporary individuals and how they evolve in time within the same population. We have seen that a large fraction ($1 - S(\infty) \simeq 80\%$) of the ancestral population constitutes the pedigree of every present individual. As a next step, one can compare two individuals and compute the degree of similarity between their trees, that is, the set of ancestors appearing at each generation in both trees simultaneously. We will see in particular that the two trees become identical after a number G_c of generations.

We start with the definition of the *overlap* between the genealogical trees of two different individuals, α and β. Let $w_\gamma^{(\alpha)}(G)$ be the weight of the ancestor γ in the tree of α at generation G in the past, and similarly let $w_\gamma^{(\beta)}(G)$ be the weight of the same ancestor γ at generation G for β. These weights evolve according to eqn (3) with $w_\gamma^{(\alpha)}(0) = \delta_{\gamma,\alpha}$ and $w_\gamma^{(\beta)}(0) = \delta_{\gamma,\beta}$ at generation $G = 0$. In order to quantify the similarity between the two trees, we introduce the quantities

$$X^{(\alpha)}(G) = \sum_{\gamma=1}^{N(G)} [w_\gamma^{(\alpha)}(G)]^2$$

and

$$Y^{(\alpha,\beta)}(G) = \sum_{\gamma=1}^{N(G)} w_\gamma^{(\alpha)}(G) w_\gamma^{(\beta)}(G).$$

$Y^{(\alpha,\beta)}(G)$ measures the correlation between the two trees at generation G in the past and $X^{(\alpha)}(G)$ acts as a normalization factor. We then define the overlap $q^{(\alpha,\beta)}(G)$ between the two trees at that generation by

$$q^{(\alpha,\beta)}(G) = \frac{Y^{(\alpha,\beta)}(G)}{[X^{(\alpha)}(G) X^{(\beta)}(G)]^{1/2}}\,.$$

This overlap is a measure of the (cosine of the) angle between the two N-dimensional vectors $w_\gamma^{(\alpha)}(G)$ and $w_\gamma^{(\beta)}(G)$.[††] When $q^{(\alpha,\beta)}(G) \simeq 0$, the two vectors are essentially orthogonal and the ancestors of α and β are all different. On the other hand, when $q^{(\alpha,\beta)}(G) \simeq 1$, the vectors are almost identical (as for brothers).

For a large enough population, the fluctuations of $X^{(\alpha)}(G)$ and $Y^{(\alpha,\beta)}(G)$ are small around the population averaged values $\langle X(G)\rangle$ and $\langle Y(G)\rangle$ for almost all choices of α and β. Of course, if α and β are brothers, $Y^{(\alpha,\beta)}(G) = X^{(\alpha)}(G)$, a value very different from its average $\langle Y(G)\rangle$; it is however very unlikely to get brothers, sisters or even cousins if one picks up two individuals at random from a large population.

The averages $\langle X(G)\rangle$ and $\langle Y(G)\rangle$ can be calculated from the evolution of the weights (3). Initially, $X(0) = 1$ and $Y(0) = 0$ since the

---

[††] Similar quantities have been proposed as an indicator of the amount of evolutionary divergence between populations (Kimura, 1983). The quantity analogous to our weight $w_\gamma^{(\alpha)}$ in the population genetics approach is the frequency of the sampled alleles, the number of ancestors γ corresponds to the number of genes (that is, the dimension of the space in which the vector $w_\gamma^{(\alpha)}$ is embedded), and our individuals α and β correspond to the compared populations (Cavalli-Sforza & Conterio, 1960).

individuals α and β in any pair are different. Using the fact that for large N the fluctuations of $X^{(\alpha)}(G)$ and $Y^{(\alpha,\beta)}(G)$ are small, the expected value of the overlap $q(G)$ between two randomly chosen individuals is given by

$$q(G) \simeq \frac{\langle Y(G)\rangle}{\langle X(G)\rangle} = \frac{1}{1 + m^{G_c - G}}, \tag{6}$$

where

$$G_c = \frac{\log((m-1)N)}{\log m} - 1. \tag{7}$$

This expression is derived in the appendix. Of course, eqn (6) is only valid with probability one with respect to the random choice of α and β and with respect to the dynamics. We see that for large N, the overlap $q(G)$ is essentially zero for a number of generations of order $G_c \simeq \log N/\log m$ and then within a number of generations ΔG which *does not depend on N*, it becomes equal to unity. Figure 4 displays the averaged overlap $q(G)$ as a function of the number of generations G for different values of N. We have chosen $m = 2$ so that the population remains constant in size. We see that changing N does not change the G dependence except for a translation of the curve. In particular, the range ΔG on which the overlap changes from 1 to 99% does not depend on N. It is easy to check from eqn (6) that for $m = 2$, the overlap should satisfy

