AP

At the Boundary between Biological and Cultural Evolution: The Origin of Surname Distributions

Susanna C. Manrubia*† and Damián H. Zanette‡

†*Centro de Astrobiología, INTA-CSIC Ctra. de Ajalvir km.* 4, 28850 *Torrejón de Ardoz, Madrid, Spain and* ‡*Consejo Nacional de Investigaciones Científicas y Técnicas, Centro Atómico Bariloche and Instituto Balseiro*, 8400 *Bariloche, Argentina*

Surnames and non-recombining alleles are inherited from a single parent in a highly similar way. A simple birth–death model with mutations can accurately describe this process. Exponentially growing and constant populations are investigated, and we study how different compositions of the founder populations can be observed in present-day diversity distributions. We analyse different quantities in the statistically stationary state, both through analytic and numerical methods. Our results compare favourably to field data for family sizes in several countries. We discuss the relationship between the distribution of surnames and the genetic diversity of a population. © 2002 Elsevier Science Ltd. All rights reserved.

1. Introduction

Biological and cultural features of human populations have been traditionally studied by separate disciplines, but the parallelisms between biological and cultural evolution have been put forward by a number of researchers. Already Darwin (1871) pointed out that "the formation of different languages and of distinct species and the proofs that both have been developed through a gradual process are curiously parallel".

Cultural traits are transmitted from ancestors to their descendence, in a process analogous to inheritance, and are subject to changes, similar to mutations, by interaction between individuals—such as teaching and imitation. Moreover, they usually fulfill a practical purpose,

which amounts to being subject to selection. In fact, they enhance the relationships within human groups, defining social entities comparable to certain biological species and populations. The quantification of cultural traits has been attempted only recently. For example, Cavalli-Sforza *et al.* (1982) applied concepts from the quantitative theory of biological evolution to construct a theory of cultural evolution. They analysed 40 traits, ranging from political preferences to superstitions. Many of these traits are subject to high mutability, since they are influenced by fashion, individual acquaintances, and personal experience, and one does not expect their quantitative properties to be directly comparable to those of any biological feature. Other traits, on the other hand, are better preserved. Among them we find languages and surnames. Language is essential to integrate the individual to society; surnames are historical—though recent—signs of identity

*Corresponding author. Tel.: +34-91-5202-089; fax: +34-91-520-1074
E-mail address: susanna@complex.upc.es (S.C. Manrubia).

in social groups. Quite early, Galton & Watson (1874) dealt with the problem of the extinction of surnames. This problem is equivalent to that of the extinction of a mutant allele in a population, although this relation was only established half a century later (Fisher, 1922; Haldane, 1927), when the first quantitative approaches to biological evolutionary processes took place.

Comparative methods analogous to biological taxonomy are used to determine the degree of similarity between languages. This method returns the *genetic classification* of linguistic diversity (see for example Greenberg, 1992; Ruhlen, 1992). Recently, a quantitative study of the taxonomy of languages has been carried out (Zanette, 2001), showing that the distribution of the number of subtaxa within a taxa displays scaling properties, quantitatively similar to those disclosed in biological taxonomy by Burlando (1990, 1993). That is, if n is the number of subtaxa belonging to a given taxa—say, the number of languages in the Indo-European family, or the number of species in the genus *Canis*—the fraction $p(n)$ of taxa that have precisely n subtaxa scales as $p(n) \sim n^{-\beta}$. The exponent is found in the interval $1 \leqslant \beta \leqslant 2$ in both cases. This gives quantitative support to Darwin's observation on the "equivalence" between the mechanisms behind biological and linguistic evolution. A complementary comparison is that of linguistic abundance, measured as the number of individuals speaking a language, with the number of individuals of a biological species. Again, the frequency as a function of the number of individuals has scaling properties both for languages (Gomes *et al.*, 1999) and for species (Pielou, 1969), see Fig. 1.

Surnames are cultural traits (Cavalli-Sforza & Feldman, 1981) whose transmission bears strong similarity with that of some biological features. They are inherited from one of the parents, usually the father, much in the same way as the Y-chromosome or the mitochondrial DNA. The extinction of a surname and the persistence of a non-recombining neutral allele are equivalent problems. This is not only a mathematical fact, but has also practical implications. Indeed, to assess the multiple or single origin of a surname one can turn to genetic measures, since males sharing the same surname might also share the
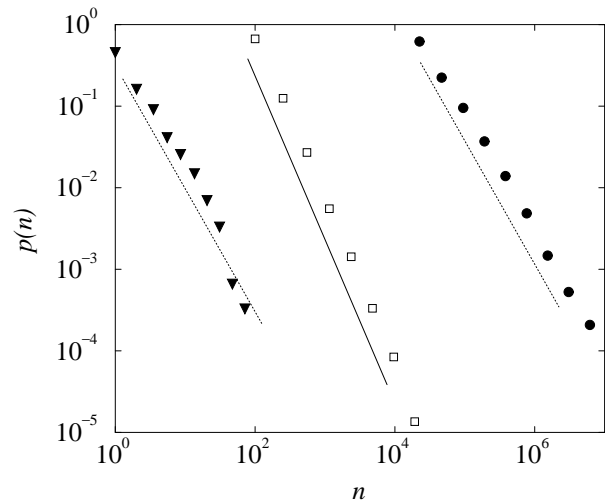

FIG. 1. Scaling behaviour of the fraction $p(n)$ of species represented by n individuals (triangles), of surnames borne by n persons (squares), and of languages with n speakers (circles). Data for the distribution of species abundance are from Poore (1968), corresponding to trees in a Malaysian rainforest [see also Solé and Alonso (1998)]. Data for surnames (beginning by A) are from the 1996 Berlin telephone book. Data for languages are from `http://www.sil.org/ethnologue/preface.html`. As a guide to the eye we draw lines with slope -2 (solid line) and $-3/2$ (dotted lines). Data sets have been mutually shifted for better visualization.

same haplotype in the non-recombining segment of the Y chromosome (Sykes & Irven, 2000). In a very large population, the statistical properties of the surname distribution can be strongly correlated with genetic diversity (Barrai *et al.*, 1996), and may even be used to understand the genetic structure of a population (Yasuda *et al.*, 1974). Recent reports on actual populations (Miyazima *et al.*, 2000; Zanette & Manrubia, 2001) show that the distribution of surnames follows the same statistical law observed for languages and biological species. Namely, if n is the number of individuals bearing a given surname, the fraction $p(n)$ of surnames decreases with n as $p(n) \sim n^{-\gamma}$ (see Fig. 1). Here, however, the exponent is always $\gamma \approx 2$. This paper focuses on a model aimed at predicting this kind of regularities, observed in disparate human populations.

2. Theoretical Approaches to Surname Evolution

At the mathematical level, several models have been proposed and analysed in order to identify quantitative properties of surnames

evolution—or, equivalently, those of non-recombining neutral alleles. The main points addressed by these studies are (i) the probability of fixation of a given surname/allele in a closed population (Galton & Watson, 1874; Fisher, 1922; Haldane, 1927; Moran, 1962; Lange, 1981; Rannala 1997; Hull 1998), and (ii) the distribution of the number of individuals bearing the same surname/allele (Kimura & Crow, 1964; Karlin & McGregor, 1967; Fox & Lasker, 1983; Panaretos, 1989; Gale 1990; Consul, 1991; Islam, 1995; Zanette & Manrubia, 2001). Indeed, these two questions cover complementary aspects of the same problem. In (i), one deals with a closed population (no immigrants enter the system), and implicitly assumes that the mutation rate is low enough, such that fixation can indeed occur before a mutant form appears. Suppose that there are N individuals in the population. We know from coalescence theory that the time n_g (in units of the number of generations) required for a neutral allele to be fixed is of order $n_g \sim N$. Now suppose that the mutation rate per generation and per individual is r. Then $M \equiv r n_g N$ is the average number of mutants after n_g generations have elapsed. Only if $M \ll 1$, that is if $r \ll N^{-2}$, will the fixation of the allele be possible. As soon as this inequality is violated, a new situation arises, in which both neutral drift and mutation play relevant roles. In this case, a broad distribution of surnames or of genetic diversity is expected. Actually, the value of r will be usually fixed by the nature of the problem, while the size N of the considered population can increase enough such that $N^2 > 1/r$. Therefore, the statistical behaviour crosses over to the second regime, where the appearance of mutants cannot be discarded. We are then in the assumptions of the models of class (ii).

All the models in class (ii) represent systems where statistical equilibrium is reached. While the neutral drift drives the less-frequent surnames out of the system, mutations generate new surnames. For large times, the number of different surnames in the system is much larger than unity. Yasuda and co-workers (Yasuda et al., 1974) used a stochastic model for a population with a fixed number of individuals. At each evolution step, a new individual with the surname of his father is added, but he acquires a new surname with a given probability. In any case, a randomly chosen individual is eliminated in order to keep the population constant. Their analytical results compared successfully to field data despite the restriction of constant population, which forces the elimination of individuals. In this model, the size of the progeny is Poisson distributed.

