ELSEVIER

BioSystems 66 (2002) 179–191

BioSystems

www.elsevier.com/locate/biosystems

Frequency distributions from birth, death, and creation processes

David L. Bartley [a,*], Trevor Ogden [b], Ruiguang Song [c]

[a] National Institute for Occupational Safety and Health, 4676 Columbia Parkway, Cincinnati, OH 45226, USA
[b] 40 Wilsham Road, Abingdon, Oxfordshire OX14 5LE, UK
[c] Centers for Disease Control and Prevention, 1600 Clifton Road, NE, Atlanta, GA 30333, USA

Received 8 August 2001; received in revised form 28 June 2002; accepted 5 July 2002

Abstract

The time-dependent frequency distribution of groups of individuals versus group size was investigated within a continuum approximation, assuming a simplified individual growth, death and creation model. The analogy of the system to a physical fluid exhibiting both convection and diffusion was exploited in obtaining various solutions to the distribution equation. A general solution was approximated through the application of a Green's function. More specific exact solutions were also found to be useful. The solutions were continually checked against the continuum approximation through extensive simulation of the discrete system. Over limited ranges of group size, the frequency distributions were shown to closely exhibit a power-law dependence on group size, as found in many realizations of this type of system, ranging from colonies of mutated bacteria to the distribution of surnames in a given population. As an example, the modeled distributions were successfully fit to the distribution of surnames in several countries by adjusting the parameters specifying growth, death and creation rates.
Published by Elsevier Science Ireland Ltd.

Keywords: Birth; Death; Creation; Surname; Skewed distribution; Zipf; Pareto; Power-law; Kummer; Bateman

1. Introduction

In this paper we present a detailed model describing the effects of growth and death on the frequency distributions of various-sized labeled groups of individuals. The effect of the introduction of (new) single-sized groups is emphasized. With the decay of transients associated with initial conditions, the distribution of small-sized groups becomes quickly dominated by the rate of influx of such *singleton* groups.

Note the rather peculiar use of the word "group". A group can, for example, have a single individual as a member. The individual can be termed a *singleton*, and the group, a *singleton group*.

Analysis is carried out within a continuum approximation, which evokes analogies to fluid convection and diffusion. This provides both calculational simplification as well as physical intuition regarding the system. For example,

* Corresponding author

0303-2647/02/$ - see front matter. Published by Elsevier Science Ireland Ltd.
PII: S0303-2647(02)00053-9

domination by the singleton influx rate will be found explainable in terms of the convection of the small groups towards the larger.

The general type of system considered in this paper is likely to have many applications, most of them enumerated by Simon (1955) with highly skewed distributions approximating the *Zipf–Pareto Law* (Johnson et al., 1995), falling off as a power law. For example, bacterial growth, death and mutation follow the assumptions adopted below, at least over limited time spans. Seemingly disparate, word usage in a developing novel is undoubtedly related. As an example application, we have considered the distribution of surnames in several countries. In this case, *individuals* refer to individual families, and a *group*, to families with a common surname.

One result of this research may be a better understanding of a related theory developed by Simon (1955) and later criticized by Mandelbrot (1959). Simon's theory resulted in power-law functions in an asymptotic (large group size) limit. For example, if $f(x)$ is the number of surnames which occur x times in a population, then $f(x)$ would be roughly proportional to $x^{-\nu}$, where the exponent ν is a constant. Then the number $xf(x)$ of people with names occurring x times has a power-law distribution as well. With a strictly power-law distribution, ν must be greater than 2 in order for $xf(x)$ to be summable, i.e. if the total number of families is finite. With many systems, however, some described by Simon, finite data are very often found to be well approximated with the exponent less than 2. This apparent dilemma is explicated within the present model.

The plan for the remainder of the paper is as follows. A time-dependent model with plausible birth-and-death processes is first derived in the discrete case and then is approximated as a continuum model. The resulting partial differential equation is then solved in several approximations. Specific exact solutions are also given. Although the mathematics may be difficult for some readers to follow, we have tried to write the paper in a way that those who wish to may skim past equations and yet get an understanding of the material. One of the more difficult sections (approximation of general solutions using Green's functions) has been relegated to an appendix, as the material is essentially a starting point for future work.

Finally, by way of illustration, the exact solutions are applied to the study of surname frequency distribution. Paternally inherited surnames have been widely studied as simulators of alleles of genes transmitted through the Y chromosome. One feature that has attracted attention is the frequency distribution of names (e.g. Fox and Lasker, 1983; Barrai et al., 1996, 1999; Rodriguez-Larralde et al., 1998), partly because this may provide information on the extent of inbreeding in the population (Lasker, 1983). It seems likely also that the frequency distribution of surnames could model the distribution of alleles subject to the same creation, death and transmission processes.

In countries in Europe and America for which data are available there are hundreds of thousands of surnames (counting every spelling variant as a different name), with between 0.2 and 2% of the population holding the commonest name. Between about 35 and 50% of names occur only once (*singletons*); the next commonest class is names that occur twice, the next is names that occur three times and so on. The usual features of the distribution are illustrated in Fig. 1, which uses the data on 88.7 million US telephone subscribers with 1.74 million different surnames, as described by Tucker (2001), Hanks and Tucker (2000). The line shown is $f(x) = 0.357x^{-1.41}$, derived by fitting a power law to the first 20 points. At larger values of x, the curve of the data becomes slowly steeper (ν increases). Other countries give similar patterns, with $1 < \nu < 2$. This pattern applies in English-speaking countries, and in Germany, Switzerland and Italy, but not necessarily in countries which have other naming histories and customs, even in western Europe.