$$q(G+1) = \frac{2q(G)}{1 + q(G)} \tag{8}$$

(plain line in the inset). The fixed point $q(G) = 0$ is unstable for this map. All the trajectories finally converge to the stable fixed point $q(G) = 1$ for large G. Also, the quantity ΔG can be estimated by counting how many generations are required for the overlap to change from 1 to 99% and this gives from eqn (6)

$$\Delta G \simeq \log(10^4)/\log m,$$

that is, $\Delta G \simeq 14$ for $m = 2$ and $\Delta G \simeq 11$ for $m = 2.3$ as in Fig. 2. Typical values of G_c are $G_c \simeq 20$ for a population of constant size $N = 10^6$. For a population increasing with $m = 2.3$ as in Fig. 2, one gets $G_c = 21$ if the size in the last generation is $N = N(0) = 75$ millions.

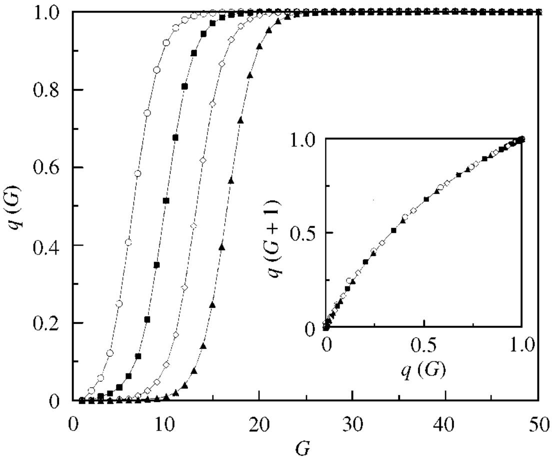

FIG. 4. The averaged overlap $q(G)$ as a function of the number of generations G. The results of simulations for different sizes of the population $N = 100$ (○), 1000 (■), 10000 (◇), 100 000 (▲) agree with this prediction, up to small finite-size corrections only visible for $N = 100$. The inset shows the results of simulations and the prediction (8).

The previous analysis can be easily extended to the hypothetical case of having an arbitrary number n of parents instead of 2. As is shown in the appendix, the statistical properties of genealogical trees in a population of constant size but arbitrary n are the same as for a population with only two parents and an expanding or shrinking size according to eqn (1). The described statistical properties are thus equivalent in (i) a system with sexual reproduction and a growth rate $m = n$ and (ii) a system with constant population size but a number m of genders.

The existence of a generation G_c around which the genealogical similarity among individuals changes from 0 to 1 and which grows logarithmically with the size of the population is one of our main results. This has to be compared with the number of generations required for the population to become genetically homogeneous (Donnelly & Tavaré, 1991; Harpending *et al.*, 1998), which grows proportionally to N. The difference is that when $G_c \ll G \ll N$, all the overlaps are 1, i.e. all the genealogical trees in the population have the same ancestors with the same weights, but the genomes are still very different: This is just an extension of the situation of brothers who have exactly the same genealogical tree but different genomes.

4. Simple Model for the Contribution of the Ancestors to the Genome

The evolution of a set of sequences subject to coalescence and recombination was first described by Hudson (1983). In this case, evolution proceeds until the most recent common ancestor for each set of homologous sites has been found. The set of MRCA sites does not necessarily belong to the genome of a single ancestor: on the contrary, it is, in general, spread on a finite fraction of the original population (Wiuf & Hein, 1997, 1999). In this section, we focus our attention on the statistical properties of the ancestry of a single extant genome. In particular, we calculate the equilibrium distribution for the fraction of material contributed by each ancestor.

Consider the whole set of genes that a present diploid organism has inherited from its parents. Although both parents contributed 50% each, it is no longer true that grandparents contributed 25% each, since independent assortment of chromosomes plus crossing over mixed in each of the parental gametes the material inherited from the previous generation. As a rough approximation to the output of genetic recombination, one might consider that each sequence is obtained as the addition of a fraction f of the genetic material of one parent and a fraction $1-f$ of the genetic material of the other parent with $f \in (0, 1)$. This would be true if the length L of the sequence was long enough (or infinitely long), so that there would be no restriction on the number of times it could be divided, and if one could forget the linear structure of the sequence. The process of coalescence and recombination (for small N) is schematically represented in Fig. 5.