Other models start with a single individual in the population and new individuals carrying the same surname or a new one are sequentially added (Panaretos, 1989; Consul, 1991; Islam, 1995). These fall in the category of branching processes (Harris, 1963) with an increasing number of individuals in the population. Panaretos (1989) rephrased a model introduced by Simon (1955) in the context of linguistics, as follows. At each step, an individual is chosen to be the father of a newborn and his surname is transmitted with probability $1 - \alpha$. With the complementary probability α, the newborn is assigned a surname not present in the population. Recently, we have modified Simon's model by (i) introducing an additional parameter μ which represents the death rate of the individuals in the population, and (ii) allowing for arbitrary initial conditions. Preliminary results have been successfully compared with large data sets (Zanette & Manrubia, 2001). This is the starting point of the present work, where we present new analytical and numerical results for this birth–death model.

We describe the model in the next section, where our analytical results are derived. Questions like the role played by the composition of the founder population and by death events in an exponentially growing population are addressed. We also analyse a system with constant population, which turns out to quantitatively differ from the previous case. In Section 4 we numerically check our analytical results and run computer simulations of the model in situations not amenable to an analytical description. In Section 5 our results are tested against field data. We finish with an overall discussion in the last section.

3. Birth–Death Model for Surname Evolution

Our model population evolves in discrete steps, each step corresponding to the birth of a

new individual. At each time step, moreover, an individual is chosen at random from the whole population and becomes removed with probability μ, representing a death event. The total population at step s, $P(s)$, is therefore a stochastic process governed by the evolution equation

$$P(s + 1) = P(s) + 1 - w(s), \qquad (1)$$

where $w(s)$ is the dichotomic stochastic process

$$w(s) = \begin{cases} 1 & \text{with probability } \mu, \\ 0 & \text{with probability } 1 - \mu. \end{cases} \qquad (2)$$

We show in Appendix A that, averaging over realizations of the stochastic process $w(s)$, the average total population $\bar{P}$ grows exponentially in time:

$$\bar{P}(t) = P_0 \exp[v(1 - \mu)t]. \qquad (3)$$

Here, v is the birth rate per individual and unit time, and the product $v\mu$ turns out to be the corresponding death rate; P_0 is the initial population.

We think of the population as divided into groups—the families—within which all individuals bear the same surname. At each birth event, the newborn is assigned a new surname, not previously present in the population, with probability α. This probability can be seen as a mutation rate for surnames, but could also be interpreted as a measure of migration towards the population of individuals with new surnames. With the complementary probability, $1 - \alpha$, a pre-existent individual is chosen at random from the population and becomes the newborn's father, i.e. his surname is given to the newborn. Surnames are therefore assigned with probabilities proportional to the size of the corresponding families, allowing, however, for fluctuations inherent to the random distribution of births among families. Consequently, the distribution of surnames is driven by a stochastic multiplicative process (Van Kampen, 1981), modulated in turn by the total population growth. This process is analogous to the mechanism proposed by Simon (1955) to model the frequency distribution of words and city sizes, described by Zipf's law (Zipf, 1949),

among other instances. Our model, in fact, reduces to Simon's model if mortality is neglected, i.e. for $\mu = 0$. Note that allowing for death events adds a relevant process in the case of surnames, namely, the possibility that a surname disappears if it is borne by a single individual and the individual dies.

Since neither the probability of becoming father of a newborn nor the death probability depend on the individual's age, the present model population can be thought of as ageless. As a consequence, if v and μ are constant, the probability that an individual dies at an age between T and $T + \mathrm{d}T$ is

$$\mathrm{d}p(T) = v\mu \exp(-v\mu T)\,dT, \qquad (4)$$

which implies a life expectancy $\bar{T} = 1/v\mu$. Moreover, the probability that an individual has exactly m children with the same family name during its whole life equals

$$p(m) = (1 - \alpha)\mu[1 + (1 - \alpha)\mu]^{-m-1}, \qquad (5)$$

giving a fertility $\bar{m} = 1/(1 - \alpha)\mu$. The exponential distribution in eqn (5) is in reasonable agreement with actual data collected over relatively short periods, for instance, for the United States (Hull, 1998), but contrasts with the Poissonian distribution found in data integrated over several centuries, as in the case of England (Dewdney, 1986). This discrepancy can presumably be ascribed to long-range variations of real birth and death rates (see Fig. 2).

We are assuming that an individual's surname is inherited from the father. Consequently, the model—as presented above—describes the evolution of the male population only. The same evolution rules apply however if it is assumed that surnames are transmitted with the same probability by either parent. In this case, there is no sex distinction and the model encompasses the whole population. The real situation is in fact intermediate between these two limiting cases: whereas some societies strongly favour inheritance of the surname either from the father or from the mother, other groups allow for a choice between the two possibilities. In some developed western countries, where the surname is preferentially transmitted by the father, opting for the mother's surname has been proposed as a
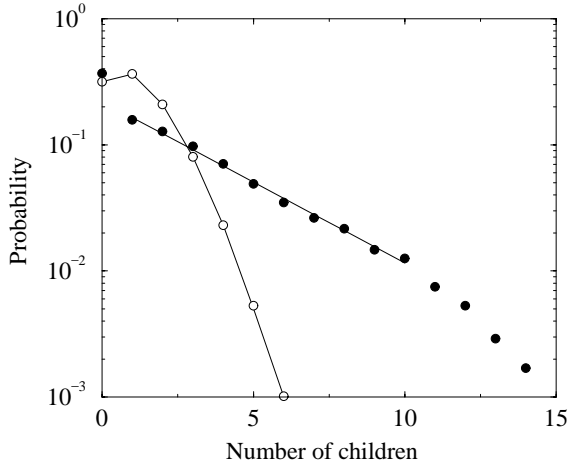
FIG. 2. Probability distribution of the number of children per male in two different populations. Open dots correspond to an average during the period 1350–1986 in England (Dewdney, 1986). The solid line joining the data points is actually a fitting with a Poisson distribution, $P(n) = \exp(-\lambda)\lambda^n/n!$ with average $\langle n \rangle = \lambda = 1.15$. Solid dots are data from the 1920 American census [from Hull (1998); originally compiled by Lotka (1931)]. The solid line fitting the first part of the data is an exponential distribution, $P(n) \propto \exp(-0.3n)$.

method to avoid the persistent extinction of surnames, due to the extremely slow population growth (Legay & Vernay, 1999).

3.1. DISTRIBUTION OF FAMILIES BY SIZE. THE CASE OF $\mu = 0$

As stated above, our model reduces to Simon's model (Simon, 1955) for $\mu = 0$. In this case, $w(s) = 0$ for all s and the total population evolves deterministically, $P(s) = P_0 + s$, since exactly one individual is added to the population at each step. The number of families with exactly i individuals at step s, $n_i(s)$, grows to $n_i(s) + 1$ when an individual is added to a family of size $i - 1$. This occurs with probability $(1 - \alpha)(i - 1)/P(s)$. On the average, thus, the number of families with i individuals varies according to

$$\bar{n}_i(s + 1) = \bar{n}_i(s) + \frac{1 - \alpha}{P(s)}[(i - 1)\bar{n}_{i-1}(s) - i\bar{n}_i(s)]$$

$$(6)$$

for $i > 1$. To the families with only one individual, on the other hand, the positive contribution comes from the creation of new surnames

with probability α. Therefore,

$$\bar{n}_1(s + 1) = \bar{n}_1(s) + \alpha - \frac{1 - \alpha}{P(s)}\bar{n}_1(s). \qquad (7)$$

Note that, since for $\mu = 0$ the stochastic process $w(s)$ becomes trivial, overlines indicate here average over realizations of the stochastic process by which each newborn's surname is chosen.