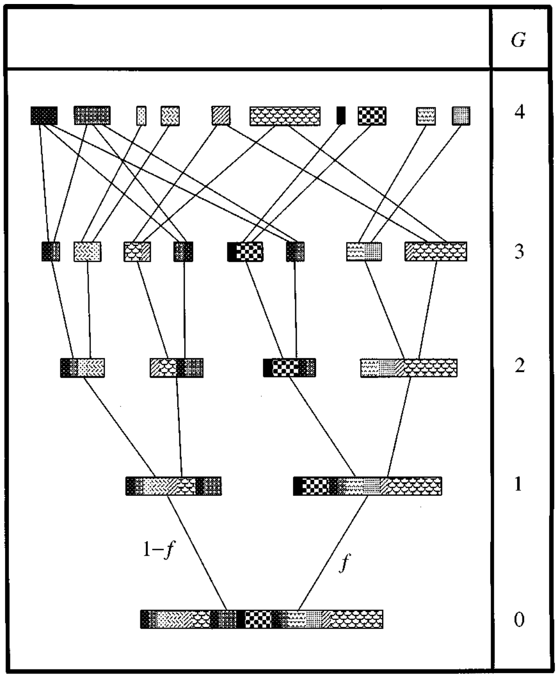

FIG. 5. Representation of the first 5 generations of the tree in Fig. 1 with a random distribution of the weight of an individual between its two parents. The fraction f of the weight contributed by each ancestor is randomly chosen from a distribution with average value $\langle f \rangle = 1/2$.

We can now repeat the analysis done previously to the present extension. We will discard the correlations between the values of f coming from a couple. This is equivalent to our assumption that fixing the pairs for k offspring or choosing the parents of each individual at random only has effects of order $O(N^{-1})$ (see the appendix), and we can therefore work in the simplest realization of the process. Hence, we assume that the fraction f takes independent values for each parent. The recursive equations (3) for the weights become

$$w_{\gamma}^{(\alpha)}(G+1) = \sum_{\gamma' \text{ children of } \gamma} f_{\gamma'} w_{\gamma'}^{(\alpha)}(G), \qquad (9)$$

where the weight $w_{\gamma}^{(\alpha)}(G)$ means now the fraction of the genetic material of individual α inherited from ancestor γ at generation G. The random fraction f is chosen anew for each offspring from a distribution $\rho(f)$ (with average value $\langle f \rangle = 1/2$). This implies that now even brothers would have different weights for their ancestors, and hence brings us slightly closer to the real genetic process.

Following the procedure described in the appendix, one can calculate the fraction S of ancestors without lines of descent in the present (as we also show in Section 2) and the exponent ξ for the distribution $P(w)$. In general, given the distribution $\rho(f)$ for the contributions of the parents, we get

$$S(\infty) = \mathrm{e}^{mS(\infty)-m}, \qquad (10)$$

$$1 = S(\infty) m^{2+\xi} \langle f \rangle^{1+\xi} \int f^{-\xi-1} \rho(f) \, \mathrm{d}f \qquad (11)$$

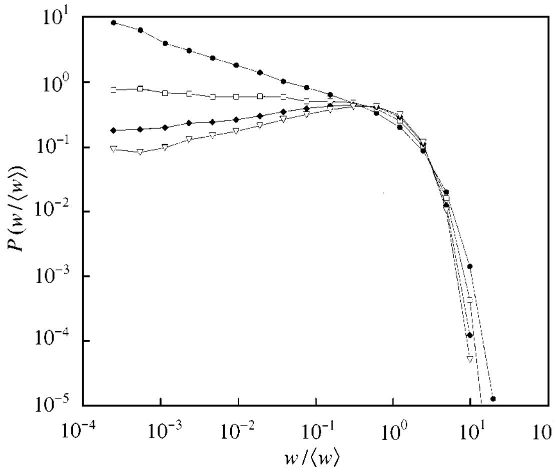

FIG. 6. Stationary distribution of weights $P(w/\langle w\rangle)$ versus $w/\langle w\rangle$ for several choices of δ. The fixed population size is $N = 4096$, and we have averaged over 10^3 independent runs. Values of δ are as shown in the legend: $\delta = 0.5$ (●——●); $\delta = 0.35$ (□——□); $\delta = 0.2$ (◆——◆); $\delta = 0.05$ (▽——▽).

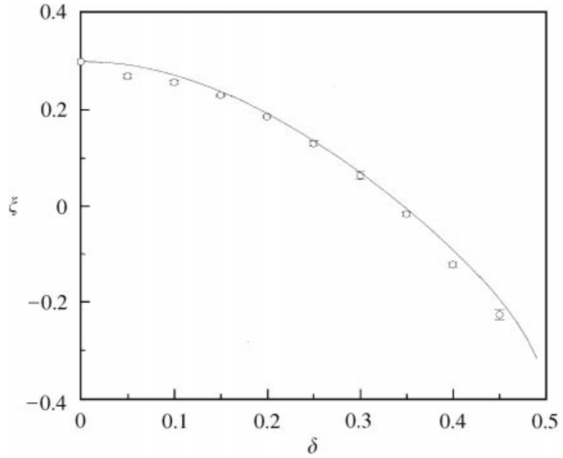

FIG. 7. Comparison between the predicted values of the exponent ξ (——) given by eqn (12) and the results of the simulations for different values of δ (○). Parameters are as in Fig. 6. For a value of $\delta \simeq 0.35$, the exponent ξ changes sign. This point is important since the typical contribution of a randomly chosen ancestor changes suddenly in a finite amount.