The system of eqns (6) and (7) can be completely solved for arbitrary initial conditions. In fact, eqn (7) is an autonomous equation for $\bar{n}_1(s)$, whose solution reads

$$\bar{n}_1(s) = \frac{\alpha}{2 - \alpha}(P_0 + s) + \left(n_1(0) - \frac{\alpha}{2 - \alpha}P_0\right)$$

$$\frac{\Gamma(P_0)\Gamma(P_0 + s - 1 + \alpha)}{\Gamma(P_0 + s)\Gamma(P_0 - 1 + \alpha)}, \qquad (8)$$

where $\Gamma(z)$ is the gamma function (Abramowitz & Stegun, 1970). Then, eqn (6) can be used to recursively obtain $\bar{n}_i(s)$ for $i > 1$. For long times, $s \to \infty$, $\bar{n}_1(s)$ is essentially given by the first term on the right-hand side of eqn (8), which grows linearly with s, as the total population. The second term, which contains all the information about the initial condition, can be seen to decrease as $s^{-(1-\alpha)}$. In this limit, the recursion from eqn (6) can be immediately solved. Thus, for $s \to \infty$, we find

$$\bar{n}_i(s) = \frac{\alpha}{2 - \alpha}\frac{\Gamma(i)\Gamma(1/(1 - \alpha) + 2)}{\Gamma((1/(1 - \alpha)) + 1 + i)}(P_0 + s). \qquad (9)$$

The asymptotic number of families of a given size i turns out to be proportional to the total population. For fixed s and sufficiently large values of i, $\bar{n}_i(s)$ decreases as a power law, $\bar{n}_i \propto i^{-z}$, with

$$z = 1 + \frac{1}{1 - \alpha}. \qquad (10)$$

In the limit of small mutation rate, $\alpha \approx 0$, we have $z \approx 2$. As advanced in Section 1, this is the value observed in actual surname distributions. This same result was obtained by Simon (1955) for a special initial condition. Elsewhere (Zanette & Manrubia, 2001) we have already discussed the fact that, though the effects of the initial

condition in Simon's model fade out for long times, transients can strongly depend on the initial distribution $n_i(0)$ (see also Section 4). In our case, this aspect could be relevant in populations where modern surnames have appeared recently, like the Japanese (Miyazima *et al.*, 2000). We shall return to this specific point later.

Equations (6) and (7) imply that, in average, the total number of surnames in the population grows linearly

$$\bar{N}(s) = \sum_{i=1}^{\infty} \bar{n}_i(s) = N_0 + \alpha s \qquad (11)$$

with N_0 the initial number of different surnames. As a function of time, the number of surnames increases exponentially:

$$\bar{N}(t) = N_0 + \alpha P_0[\exp(vt) - 1]. \qquad (12)$$

We have already pointed out that, in contrast, the number of surnames in some real populations at present times is decreasing (Cavalli-Sforza & Feldman, 1981; Legay & Vernay, 1999). Under suitable conditions, adding mortality allows to reproduce this particular behaviour.

3.2. EFFECT OF DEATH EVENTS

Under the action of mortality, the growth of the total population $P(s)$ fluctuates stochastically, according to eqn (1), depending on the occurrence of death events at each evolution step. The evolution of $\bar{n}_i(s)$ can be implemented in two substeps, as follows. First, eqns (6) and (7) are used to calculate the intermediate values

$$\bar{n}_i'(s) = \bar{n}_i(s) + \frac{1-\alpha}{P(s)}[(i-1)\bar{n}_{i-1}(s) - i\bar{n}_i(s)] \quad (13)$$

and

$$\bar{n}_1'(s) = \bar{n}_1(s) + \alpha - \frac{1-\alpha}{P(s)}\bar{n}_1(s) \qquad (14)$$

and the population is updated to $P'(s) = P(s) + 1$. Second, the evolution due to mortality is applied. In terms of the random process $w(s)$ of

eqn (2), we have

$$\bar{n}_i(s+1) = \bar{n}_i'(s) + \overline{\left[\frac{w(s)}{P'(s)}\right]}[(i+1)\bar{n}_{i+1}'(s) - i\bar{n}_i'(s)]$$

$$(15)$$

and $P(s+1) = P'(s) - w(s)$ [cf. eqn (1)].

Naive replacement of the stochastic process $w(s)$ by its average value, $w(s) \to \mu$, in order to obtain a deterministic approximation to the problem, would lead to an equation for which positive solutions cannot be insured. In fact, it is possible to give initial conditions for such deterministic equation which lead to negative values of $\bar{n}_i(s)$ at sufficiently large s and i. A satisfactory deterministic approximation can however be proposed by assuming that the solution $\bar{n}_i(s)$ varies slowly both in s and i. This assumption is generically verified for large populations with smooth initial conditions. In this situation, eqns (13) and (15) admit a continuous approximation in terms of real variables z and y, which replace the integer variables s and i, respectively. Meanwhile, $\bar{n}_i(s)$ is replaced by a continuous function $n(y, z)$.

The continuous approximation can be analyzed at different truncation orders, as discussed in Appendix B. At the first order, the approximate solution to eqns (13) and (15) reads

$$n(y, z) = \alpha \frac{P_0 + (1-\mu)z}{1-\alpha-\mu} y^{-1-(1-\mu)/(1-\alpha-\mu)} \quad (16)$$

for $y < y_T(z)$, and

$$n(y, z) = y_T^{-1} n(y/y_T(z), 0) \qquad (17)$$

for $y > y_T(z)$, with

$$y_T(z) = \left(1 + \frac{1-\mu}{P_0}z\right)^{(1-\alpha-\mu)/(1-\mu)}. \qquad (18)$$

In this first-order continuous approximation, thus, the average number of families $n(y, z)$ exhibits two separated regimes. For large values of the size variable, $y > y_T(z)$, the distribution is essentially determined by the initial condition. At the first evolution steps, where $z \to 0$ and $y_T \to 1$, this regime covers practically all the domain of variable y. As time elapses and y_T grows, however, this regime recedes and is

replaced for $y < y_T$ by a power-law distribution, $n \propto y^{-\zeta}$, with

$$\zeta = 1 + \frac{1-\mu}{1-\alpha-\mu}. \tag{19}$$

Note that, as the exponent z in eqn (10), this new exponent coincides with the observed value, $\zeta \approx 2$, in the relevant limit $\alpha \approx 0$. This result is independent of the death probability μ. For sufficiently long times, the power-law regime will be observed for all the relevant family sizes. All the contribution of the initial condition will become restricted to the domain of the largest families.

The boundary $y_T(z)$ between the two regimes, eqn (18), grows as a power of the ratio between the average current population $P_0 + (1 - \mu)z$ and the initial population P_0. For $\alpha \approx 0$, the exponent is practically equal to unity. Consequently, for the boundary to reach a given value y_0, in such a way that all the families with sizes below y_0 are in the power-law regime, the total population must grow by a factor practically equal to y_0. Taking into account eqn (3), the position of the boundary as a function of real time reads

$$y_T(t) = \exp[v(1-\alpha-\mu)t]. \tag{20}$$

The transient t_0 needed for the power-law regime to develop up to a given size y_0 is therefore logarithmic, $t_0 \propto \ln y_0$.

The average total number of surnames in the continuous approximation is calculated by analogy with eqn (11):

$$\bar{N}(z) = \int_1^\infty n(y,z) \, dy. \tag{21}$$

Using the above first-order solution for $n(y,z)$, we find $\bar{N}(z) = N_0 + \alpha z$ or, as a function of time,

$$\bar{N}(t) = N_0 + \frac{\alpha P_0}{1-\mu}\{\exp[v(1-\mu)t]-1\}, \tag{22}$$

as shown in Appendix B [cf. eqn (12)]. In consequence, within this approximation, the surname diversity always grows exponentially. We point out, however, that this conclusion is valid for sufficiently smooth distributions and for $\mu < 1 - \alpha$, two necessary conditions for the

continuous approximation to apply to our problem. It could therefore be argued that having a decreasing number of surnames, as found in some modern developed societies (Cavalli-Sforza & Feldman, 1981; Legay & Vernay, 1999), requires violation of the above conditions. For instance, an initial condition with sharp variations would violate the smoothness condition during the first evolution stages. A death probability $\mu \approx 1$ would also threaten the validity of the continuous approximation. This is precisely the case of modern developed societies, where the population growth rate is practically vanishing and, as a matter of fact, reaches occasionally negative values—a situation not described by the present model. In Section 3.3 we treat the special case $\mu = 1$ and show that, in this limit, the total number of surnames can in fact decrease.

Though the first-order continuous approximation gives a rather rough description of the solution to our model, as a piecewise function with a moving discontinuity, it provides a quite clear qualitative picture of how the solution behaves. The growth of the power-law regime as the initial condition recedes is in good qualitative agreement with the evolution observed in numerical realizations of the model. A better analytical approximation is obtained from the second-order truncation. Second-order derivatives would in fact introduce diffusive-like effects in the variable y, with the consequent smoothing of the discontinuities of the first-order approximation. The second-order equation, however, cannot be analytically solved for arbitrary initial conditions. Nevertheless, it is possible to give an asymptotic approximation for long times, as

$$n(y,z) = \frac{\alpha P(z)}{1-\alpha-\mu}\left(2\frac{1-\alpha-\mu}{1-\alpha+\mu}\right)^{\zeta-1}$$
$$y^{-1}U\left(\zeta-1,0,2\frac{1-\alpha-\mu}{1-\alpha+\mu}y\right), \tag{23}$$

where $U(a,b,x)$ is the Kummer's function (see Appendix B). With respect to the first-order approximation, eqn (16), this solution predicts a lower value of $n(y,z)$ for small y. For larger family sizes, however, it behaves as a power law, $n(y,z) \propto y^{-\zeta}$ with exactly the same exponent as in

eqn (19). The asymptotic behaviour for large y in the power-law regime is therefore not modified. It must be pointed out that, for small values of α, the profile of $n(y,z)$ given by eqn (23) results to be quite independent of the death probability over a considerable range of values of μ. Only for $\mu \approx 1 - \alpha \approx 1$, where the exponent ζ strongly depends on μ, does $n(y,z)$ change sensibly as μ is varied. We take advantage of this feature in Section 5, where the second-order approximation is compared with actual data of surname distributions.

3.3. THE CASE OF CONSTANT POPULATION, $\mu = 1$

As argued above, the limit $\mu = 1$ is relevant to the discussion of the evolution of surname distributions in some modern developed societies. In fact, this limit corresponds to the case where the population growth rate vanishes, a situation which is closely approached, for instance, in many European countries. In this case, eqn (15) is again deterministic, with $w(s) = 1$ and $P'(s) = P_0 + 1$. The total population at any time is $P(s) = P_0$.

The main difference between this case and that of $\mu < 1$ is that, now, the distribution $\bar{n}_i(s)$ becomes independent of time for asymptotically large times. This feature is in agreement with the fact that the asymptotic distribution for $\mu < 1$ is proportional to the total population $P(s)$. The asymptotic surname distribution $\bar{n}_i^\infty$ is given, for $\mu = 0$, by the recurrence equations

$$\bar{n}_{i+1}^\infty = \frac{[(2-\alpha)P_0 - 2i(1-\alpha)]i\bar{n}_i^\infty - (1-\alpha)(P_0 + 1 - i)(i-1)\bar{n}_{i-1}^\infty}{(i+1)[P_0 - (1-\alpha)(i+1)]} \tag{24}$$

for $i > 1$, and

$$\bar{n}_2^\infty = \frac{[(2-\alpha)P_0 - 2(1-\alpha)]\bar{n}_1^\infty - \alpha P_0^2}{2[P_0 - 2(1-\alpha)(i+1)]}. \tag{25}$$

Note that, since the total population is always P_0, we have $\bar{n}_i^\infty = 0$ for $i > P_0$, as no family can be larger than P_0.

The solution reveals two well-defined regimes, depending on how the product αP_0 compares with unity. For $\alpha P_0 > 1$, the asymptotic distribu-

tion behaves as

$$\bar{n}_i^\infty \approx \frac{\alpha P_0}{i}(1-\alpha)^{i-1} \tag{26}$$

for a vast range of family sizes. Departures from this behaviour are found very close to $i = P_0$ only. For $i > P_0$, in fact, the distribution must vanish. We stress the remarkable difference between the exponential stationary distribution (26) and the long-time power-law solution obtained for $\mu < 1$. In this regime, the stationary total number of surnames is

$$\bar{N}^\infty \approx \frac{\alpha P_0}{1-\alpha}|\ln \alpha|. \tag{27}$$

For $\alpha P_0 < 1$, on the other hand, the distribution behaves as a power law, $\bar{n}_i^\infty \sim i^{-1}$, over practically the whole range of family sizes. Note that the exponent of this power law is in agreement with eqn (16) in the limit $\mu \to 1$. For $i \approx P_0$, however, the distribution deviates from the power law and exhibits a sharp peak. In the limit $\alpha \to 0$, the distribution becomes an isolated peak at $i = P_0$, namely, $\bar{n}_i^\infty = 0$ for $i < P_0$ and $\bar{n}_i^\infty = 1$ for $i = P_0$. Therefore, the total number of surnames is $\bar{N}^\infty = 1$. This special solution describes the well-known case of a closed population with no surname mutations where, by random drift, a single surname survives for asymptotically large times (Cavalli-Sforza & Feldman, 1981). Notice also that this limit corresponds to the birth–death model introduced by Moran (1962) to study probabilities of fixation of alleles when generations overlap.

We conclude that, for a given population P_0, the asymptotic number of surnames can be very small if the mutation probability α is, in turn, small enough. Consequently, in a steady population with many surnames at the initial stage and with a sufficiently low mutation probability, the number of surnames will decrease towards the stationary value as time elapses. Numerical simulations, discussed in detail in the next

section, show that for μ just below unity and small α, an initial transient where the number of surnames decreases temporarily can be observed. Since $\mu < 1$, however, the population grows steadily and, as a consequence of mutations, $\bar{N}$ will also increase in the long run. The situation in modern developed societies is that the population growth rate has been constantly decreasing, to reach values around zero at present times. Starting from a state with a wealth of surnames—due to the combined effect, a few centuries ago, of a high population growth and the frequent appearance of new surnames—these societies have now reached a regime of almost stationary population where the number of surnames decreases. This situation would be reverted if the growth rate could be maintained above zero during substantial periods. Presently, the only mechanism acting in this direction seems to be immigration, which is in addition an effective source of new surnames.

4. Numerical Results

In this section we present results of numerical realizations of our model, in order to compare with the analytical approximations presented in Section 3.2, and to illustrate the behaviour within the regimes where such approximations do not hold. Emphasis is put on the role of the initial conditions, the duration of transients, and the evolution of the number of different surnames. Also, we discuss actual surname distributions of modern populations in the light of our results.

4.1. ROLE OF INITIAL CONDITIONS

First, it is worthwhile to illustrate how the shape of the distribution $\bar{n}_i(s)$ depends on the initial condition $n_i(0)$ (see also Zanette & Manrubia, 2001). We focus the attention on initial distributions of the form $n_i(0) = N_0$ for $i = i_0$, and $n_i(0) = 0$ for $i \neq i_0$, in which there are N_0 families of equal size formed by i_0 individuals each. Consequently, the initial population is $P_0 = i_0 N_0$. In the following, we denote such an initial condition by the pair (i_0, N_0). Figure 3 shows four normalized distributions of family sizes

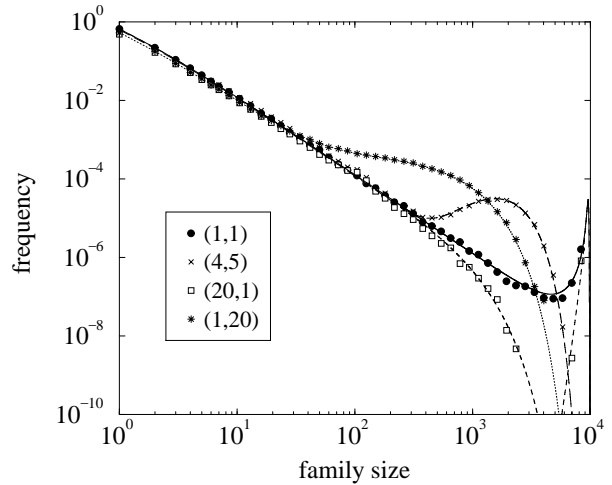$$p_i(s) = \frac{\bar{n}_i(s)}{\sum_i \bar{n}_i(s)} \qquad (28)$$

FIG. 3. Effect of the initial condition on the tail of the surname distribution. The normalized distribution $p_i(s)$ is shown for $s = 10^4$, $\mu = 0$, $\alpha = 10^{-2}$, and four different initial conditions, as specified in the legend. Symbols stand for numerical simulations of the birth–death stochastic model, averaged over 2000 independent realizations. Curves correspond to the numerical solution of the deterministic eqns (6) and (7).

obtained both from the numerical realization of the stochastic birth–death model and from the iterative solution of the deterministic eqns (6) and (7), for $\mu = 0$. We find a very good agreement between both methods and, at the same time, clearly realize the relevant role of the initial condition in determining the profile of the distribution for large family sizes. These same features are found for other values of the death probability.

The solution of the first-order approximation to our model, eqns (16)–(18), predicts that in the zone of small family sizes—i.e. for $y < y_T$ in the continuous variables—the only dependence on the initial condition appears through the quantity P_0. This means that the distribution of family sizes is not sensitive to a variation of i_0 and N_0 as far as their product is kept constant. Note that the three cases $(4, 5)$, $(1, 20)$, and $(20, 1)$ of Fig. 3 share the same value of $P_0 = 20$. For the parameters of the figure, the crossover between the regions of small and large family sizes, given for the continuous variables by eqn (18), should occur at $i_T \approx 470$. Though two of the distributions indeed have the same profile—with the expected power-law decay—up to that value, the distribution corresponding to the initial

condition $(1, 20)$ deviates considerably below i_T. This deviation can be attributed to the fact that the initial condition $(1, 20)$ corresponds to a quite singular distribution, with a high peak at $i = 1$. A continuous approximation for such a distribution is arguably expected to give a poor description of the real situation.