as one can easily show from eqn (9) that the generating function $h_G(\lambda)$ defined by $h_G(\lambda) = \langle \exp[\lambda w(G)/\langle w(G)\rangle]\rangle$ has a limit $h_\infty(\lambda)$ for large G which satisfies

$$h_\infty(\lambda) = \exp\left[-m + m\int \rho(f)\,\mathrm{d}f h_\infty\left(\frac{\lambda f}{m\langle f\rangle}\right)\right].$$

Figure 6 summarizes the changes in the distribution $P(w)$ for different distributions $\rho(f)$ of the random variable f. We have considered a simple case of a population of constant size (i.e. $m = 2$) and with $\rho(f) = 1/(2\delta)$ uniform in the interval $(1/2 - \delta, 1/2 + \delta)$. In this particular case, an implicit relation between δ and the exponent ξ can be obtained,

$$\delta\xi = S[(\tfrac{1}{2} - \delta)^{-\xi} - (\tfrac{1}{2} + \delta)^{-\xi}]. \qquad (12)$$

As δ varies, $P(w)$ remains a power law at small w (i.e. $P(w) \propto w^{\xi}$), and the exponent ξ monotonously decreases with δ. In particular, for $\delta \simeq 0.35$, ξ changes sign: the maximum of $P(w)$ moves discontinuously from $w/\langle w\rangle \simeq 1$ to $w/\langle w\rangle \simeq 0$. The exponents obtained through simulations of the process are represented in Fig. 7 together with the numerical solution of eqn (12), showing a good agreement.

5. Discussion

We have analysed the statistical properties of genealogical trees generated inside a closed sexual population. We focused our interest on the distribution of the repetitions of ancestors in the trees and on the amount of genetic material contributing to an extant genome. The precise values of ξ, $S(\infty)$, G_c and ΔG depend only weakly on the details of the model and do not change qualitatively if for instance a non-Poissonian distribution of offspring is used. Moreover, we have shown how our results can be extended to the hypothetical case of having an arbitrary number n of parents: indeed, this case proves to be equivalent to a biparental population with a growth rate $m/2 = n/2$.

The problem analysed here presents a number of connections to other fields. Equations similar to eqn (3) appear also in the distribution of constraints in granular media where the variables w represent the force acting on each grain and the recursion (3) expresses the way in which constraints are transmitted from one layer to the next (Coppersmith *et al.*, 1996). In this case, $n \neq 2$ and even fluctuating n would be perfectly realistic. The fact that the overlap changes from 0 to 1 within a small number of generations ΔG independent of the size of the population and after

$G_c \simeq \log N$ generations is also very reminiscent of the sharp cutoff phenomenon characteristic of some natural mixing processes modelled by Markov chains. One example of such systems is the shuffling of cards, where the stationary state in which the system has lost almost all information about the initial ordering of the Q cards is reached through a sharp cutoff after about log Q riffle shuffles (Diaconis, 1996).

It is clear that the study of the interplay between the weights calculated in our generalized model and the structure of the genome would require more sophisticated approaches (Derrida & Jung-Muller, 1999; Wiuf & Hein, 1997, 1999). We have discarded the correlations between the history of neighbouring sites in a sequence and assumed the independence of the factors f. Actually, the closer in the sequence two positions are, the more correlated are their genealogical histories (Kaplan & Hudson, 1985). This fact constrains the possible breaking points for our simulated sequences, implying that the random factors f in eqn (9) are a crude approximation to reality.

Since we have faced the problem from a statistical perspective, our results represent the average, typical behaviour, and are only valid with probability one when the population size is large. We did not study fluctuations due to the finite size of the population. Nonetheless, we hope that our results contribute to a better understanding of the role of genealogy in the degree of diversity of finite-size interbreeding populations.

The authors acknowledge discussions with Jordi Bascompte, Ugo Bastolla and Julio Rozas. SCM thanks the Alexander von Humboldt Foundation for support.

REFERENCES

Austerlitz, F. & Heyer, E. (1998). Social transmission of reproductive behavior increases frequency of inherited disorders in a young-expanding population. *Proc. Nat. Acad. Sci. U.S.A.* **95,** 15140–15144.

Cavalli-Sforza, L. L. & Conterio, F. (1960). Analisi della fluttuazione di frequenze geniche nella popolazione della Val Parma. *Atti Assoc. Genet. Ital.* **5,** 333–344.

Coppersmith, S. N., Liu, C.-H., Majumdar, S. N., Narayan, O. & Witten, T. A. (1996). Model for force fluctuations in bead packs. *Phys. Rev. E* **53,** 4673–4685.