4.2. COMPUTATIONAL MEASUREMENT OF TRANSIENTS

Equation (18) defines, within the first-order approximation, the boundary that separates the asymptotic regime and the zone dominated by the initial condition. Alternatively, for a fixed family size, it can be used to determine the transient before the asymptotic distribution builds up at that size. In order to test this aspect of the first-order approximation, we have devised an independent computational method to evaluate the point at which the cut-off between the two regimes actually takes place. According to eqns (16) and (28), the asymptotic normalized distribution of family sizes is $p_i^\infty \equiv p_i(s \to \infty) \approx (\zeta - 1)i^{-\zeta}$. In numerical realizations, the distribution is expected to adopt values very close to p_i^∞ for large i and large enough s. In the simulations, we fix a certain family size i_T and stop the calculation at the step s_T when the measured value of p_{i_T} satisfies, for the first time, $|p_{i_T} - p_{i_T}^\infty| < \Delta$. The results presented in the

following correspond to the choice $i_T = 10^2$ and $\Delta = 10^{-1}$. Since, as shown in Section 4.1, the agreement between the numerical simulations of the stochastic birth–death model and the iterative solution of the corresponding average equations is very good, we use this second description to measure the transient s_T.

We focus the analysis on the dependence with the initial conditions, and keep the values of $\alpha = 10^{-2}$ and $\mu = 0$ fixed. As far as $\alpha \approx 0$, the results are qualitatively the same for other values of μ. We consider initial conditions of the form (i_0, N_0), as defined in Section 4.1. Figure 4 shows the transient s_T measured from the numerical solution to eqns (6) and (7) according to the criterion introduced above, as a function of i_0 and N_0. On the one hand, for small N_0 and, especially, for small i_0, s_T varies quite irregularly. On the other hand, we see that for larger values of i_0 and N_0, say $i_0 > 10$ and $N_0 > 5$, s_T exhibits a well-defined linear dependence with both parameters. In this regime, the transient is well approximated by

$$s_T = ai_0N_0. \tag{29}$$

Linear fitting of s_T as a function of i_0 and N_0 for the data shown in Fig. 4 yields $a = 237 \pm 4$. Note that eqn (29) implies that, for large i_0 and N_0 the transient length is proportional to the initial population $P_0 = i_0N_0$. This is to be compared
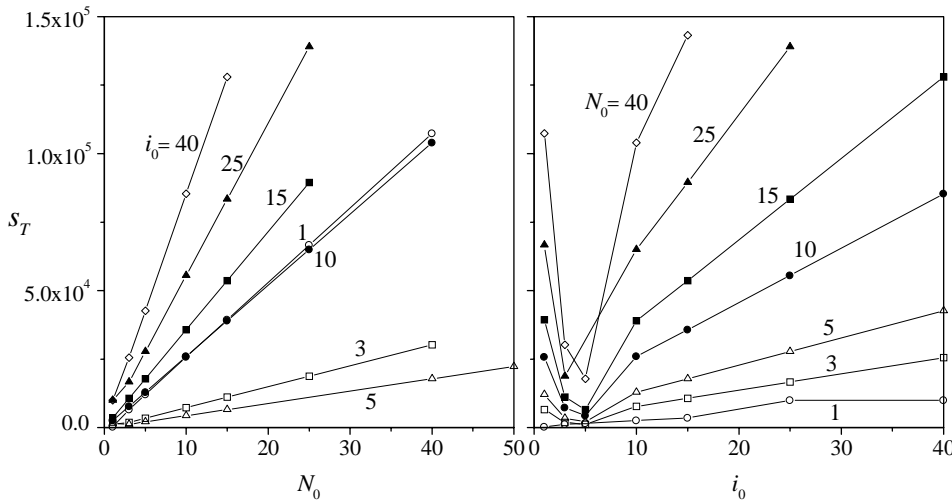

FIG. 4. Numerically measured transient s_T as a function of the initial parameters i_0 and N_0, for $\alpha = 10^{-2}$ and $\mu = 0$. The parameters that define the transient (see text) are $i_T = 10^2$ and $\Delta = 10^{-1}$.

with the analogous result within the first-order approximation, eqn (18), which implies

$$s_T = \frac{P_0}{1 - \mu}\left[i_T^{(1-\mu)/(1-\alpha-\mu)} - 1\right].\qquad(30)$$

The first-order approximation predicts then a proportionality between s_T and P_0—independently of the values of α and μ—in full agreement with the numerical results. We point out, however, that the corresponding proportionality coefficients cannot be directly compared. In fact, the coefficient a in eqn (29) is evaluated following the computational definition of the transient, and thus depends on the parameter Δ, which has no correlate in the analytical approach. Note, nevertheless, that for $\alpha \approx 0$ and $\mu = 0$ eqn (30) predicts a coefficient approximately equal to $i_T = 100$, which is of the same order as the numerical result.

4.3. EVOLUTION OF SURNAME DIVERSITY

As discussed in Section 3, the evolution of the average number $\bar{N}(s)$ of different surnames is driven in our model by two competing mechanisms. The diversity increases due to the appearance of new surnames at rate α, and surnames borne by single individuals disappear when the individual in question dies. Within the first-order continuous approximation to our model, for $\mu < 1$, the average number of different surnames increases linearly with s as $\bar{N}(s) = N_0 + \alpha s$ (see Section 3.2). In this approximation, the death probability μ plays a role in the variation of diversity as a function of time, eqn (22). On the one hand, increase of the surname diversity in a steadily growing population is generally expected when new surnames are created at a constant rate. On the other, it is possible to conceive special situations where the number of different surnames should temporarily decrease, violating the first-order approximation. Imagine, for instance, that the initial population consists of an ensemble of families with only one individual each. For moderate values of μ and small α, the initial stage would be characterized by the death of some individuals, with the consequent disappearance of their surnames, and no significative appearance of new surnames. Since the total population grows, how-
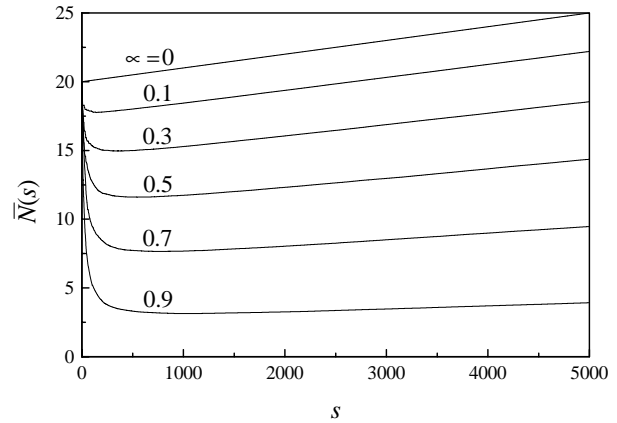

FIG. 5. Evolution of the average number of different surnames for the initial condition $(1, 20)$, with $\alpha = 10^{-3}$ and different values of the death probability μ.

ever, $\bar{N}(s)$ will eventually attain a minimum and, from then on, will increase.

To illustrate this situation, we numerically solve eqns (13)–(15) for the initial condition $(1, 20)$, with 20 families of one individual each. Note that this is the initial condition for which we detected the largest deviations from the first-order approximation in Section 4.1. Figure 5 shows the evolution of $\bar{N}(s)$ for different values of μ. As expected, an initial transient where the surname diversity drops is found for $\mu > 0$. The transient is longer and the minimum in $\bar{N}(s)$ is deeper as μ grows. In all cases, however, the subsequent growth of $\bar{N}(s)$ is clearly seen. Note that the slope of this growth depends slightly on μ, a feature not predicted by the first-order continuous approximation.

These long transients where (for large μ and suitable initial conditions) the number of different surnames is expected to decrease could be relevant to the description of modern populations with declining diversity (Cavalli-Sforza & Feldman, 1981; Legay & Vernay, 1999). A realistic description of this situation should however take into account that, in recent times, the relative values of birth and death rates in actual populations have considerably changed. The general trend to population growth observed worldwide in the 19th century has by now been reversed in many developed areas, such as in Europe, where the total population is practically stationary. In our model, this corresponds to an increment in the value of μ to
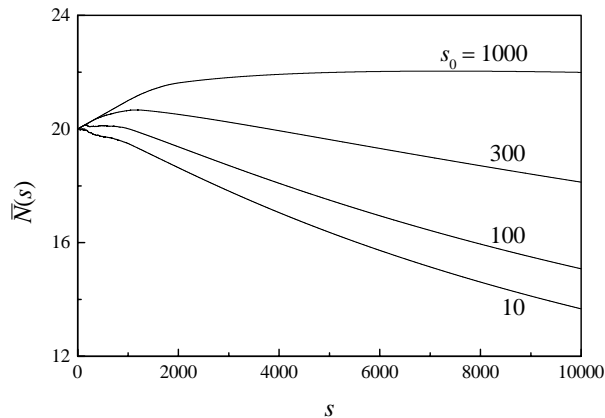
FIG. 6. Evolution of the diversity of surnames under variation of the death probability μ, for $\alpha = 10^{-3}$. See main text for further explanations.

values close to unity. In particular, the death probability must be allowed to vary with time. Additionally, if the description is expected to encompass the periods where the appearance of new surnames was frequent—specifically, the Middle Ages in the case of Europe—α should also change during the evolution.