Derrida, B. & Jung-Muller, B. (1999). The genealogical tree of a chromosome. *J. Stat. Phys.* **94,** 277–298.

Derrida, B., Manrubia, S. C. & Zanette, D. H. (1999). Statistical properties of genealogical trees. *Phys. Rev. Lett.* **82,** 1987–1990.

Dewdney, A. K. (1986). Computer recreations: Branching phylogenies of the Paleozoic and the fortunes of English family names. *Scientific American* **254,** 12–16.

Diaconis, P. (1996). The cutoff phenomenon in finite Markov chains. *Proc. Nat. Acad. Sci. U.S.A.* **93,** 1659–1664.

Donnelly, P. & Tavaré, S. (1995). Coalescents and genealogical structure under neutrality. *Annu. Rev. Genet.* **29,** 401–421.

Fisher, R. A. (1922). On the dominance ratio. *Proc. Roy. Soc. Edin.* **42,** 321–341.

Fisher, R. A. (1930). *The Genetical Theory of Natural Selection.* Oxford: Clarendon (1958, 2nd edn, New York: Dover).

Gale, J. S. (1990). *Theoretical Population Genetics.* London: Unwin Hyman Ltd.

Gouyon, P. H. (1999). Darwin et Mendel on du vertige qui saisit l'évolutionniste qui tente de remonter dans l'arbre. *Mag. Littéraire* **374,** 58–61.

Haldane, J. B. S. (1927). A mathematical theory of natural and artificial selection. Part V: selection and mutation. *Proc. Camb. Phil. Soc.* **26,** 838–844.

Harpending, H. C., Batzer, M. A., Gurven, M., Jorde, L. B., Rogers, A. R. & Sherry, S. T. (1998). Genetic traces of ancient demography. *Proc. Nat. Acad. Sci. U.S.A.* **95,** 1961–1967.

Harris, T. E. (1963). *The Theory of Branching Processes.* Berlin: Springer-Verlag OHG.

Hudson, R. R. (1983). Properties of the neutral allele model with intergenic recombination. *Theor. Pop. Biol.* **28,** 183–201.

Hudson, R. R. (1991). Gene genealogies and the coalescent process. In: *Oxford Surveys in Evolutionary Biology* (Futuyma D. and Antonovics, J., eds), Vol. 7, pp. 1–44. Oxford: Oxford University Press.

Kaplan, N. & Hudson, R. R. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111,** 147–164.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution.* Cambridge: Cambridge University Press.

Kingman, J. F. C. (1982). The coalescent. *Stochast. Proc. Appl.* **13,** 235–248.

Kingman, J. F. C. (1993). *Poisson Processes.* Oxford Studies in Probability, vol. 3. Oxford: Oxford University Press.

Ohno, S. (1996). The Malthusian parameter of ascents: what prevents the exponential increase of one's ancestors? *Proc. Nat. Acad. Sci. U.S.A.* **93,** 15276–15278.

Serva, M. & Peliti, L. (1991). A statistical model of an evolving population with sexual reproduction. *J. Phys. A: Math. Gen.* **24,** L705–L709.

Wiuf, C. & Hein, J. (1997). On the number of ancestors to a DNA sequence. *Genetics* **147,** 1459–1468.

Wiuf, C. & Hein, J. (1999). The ancestry of a sample of sequences subject to recombination. *Genetics* **151,** 1217–1228.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16,** 97–159.

APPENDIX A

In this appendix we have regrouped the technical aspects of the derivations of the main eqns (4)–(6), (10) and (11) presented in the body of the paper.

One may consider several variants of the model which all give a Poisson distribution for the number of offspring when the size of the population is large. For instance, the population size could be strictly multiplied by a factor $m/2$ at each generation or it could fluctuate (if we take the number of offspring from the Poisson distribution). One might decide that each individual has two parents chosen at random in the previous generation or form fixed couples and assign each couple some children. All these variants do not change the results when the population size is large, but might affect some finite-size corrections that we compute in this appendix.

We will choose the following version of the model, which makes the calculation of the finite-size corrections not too difficult. Our population has a given size $N(G)$ at each generation G in the past, and we will assume that all the $N(G)$ are very large, at least in the range of generations G that we will consider. Now, to construct the ancestors of all the $N(G)$ individuals at generation G in the past, we choose for each of them a pair of parents at random among the $N(G+1)$ individuals at the previous generation (to facilitate the calculation, we do not even exclude that the two parents might be equal). Within this model, the number k of children of an individual at generation $G+1$ is random and can be written as

$$k = \sum_{i=1}^{2N(G)} z_i$$

where $z_i = 1$ with probability $1/N(G+1)$ and $z_i = 0$ otherwise. It follows that the whole distribution of k can be calculated. The probability p_k that an individual at generation $(G+1)$ has exactly k children is given by the binomial distribution

$$p_k = \frac{(2N(G))!}{k!(2N(G)-k)!}\left(\frac{1}{N(G+1)}\right)^k \times \left(1 - \frac{1}{N(G+1)}\right)^{2N(G)-k} \qquad \text{(A.1)}$$