As a qualitative demonstration of the effect of varying the model parameters with time, we focus the attention on a change of μ from a low value to a high value. For the initial condition $(1, 20)$ we fix $\mu = 0$ during the first s_0 evolution steps. Then, μ is left to increase linearly with s, such that it reaches unity after s_1 additional steps. In the numerical calculations shown in Fig. 6 we fix $s_1 = 10^3$ and consider several values of s_0. For $s < s_0$, the number of different surnames increases (cf. Fig. 5). Then, as the death probability grows, $\bar{N}(s)$ attains a maximum and begins to decrease. For asymptotically large times, it is expected to approach a stationary value, as predicted for the case $\mu = 1$ in Section 3.3. Interestingly, $\bar{N}(s)$ responds faster to the effects of growing μ for smaller values of s_0, where the contribution of the initial condition is still important during the variation of the death probability. The asymptotic surname distribution seems to be, in this sense, quite robust to the action of varying μ. This feature should be related to the fact that the asymptotic distribution is rather insensitive to the value of μ, as discussed at the end of Section 3.2.

5. Comparison of the Model with Field Data

Finally, we compare some of the analytical results for our model, derived under several approximations in Section 3.2, with actual data from three modern populations. Specifically, we focus the attention on the second-order approximation, eqn (23), for the asymptotic distribution of surnames in the range of small family sizes, where the effect of the initial condition is negligible. In fact, it is virtually impossible to extract information on the distribution of surnames in historical times from the data presently available. We recall that eqn (23) correctly describes the asymptotic power law $n(y, z) \propto y^{-\zeta}$ with $\zeta \approx 2$, as observed in real data. The validity of our second-order approximation is therefore to be especially evaluated in the region of very small family sizes, where the distribution differs from the power law.

Our three data sets were obtained from surname counts in telephone books. They correspond to (i) the almost $350\,000$ different surnames of the whole 1996 Argentine telephone book, (ii) the 6400 surnames beginning by A in the 1996 Berlin telephone book, and (iii) the surnames in five Japanese cities, with populations ranging between 2×10^3 and 2×10^5, reported by Miyazima *et al.* (2000). Let us point out that the three populations involved here have considerably different demographic history. The modern Argentine population has predominantly European ancestors, who immigrated mainly in the period 1880–1915 and just after the World War II. Their surname distribution has therefore to be considered as a combined sample from the countries that contributed the immigrants. From the times of the largest immigration waves to the present, both the birth and death rates have substantially changed. As for Berlin, this city was practically abandoned in the late stages of World War II and subsequently repopulated with a mixture of the ancient inhabitants and newcomers from other cities of Germany. In this case, the surname distribution has consequently resulted from a combination of several German regions. The modern population in Berlin, in addition, presents the particularity of having been artificially separated into two practically immiscible

groups during several decades, ending in 1989. Since European surnames appeared mostly during the Middle Ages (Legay & Vernay, 1999), the populations that contributed surnames to both Argentina and Berlin are expected to have developed the asymptotic surname distribution on a substantial range of family sizes. This situation contrasts with the case of modern Japanese surnames, that were originated some 120 years ago (Miyazima *et al.*, 2000). In Japan, moreover, the contribution of immigration in the relevant period should have been considerably less important than in Berlin and Argentina.

Explicit evaluation of eqn (23) requires fixing the values of the mutation rate α and the death probability μ. We know that the mutation rate, i.e. the probability that a new individual acquires a surname not previously present in the population, is very low for any modern society. Since, as far as $\alpha \neq 0$, the limit of eqn (23) for $\alpha \rightarrow 0$ is well defined, fixing any sufficiently small value for α gives a correct description of $n(y, z)$. In our comparison with real data, we have taken $\alpha = 10^{-3}$. As for the death probability μ, unfortunately, we have found it impossible to fix a reliable value. For all real populations, mortality has considerably changed during the periods relevant to the evolution of surname frequencies. Moreover, since μ measures the relative frequency of death and birth events—the latter also including, in real populations, arrival of new individuals by immigration—an evaluation of the death probability should also involve a detailed description of immigration effects. On the other hand, as discussed at the end of Section 3.2, the asymptotic profile of $n(y, z)$, given by eqn (23), is practically independent of the death probability as far as μ is not close to unity. We therefore decided to fit the field data with eqn (23) allowing μ to vary in order to get the best approximation in the relevant zone of small family sizes.

Figure 7 shows the three data sets and the corresponding fittings with eqn (23). Fittings have been optimized in the domain of small family sizes, where our analytical approximation is expected to hold. In the case of Argentina, the agreement with real data is excellent up to family sizes above $i = 100$. Only for $i > 200$ a systematic
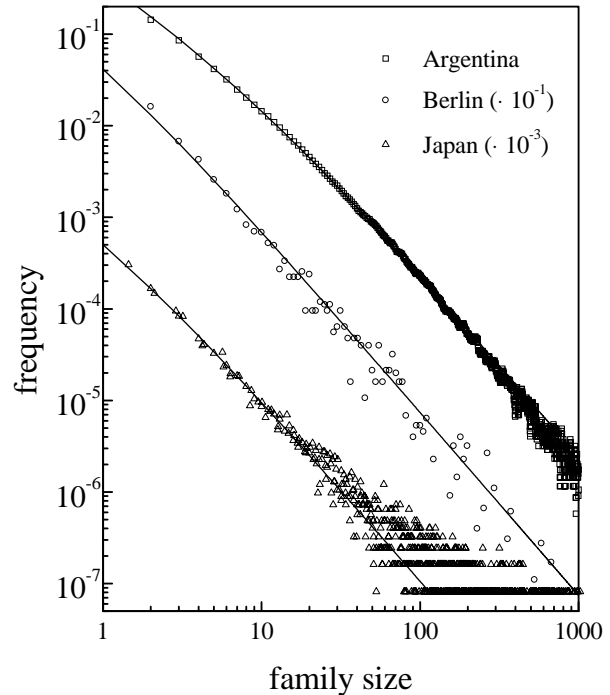

FIG. 7. Frequency of family names as a function of the family size for three modern populations. Dots correspond to real data from the sources quoted in the text. Curves stand for the fitting of each data set with eqn (23), taking $\mu = 0.7$ for Argentina and $\mu = 0.2$ for Berlin and Japan. For clarity, the data for Berlin and Japan have been shifted by a constant factor.

deviation is observed, where the analytical result overestimates the frequency. According to our discussion in Section 3.2, this deviation would be the remnant contribution of initial conditions. For the Berlin data, the statistics are poorer. However, it is clear that the analytical approximation fits the data with good precision in the whole range shown here. In the case of the Japanese data the fitting is very good for relatively small family sizes, $i < 10$, but, on the other hand, noticeable systematic deviations appear for $i > 20$. This agrees with our expectation on the effect of initial conditions. In the 120 years elapsed from the appearance of modern Japanese family names, the population in the cities from which the present data were obtained has increased by a factor of, at most, order 10. Therefore, according to eqn (18), initial conditions should contribute to the distribution for relatively small family sizes, $i > 10$, just as observed.

6. Discussion

We have analysed a birth–death model with overlapping generations, in order to study several statistical properties of monoparental inheritance in large populations. We have focused the analysis on a cultural trait, namely, the distribution of surnames, taking advantage of the availability of big corpora of real data. However, the model applies generally to biological traits associated with non-recombining alleles. Our model has two parameters, namely, the probability that a new individual carries a surname not present in the population, α, and the average number of death events per birth μ. For any value of $\alpha > 0$ the system attains a broad, stationary distribution of surname diversity. If $\mu < 1$ the total population grows exponentially in time and so does the total number of different surnames. The marginal case $\mu = 1$ corresponds to constant total population and, on average, to constant surname diversity for asymptotically large times.