In particular,

$$\langle k \rangle = \frac{2N(G)}{N(G+1)},$$

$$\langle k(k-1) \rangle = \frac{2N(G)[2N(G)-1]}{N(G+1)^2},$$

$$\langle k(k-1)(k-2) \rangle = \frac{2N(G)[2N(G)-1][2N(G)-2]}{N(G+1)^3}. \qquad \text{(A.2)}$$

If the population size is multiplied by a factor $m/2$ at each generation, i.e. if $N(G) = N(G+1)m/2$ (as G counts the number of generations in the past), one recovers from eqn (A.1) the Poisson distribution $p_k = m^k e^{-m}/k!$ for large $N(G)$.

Calculation of the Density of Individuals without Long-term Descendants and Derivation of eqn (4)

To establish eqn (4), one simply needs to notice that for an individual to have no descendants after $G+1$ generations, all his children should have no descendants after G generations. Let $M(G)$ be the number of individuals with no descendants at generation G in the past. Given $M(G)$, one can write $M(G+1)$ as

$$M(G+1) = \sum_{\gamma=1}^{N(G+1)} y_\gamma,$$

where $y_\gamma = 1$ if all the children of γ are among the $M(G)$ and $y_\gamma = 0$ otherwise. It can be shown that

$$\langle y_\gamma \rangle = \left(1 - \frac{1}{N(G+1)}\right)^{2N(G)-2M(G)},$$

and

$$\langle y_\gamma y_{\gamma'} \rangle = \left(1 - \frac{2}{N(G+1)}\right)^{2N(G)-2M(G)}$$

for $\gamma \neq \gamma'$. This gives

$$\langle M(G+1) \rangle = N(G+1)\left(1 - \frac{1}{N(G+1)}\right)^{2N(G)-2M(G)}, \qquad \text{(A.3)}$$

$$\langle M^2(G+1)\rangle = \langle M(G+1)\rangle + N(G+1)[N(G+1)-1] \times \left(1-\frac{2}{N(G+1)}\right)^{2N(G)-2M(G)}. \qquad \text{(A.4)}$$

When all the M's and N's are large, we see from eqns (A.3) and (A.4) that the fluctuations of $M(G+1)$ are small (as $\langle M^2(G+1)\rangle - \langle M(G+1)\rangle^2 \ll \langle M(G+1)\rangle^2$), and one finds from eqn (A.3) that the ratio $M(G)/N(G) \equiv S^{(\alpha)}(G)$ satisfies

$$S^{(\alpha)}(G+1) = \exp\left[\frac{2N(G)}{N(G+1)}\left(S^{(\alpha)}(G)-1\right)\right],$$

which is identical to eqn (4) for $N(G) = N(G+1)m/2$.

Time Evolution of the Distribution of the Weights

From the recursion (3) and from the known distribution (A.1) of k one can write recursions for the moments of the weights

$$\langle w_\gamma^{(\alpha)}(G+1)\rangle = \frac{\langle k\rangle}{2}\langle w_\gamma^{(\alpha)}(G)\rangle, \qquad \text{(A.5)}$$

$$\langle [w_\gamma^{(\alpha)}(G+1)]^2\rangle = \frac{\langle k\rangle}{4}\langle [w_\gamma^{(\alpha)}(G)]^2\rangle + \frac{\langle k(k-1)\rangle}{4}\langle w_\gamma^{(\alpha)}(G) w_{\gamma'}^{(\alpha)}(G)\rangle, \qquad \text{(A.6)}$$

where $\gamma \neq \gamma'$. The normalization $\sum_\gamma w_\gamma^{(\alpha)} = 1$ allows one to rewrite

$$\langle w_\gamma^{(\alpha)}(G) w_{\gamma'}^{(\alpha)}(G)\rangle = \frac{1}{N(G)-1} \times [\langle w_\gamma^{(\alpha)}(G)\rangle - \langle [w_\gamma^{(\alpha)}(G)]^2\rangle]$$

and together with the known moments (A.1) gives that

$$\langle w_\gamma^{(\alpha)}(G+1)\rangle = \frac{N(G)}{N(G+1)}\langle w_\gamma^{(\alpha)}(G)\rangle = \frac{1}{N(G+1)}, \qquad \text{(A.7)}$$

$$\langle [w_\gamma^{(\alpha)}(G+1)]^2\rangle = \left[\frac{N(G)}{2N(G+1)} - \frac{N(G)[2N(G)-1]}{2N(G+1)^2[N(G)-1]}\right] \times \langle [w_\gamma^{(\alpha)}(G)]^2\rangle + \frac{2N(G)-1}{2N(G+1)^2[N(G)-1]}, \qquad \text{(A.8)}$$

where $\gamma \neq \gamma'$.