Our analytical results for the stationary distribution of surnames frequency are in good agreement with field data for modern human populations in different countries. Through an analysis of the transient time required for this distribution to reach its asymptotic shape, we have shown that some deviations observed in real data might actually reflect the composition of the founder population. This result has implications in the study of polyphyletism. Indeed, if the same surname can have multiple origins and thus the individuals carrying it are not always phylogenetically related, this will affect the shape of the surname distribution. In particular, it is not difficult to estimate the time when surnames originated in a population (using historical records) or, in the biological counterpart, when a mutant allele first appeared (through molecular clock analysis). Then, the approximate cut-off until which the stationary distribution follows the asymptotic shape can be calculated. If the distribution underestimates the frequency of values larger than the cut-off, the system is mainly polyphyletic. If it overestimates that frequency, then the simultaneous appearance of many individual carrying different surnames took place in the past.

The strong resemblance between the cultural inheritance of the surname and the biological process in which non-recombining neutral alleles are passed to offspring has justified to apply results from field data in the former case to the latter (Barrai et al., 1996). In the few cases where data on genetic diversity was available, it was possible to retrieve information on past populations by comparing both sets of data (Sykes & Irven, 2000). A specific example comes from the small island of Tristan da Cunha, where the fact that its 300 inhabitants represent only seven surnames and five mitochondrial lineages reflects without doubt the small size of the founder population (Soodyall et al., 1997).

Our results could be applied to population genetics under some hypotheses. Indeed, we are assuming that the number of different alleles at a given locus is practically infinite [this is analogous to the assumption made by Kimura & Crow (1964), in their model of infinitely many alleles], since the possibility of backward mutations is discarded. Nonetheless, this factor could be accounted for just by lowering the value of α, because this process implies that the "new" mutant is in fact identical to one of the forms present in the population. Since, as long as α is small, our results are not sensibly modified, we could also work with a finite but large number R of different surnames (equivalently, large genetic polymorphism), and use the same model as long as the current diversity is lower than R. We have also assumed that the death rate is constant during the lifetime of individuals, while it is known that the life expectancy depends not only on the age of individuals, but also varies as a function of time (Vaupel et al., 1998). A more realistic model of human inheritance could be constructed by taking into account the variation of μ along the lifetime of each individual.

The evolution of surname diversity follows truly neutral evolution: family names do not fulfill any practical purpose but identifying the lineage of each individual. They cannot be selected for through natural selection. Hence, as we have shown, their statistical distribution closely follows the predictions of a neutral model of monoparental inheritance. It would be interesting to test our theoretical results against the genetic diversity of a large population sample.

Unfortunately, however, data on frequencies of allelic polymorphisms are still scarce to carry out a massive study like the one presented here for actual surname distributions. Moreover, one should make sure that the different haplotypes in the sample do not confer any advantage to the individuals, in which case positive feedbacks and deviations from neutral statistics would be expected. Different tests have been proposed in the literature to detect deviations from neutrality (Nielsen, 2001; Fu & Li, 1993), and a number of neutral haplotypes have been positively identified (Sánchez Mazas et al., 1994; Stenoien, 1999). Hopefully, enough data will be available in a near future to calculate reliable diversity distributions in human populations.

REFERENCES

ABRAMOWITZ, M. & STEGUN, I. A. (1970). *Handbook of Mathematical Functions.* New York: Dover.

BARRAI, I., SCAPOLI, C., BERETTA, M., NESTI, C., MAMOLINI, E. & RODRIGUEZ-LARRALDE, A. (1996). Isonymy and the genetic structure of Switzerland. I. The distribution of surnames. *Ann. Hum. Biol.* **23,** 431–455.

BURLANDO, B. (1990). The fractal dimension of taxonomic systems. *J. theor. Biol.* **146,** 99.

BURLANDO, B. (1993). The fractal geometry of evolution. *J. theor. Biol.* **163,** 161.

CAVALLI-SFORZA, L. L. & FELDMAN, M. W. (1981). *Cultural Transmission and Evolution: A Quantitative Approach.* Princeton: Princeton University Press.

CAVALLI-SFORZA, L. L., FELDMAN, M. W., CHEN, K. H. & DORNBUSCH, S. M. (1982). Theory and observation in cultural transmission. *Science* **218,** 19–27.

CONSUL, P. C. (1991). Evolution of surnames. *Int. Stat. Rev.* **59,** 271–278.

DARWIN, C. (1871). *The Descent of Man and Selection in Relation to Sex.* London: Murray.

DEWDNEY, A. K. (1986). Computer recreations: branching phylogenies of the Paleozoic and the fortunes of English family names. *Sci. Am.* **254,** 12–16.

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

FOX, W. R. & LASKER, G. W. (1983). The distribution of surname frequencies. *Int. Stat. Rev.* **51,** 81–87.

FU, Y. X. & LI, W. H. (1993). Statistical tests of neutrality of mutations. *Genetics* **133,** 693–709.

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

GALTON, F. & WATSON, H. W. (1874). On the probability of the extinction of families. *J. Roy. Anthropol. Inst.* **4,** 138–144.

GOMES, M. A. F., VASCONCELOS, G. L., TSANG, I. J. & TSANG, I. R. (1999). Scaling relations for diversity of languages. *Physica A* **271,** 489–495.

GREENBERG, J. H. (1992). Preliminaries to a systematic comparison between biological and linguistic evolution. In: *The Evolution of Human Languages* (Hawkins J. A. & Gell-Mann M., eds) Reading: Addison-Wesley.

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

HARRIS, TH. E. (1963). *The Theory of Branching Processes.* Berlin: Springer.

HULL, D. M. (1998). A reconsideration of Galton's problem (using a two-sex population). *Theor. Popul. Biol.* **54,** 105–116.

ISLAM, M. N. (1995). A stochastic model for surname evolution. *Biom. J.* **37,** 119–126.

KARLIN, S. & MCGREGOR, J. (1967). The number of mutant forms maintained in a population. *Proc. fifth Berkeley Symp. Math. Stat. Prob.* **4,** 415–438.

KIMURA, M. & CROW, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics* **49,** 725–738.

LANGE, K. (1981). Minimum extinction probability for surnames and favorable mutations. *Math. Biosci.* **54,** 71–78.

LEGAY, J. M. & VERNAY, M. (1999). L' instructive histoire des noms de familie. *Pour Sci.* **255,** 58–65.

LOTKA, A. J. (1931). Population analysis—the extinction of families—I. *Washington Acad. Sci.* **21,** 377–380, 453–459.

MIYAZIMA, S., LEE, Y., NAGAMINE, T. & MIYAJIMA, H. (2000). Power-law distribution of family names in Japanese societies. *Physica A* **278,** 282–288.

MORAN, P. A. P. (1962). *The Statistical Processes of Evolutionary Theory.* Oxford: Clarendon Press.

NIELSEN, R. (2001). Statistical tests of selective neutrality in the age of genomics. *Heredity* **86,** 641–647.

PANARETOS, J. (1989). On the evolution of surnames. *Int. Stat. Rev.* **57,** 161–167.

PIELOU, E. C. (1969). *An Introduction to Mathematical Ecology.* New York: Wiley.

POORE, M. E. D. (1968). Studies in Malaysian rainforest. *Ecology* **56,** 143–196.

RANNALA, B. (1997). Gene genealogy in a population of variable size. *Heredity* **78,** 417–423.

RUHLEN, M. (1992). An overview of genetic classification. In: *The Evolution of Human Languages* (Hawkins, J. A. & Gell-Mann M., eds) Reading: Addison-Wesley.

SÁNCHEZ MAZAS, A., BUTLER BRUNNER, E., EXCOFFIER, L., GHANEM, N., SALEM, M. B., BREGUET, G., DARD, P., PELLEGRINI, B., TIKKANEN, M. J., LANGANEY, A., LEFRANC, G. & BUTLER, R. (1994). New data for AG haplotype frequencies in caucasoid populations and selective neutrality of the AG polymorphism. *Hum Biol* **66,** 27–48.

SIMON, H. A. (1955). On a class of skew distribution functions. *Biometrika* **42,** 425–440.

SOODYALL, H., JENKINS, T., MUKHERJEE, A. DU TOIT, E., ROBERTS, D. F. & STONEKING, M. (1997). The founding mitochondrial DNA lineages of Tristan da Cunha Islanders. *Am. J. Phys. Anthropol.* **104,** 157–166.

SOLÉ, R. V. & ALONSO, D. (1998). Random walks, fractals, and the origin of rainforest diversity. *Adv. Complex Systems* **1,** 203–220.

STENOINEN, H. K. (1999). Are enzyme loci selectively neutral in haploid populations of nonvascular plants? *Evolution* **53**, 1050–1059.

SYKES, B. & IRVEN, C. (2000). Surnames and the Y chromosome. *Am. J. Hum. Genet.* **66**, 1417–1419.

VAN KAMPEN, N. G. (1981). *Stochastic Processes in Physics and Chemistry.* Amsterdam: North-Holland.