For large $N(G)$, if the ratio $N(G+1)/N(G) = 2/m$, as in the case of a population increasing by a factor $m/2$ at each new generation, expression (A.8) becomes simpler and one gets

$$\langle [w_\gamma^{(\alpha)}(G+1)]^2\rangle = \frac{m}{4}\langle [w_\gamma^{(\alpha)}(G)]^2\rangle + \frac{m^2}{4}\left(\frac{1}{N(G)}\right)^2. \qquad \text{(A.9)}$$

In this limit, we have from eqn (A.2) that $\langle k\rangle = m$ and $\langle k(k-1)\rangle = m^2$, and we see that eqn (A.9) means that in eqn (A.4) the weights $w_\gamma^{(\alpha)}$ and $w_{\gamma'}^{(\alpha)}$ are, for large $N(G)$, uncorrelated. The calculation of higher moments of the weights can be done in the same manner and for large $N(G)$ the weights of different ancestors become again uncorrelated.

If the population size changes in time, the distribution of the weights cannot be stationary. This is already visible in expression (A.4) which shows that even the first moment of the weights changes with G. One can, however, check from eqns (A.4) and (A.9) that the ratio $\langle [w_\gamma^{(\alpha)}(G)]^2\rangle / \langle w_\gamma^{(\alpha)}(G)\rangle^2$ which satisfies

$$\frac{\langle [w_\gamma^{(\alpha)}(G+1)]^2\rangle}{\langle w_\gamma^{(\alpha)}(G+1)\rangle^2} = \frac{1}{m}\frac{\langle [w_\gamma^{(\alpha)}(G)]^2\rangle}{\langle w_\gamma^{(\alpha)}(G)\rangle^2} + 1 \qquad \text{(A.10)}$$

has a limit $m/(m-1)$ as G increases. Moreover, as the initial value of this ratio is $N(0)$, the number of generations G_c to converge to this limit is $G_c \sim \log N(0)/\log m$. Higher moments of the weights behave in a similar way and one can write recursions for ratios which generalize eqn (A.10) and which show that all the ratios have limits.

This indicates that the distribution of the ratio $w/\langle w \rangle$ becomes stationary. In the limit of large $N(G)$ [considering that the weights of the different children γ' in eqn (3) can be taken as independent and that the distribution of k becomes Poissonian], one finds that the generating function $h_G(\lambda)$ defined by

$$h_G(\lambda) = \left\langle \exp\left[\lambda \frac{w_\gamma^{(\alpha)}(G)}{\langle w(G) \rangle}\right]\right\rangle \tag{A.11}$$

satisfies

$$h_{G+1}(\lambda) = \sum_k p_k \left[h_G\left(\frac{\lambda \langle w(G) \rangle}{2\langle w(G+1) \rangle}\right)\right]^k$$

$$= \exp[-m + m h_G(\lambda/m)]. \tag{A.12}$$

Recursion (A.12) generalizes to the case $m \neq 2$ (i.e. the case of an exponentially increasing population) the result of our previous work obtained for a population of constant size ($m = 2$). Similar recursions have been studied in the theory of branching processes (Harris, 1963). The use of generating functions in population genetics is well illustrated in the book by Gale (1990), where this method is for example applied to the calculation of the probability of fixation of a mutant allele.

It is remarkable, that if one considers an imaginary world where each individual would have n parents (instead of 2), the generating function (A.11), in the case of a population of constant size, would satisfy the recursion (A.12) with $m = n$. This means that as long as the distribution of weights is concerned, the problem of a large population of constant size with m parents per individual is identical to the problem of a population of size increasing at each generation by a factor $m/2$ with two parents per individual.

Stationary Distribution

For large G, if we fix the ratio $N(G)/N(G+1) = m/2$, the generating function $h_G(\lambda)$ converges to $h_\infty(\lambda)$ solution of

$$h_\infty(\lambda) = \exp[-m + m h_\infty(\lambda/m)]. \tag{A.13}$$

If one expands the solution around $\lambda = 0$, one finds that

$$h_\infty(\lambda) = 1 + \lambda + \frac{1}{2}\frac{m}{m-1}\lambda^2 + \frac{1}{6}\frac{m^2(m+2)}{(m^2-1)(m-1)}\lambda^3 + \cdots$$

and the comparison with eqn (A.11) gives for large G

$$\frac{\langle w^2 \rangle}{\langle w \rangle^2} \to \frac{m}{m-1}; \quad \frac{\langle w^3 \rangle}{\langle w \rangle^3} \to \frac{m^2(m+2)}{(m^2-1)(m-1)},$$

which means that, in principle, the whole shape of $P(w)$ can be extracted from eqn (A.13). In particular, one can predict the power law of $P(w)$ at small w. Trying to solve eqn (A.13) for large negative λ, if one writes

$$h_\infty(\lambda) - S(\infty) \simeq \frac{1}{|\lambda|^{\xi+1}} \tag{A.14}$$

one finds, as expected, that $S(\infty)$ is the fixed point of eqn (4). Equation (A.14) is equivalent to the assumption that $P(w) \sim w^\xi$ at small w, where the exponent ξ should satisfy

$$1 = S(\infty) m^{\xi+2}.$$

This completes the derivation of eqn (5) which has already been discussed in our previous work (Derrida *et al.*, 1999).

Overlap between Two Trees

Let us now show how eqn (6) can be derived. Starting from recursion (3), one obtains by averaging over all the links relating generation G to generation $G + 1$

$$\langle w_\gamma^{(\alpha)}(G+1) w_\gamma^{(\beta)}(G+1)\rangle = \frac{\langle k\rangle}{4} \langle w_\gamma^{(\alpha)}(G) w_\gamma^{(\beta)}(G)\rangle + \frac{\langle k(k-1)\rangle}{4} \langle w_\gamma^{(\alpha)}(G) w_{\gamma'}^{(\beta)}(G)\rangle, \qquad (A.15)$$

where $\gamma \neq \gamma'$ and the averages over k are carried out with respect to eqn (A.1). This gives

$$\langle w_\gamma^{(\alpha)}(G+1) w_\gamma^{(\beta)}(G+1)\rangle = \frac{m}{4} \langle w_\gamma^{(\alpha)}(G) w_\gamma^{(\beta)}(G)\rangle + \frac{1}{4}\left(m^2 - \frac{m}{N(G+1)}\right) \langle w_\gamma^{(\alpha)}(G) w_{\gamma'}^{(\alpha)}(G)\rangle. \qquad (A.16)$$

Using the fact that the sum $\sum_{\gamma'} w_{\gamma'}^{(\beta)}(G) = 1$, so that $\langle w_\gamma^{(\alpha)}(G)\rangle = 1/N(G)$ at all generations, one gets that

$$\langle w_\gamma^{(\alpha)}(G+1) w_\gamma^{(\beta)}(G+1)\rangle = \frac{m}{4} \langle w_\gamma^{(\alpha)}(G) w_\gamma^{(\beta)}(G)\rangle + \frac{1}{4}\left(m^2 - \frac{m}{N(G+1)}\right) \times \frac{\frac{1}{N(G)} - \langle w_\gamma^{(\alpha)}(G) w_\gamma^{(\beta)}(G)\rangle}{N(G) - 1}. \qquad (A.17)$$

Keeping only the dominant contributions for large N'*s* we arrive at

$$\langle w_\gamma^{(\alpha)}(G+1) w_\gamma^{(\beta)}(G+1)\rangle = \frac{m}{4} \langle w_\gamma^{(\alpha)}(G) w_\gamma^{(\beta)}(G)\rangle + \frac{m^2}{4} \frac{1}{N(G)^2}.$$

Comparing this expression with eqn (A.15), one sees that for large N, one could have simply neglected the correlations between the weights of different individuals [i.e. directly replaced $\langle w_\gamma^{(\alpha)}(G) w_{\gamma'}^{(\beta)}(G)\rangle$ by $\langle w_\gamma^{(\alpha)}(G)\rangle \langle w_{\gamma'}^{(\beta)}(G)\rangle$ and used the Poisson distribution instead of eqn (A.1)]. The previous recursion can be integrated

$$\langle w_\gamma^{(\alpha)}(G) w_\gamma^{(\beta)}(G)\rangle = \left[\langle w_\gamma^{(\alpha)}(0) w_\gamma^{(\beta)}(0)\rangle + \frac{1}{N^2} \frac{m}{m-1} (m^G - 1) \right] \left(\frac{m}{4}\right)^G, \qquad (A.18)$$

and using the fact that $\langle w_\gamma^{(\alpha)}(G) w_\gamma^{(\beta)}(G)\rangle$ is equal to $Y(G)/N(G)$ when $\alpha \neq \beta$ and to $X(G)/N(G)$ when $\alpha = \beta$, one finds (with $X(0) = 1$ and $Y(0) = 0$) that

$$\frac{\langle Y(G)\rangle}{\langle X(G)\rangle} = \frac{(m^G - 1) m^{-G_c}}{1 + (m^G - 1) m^{-G_c}},$$

where G_c is given by eqn (7). For large N, that is, for large G_c this reduces to eqn (6) in the whole range where the expression departs from 0 or 1, that is, for G of order G_c. Finally, one can check that with the value of G_c given by eqn (7), $N(G)$ is always large, as long as N is large, so that our assumption that all the N's are large is legitimate.