VAUPEL, J. W., CAREY, J. R., CHRISTENSEN, K., JOHNSON, T. E., YASHIN, A. I., HOLM, N. V., IACHINE, I. A., KANNISTO, V., KHAZAELI, A. A., LIEDO, P., LONGO, V. D., ZENG, Y., MANTON, K. G. & CURTSINGER, J. W. (1998). Biodemographic trajectories of longevity. *Science* **280**, 855–860.

YASUDA, N., CAVALLI-SFORZA, L. L., SKOLNICK, M. & MORONI, A. (1974). The evolution of surnames: An analysis of their distribution and extinction. *Theor. Popul. Biol.* **5**, 123–142.

ZANETTE, D. H. (2001). Statistical regularities in the taxonomic classification of human languages. *Adv. Complex Systems* **4**, 281–286.

ZANETTE, D. H. & MANRUBIA, S. C. (2001). Vertical transmission of culture and the distribution of family names. *Physica A* **295**, 1–8.

ZIPF, G. K. (1949) *Human Behavior and the Principle of Least Effort.* Cambridge: Addison-Wesley.

Appendix A

Average Evolution of the Total Population

As explained in the main text, at each step, the total population $P(s)$ in our model either increases by 1 with probability $1 - \mu$, due to the occurrence of a birth event and no death event, or remains constant with probability μ, due to the occurrence of both a birth and a death event. Equations (1) and (2) quantify the corresponding stochastic process.

Since the total population changes during the evolution, the time interval δt to be associated with each step—which corresponds to the interval between consecutive births—must also change. In fact, the birth frequency is proportional to the total population, so that δt is inversely proportional to $P(t)$. We write, at step s,

$$\delta t(s) = \frac{1}{vP(s)}, \qquad (A.1)$$

where the frequency v fixes time units.

In consequence, the real-time average variation in $P(s)$ can be obtained from

$$\frac{d\bar{P}}{dt} \approx \overline{\left[\frac{P(s+1) - P(s)}{\delta t(s)}\right]}$$

$$= v\overline{[1 - w(s)]P(s)} = v(1 - \mu)\bar{P}, \quad (A.2)$$

where overlines indicate average over realizations of the stochastic process $w(s)$. To obtain this average evolution equation we have used eqn (1), and have taken into account that $w(s)$ and $P(s)$ are independent stochastic processes at each step s, so that $\overline{w(s)P(s)} = \bar{w}(s)\bar{P}(s) = \mu\bar{P}(s)$.

If v is identified with the birth rate per individual and unit time, the product $v\mu$ is the mortality rate. For constant v and μ, the solution to eqn (A.2) is

$$\bar{P}(t) = P_0 \exp[v(1 - \mu)t], \qquad (A.3)$$

with P_0 the initial population.

Appendix B

Continuous Approximation to the Distribution of Family Sizes

Under the action of both birth and death events, the evolution of the average number of families of size i, $\bar{n}_i$, is given by eqns (13) and (15). These equations can be approximately solved assuming that the solution varies slowly on s and i, so that these two discrete variables admit replacement by continuous variables z and y, respectively. Accordingly, $\bar{n}_i(s)$ is replaced by a continuous function $n(y,z)$. The approximation is based on the expansions $\bar{n}_{i\pm1}(s) \equiv n(y,z) \pm \partial_y n(y,z) + \cdots$ and $\bar{n}_i(s+1) \equiv n(y,z) + \partial_z n(y,z) + \cdots$, to be introduced in eqns (13) and (15) at different truncation orders.

B.1. FIRST-ORDER APPROXIMATION

To the first order, we obtain the differential equation

$$\frac{\partial n}{\partial z} + \frac{1 - \alpha - \mu}{P_0 + (1 - \mu)z}\frac{\partial}{\partial y}(yn) = 0 \qquad (B.1)$$

for $y > 1$. The contribution for $y = 1$, given by eqn (14), can be incorporated to (B.1) either as a boundary condition or as a singular inhomogeneity in the equation itself. The latter yields

$$\frac{\partial n}{\partial z} + \frac{1 - \alpha - \mu}{P_0 + (1 - \mu)z}\frac{\partial}{\partial y}(yn) = \alpha\delta(y - 1), \quad (B.2)$$

where $\delta(x)$ stands for the Dirac delta distribution. In the following, we obtain and analyse the solution to this equation.

Introducing the auxiliary variables

$$\xi = \ln\left(1 + \frac{1-\mu}{P_0}z\right), \quad \eta = \ln y, \quad \text{(B.3)}$$

eqn (B.2) becomes

$$\frac{\partial f}{\partial \xi} + \frac{1-\alpha-\mu}{1-\mu}\frac{\partial f}{\partial \eta} = \frac{\alpha}{1-\mu}\delta(\eta). \quad \text{(B.4)}$$

with $f(\eta,\xi) \equiv yn(y,z)$. This is a one-dimensional wave equation, with "spatial" variable η and "temporal" variable ξ. It describes shape-preserving advection of the "density" f with constant velocity $v = (1-\alpha-\mu)/(1-\mu)$, subject to the action of a point source of intensity $\alpha/(1-\mu)$ at $\eta = 0$. In our problem, thus, this equation makes sense for $v > 0$ ($\mu < 1-\alpha$) since η must be non-negative ($y \equiv i \geqslant 1$). The general solution for $\eta \neq 0$ is given by an arbitrary combination of functions of the form $f(\eta,\xi) = f_0(\eta - v\xi)$, where f_0 is in principle arbitrary. The combination must be chosen in such a way that the boundary and initial conditions are satisfied. In our case, this is achieved with a combination of two such functions. In the original variables, the solution is given piecewise as

$$n(y,z) = \alpha \frac{P_0 + (1-\mu)z}{1-\alpha-\mu} y^{-1-(1-\mu)/(1-\alpha-\mu)} \quad \text{(B.5)}$$

for $y < y_T(z)$, and

$$n(y,z) = y_T^{-1} n(y/y_T(z), 0) \quad \text{(B.6)}$$

for $y > y_T(z)$. Here, $n(y,0)$ is the initial distribution. The transition point between the two pieces is located at

$$y_T(z) = \left(1 + \frac{1-\mu}{P_0}z\right)^{(1-\alpha-\mu)/(1-\mu)}. \quad \text{(B.7)}$$

These expressions give the first-order continuous approximation $n(y,t)$ to the distribution of families by size.

The average total number of surnames in the continuous approximation is defined in (21). For the first-order approximation we

find

$$\bar{N}(z) = \int_1^\infty n(y,z)\,\mathrm{d}y$$

$$= \alpha\frac{P_0 + (1-\mu)z}{1-\alpha-\mu}\int_1^{y_T} y^{-\zeta}\mathrm{d}y$$

$$+ \int_{y_T}^\infty n(y/y_T,0)\frac{\mathrm{d}y}{y_T}. \quad \text{(B.8)}$$

The last integral is clearly equal to the initial number of surnames, N_0, as may be immediately realized by the change of variables $y/y_T \to y$. Explicit calculation of the first integral shows that $\bar{N}(z) = N_0 + \alpha z$, i.e. exactly the same result as for the case of $\mu = 0$, eqn (11). As a function of time, we have

$$\bar{N}(t) = N_0 + \frac{\alpha P_0}{1-\mu}\{\exp[v(1-\mu)t] - 1\}. \quad \text{(B.9)}$$

B.2. SECOND-ORDER APPROXIMATION

Truncation to the second order in the continuous approximation to eqns (13) and (15) yields

$$\frac{\partial n}{\partial z} + \frac{1-\alpha-\mu}{P_0 + (1-\mu)z}\frac{\partial}{\partial y}(yn)$$

$$+ \frac{1-\alpha+\mu}{2[P_0 + (1-\mu)z]}\frac{\partial^2}{\partial y^2}(yn)$$

$$= \alpha\delta(y-1). \quad \text{(B.10)}$$

Unfortunately, this equation cannot be analytically solved for arbitrary initial conditions. However, a particular solution can be found in the form of a separate function, $n(y,z) = h(y)P(z) = h(y)[P_0 + (1-\mu)z]$. Comparison with eqns (9) and (16) suggests that this particular solution will correspond to the long-time asymptotic evolution. It reads

$$n(y,z) = \frac{\alpha P(z)}{1-\alpha-\mu}\left(2\frac{1-\alpha-\mu}{1-\alpha+\mu}\right)^{\zeta-1}$$

$$y^{-1}U\left(\zeta-1,0,2\frac{1-\alpha-\mu}{1-\alpha+\mu}y\right), \quad \text{(B.11)}$$

where $U(a,b,x)$ is the logarithmic Kummer's function (Abramowitz & Stegun, 1970).