2. Methods

2.1. Frequency distribution: rate of change

Let $n(x, t)$ represent the mean number of individuals in labeled groups of size x at time t. For example, a group may consist of all individuals of a colony of bacteria with common

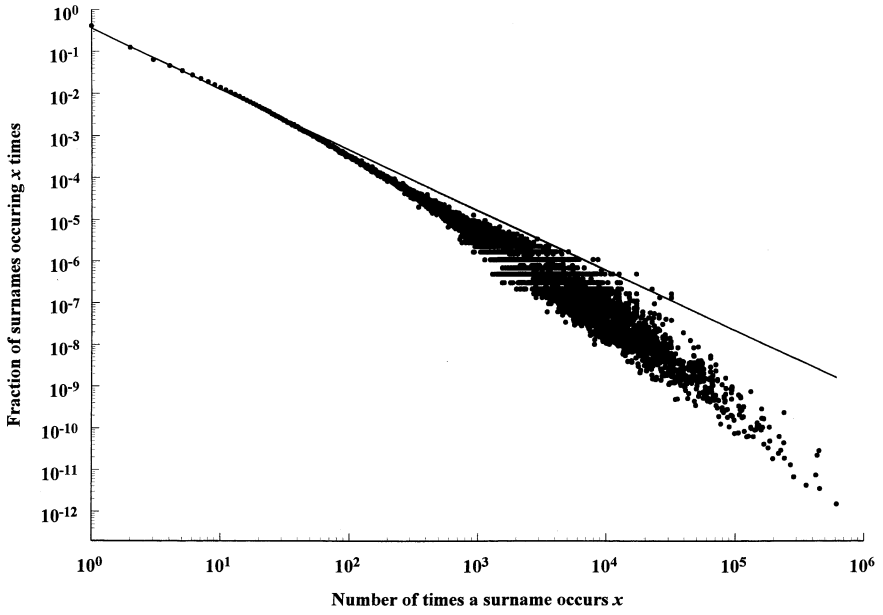

Fig. 1. US surname data and small-x fit, $f(x) = 0.3574/x^{1.4435}$. Note lack of fit at large values of the family group size x.

genotype or it could represent a set of all families of a given region with common surnames, each such family, in this case, being considered an "individual". Then $f(x, t)$ defined by:

$$f(x, t) \equiv n(x, t)/x, \tag{1}$$

is the mean number of groups of size x and, normalized, is the distribution of groups versus group-size x.

Let k_B and k_D represent birth and death rates (/individual/unit time), assumed constant over all the individuals. Then with a birth or death event, at any particular value of $x > 1$, the function $n(x, t)$ changes as follows. $n(x, t)$ either

decreases by x with a birth or death in an x-group,
increases by x with a birth in an $(x-1)$-group, or
increases by x with a death in an $(x+1)$-group.

In other words, the time rate of change in $n(x, t)$ is given by:

$$\frac{\partial n(x, t)}{\partial t} = x[-k_B n(x, t) - k_D n(x, t) + k_B n(x-1, t) + k_D n(x+1, t)], \tag{2}$$

$x = 2, 3, \ldots$

Furthermore, suppose it is possible to *create* individuals at $x = 1$. Denoting the singleton influx rate by ϕ,

$$\frac{\partial n(1, t)}{\partial t} = \phi - k_B n(1, t) - k_D n(1, t) + k_D n(2, t). \tag{3}$$

Note that the birth and death rates are approximated here as constant in time. In reality, the rates are generally time-dependent. For instance, conditions surrounding a bacterial colony may vary in time, or medical procedures can significantly change over many generations of humans. Furthermore, specific dependence of the rates on $n(x, t)$ itself, leading to non-linear equations, describes more realistic growth than the Malthusian population explosion implied by linear equations.

2.2. Indiscretion

Although exact solutions of Eq. (2) are difficult to obtain in its present discrete form, approximate solutions may be obtained through a continuum approximation, because of the large body of knowledge existent concerning second order differential equations. Eq. (3) then becomes a boundary condition at $x=1$ on the function $n(x, t)$. To this end, the right-hand side of Eq. (2) is approximated by Taylor's expansion about x:

$$\frac{\partial n}{\partial t} = x\{-k_B n - k_D n + k_B[n - \frac{\partial n}{\partial x} + \frac{1}{2}\frac{\partial^2 n}{\partial x^2} + \ldots] + k_D[n + \frac{\partial n}{\partial x} + \frac{1}{2}\frac{\partial^2 n}{\partial x^2} + \ldots]\}. \quad (4)$$

Therefore, truncating at second derivatives,

$$\frac{\partial n}{\partial t} = -kx\frac{\partial n}{\partial x} + \frac{1}{2}k_E x\frac{\partial^2 n}{\partial x^2}, \quad (5)$$

where the growth rate k and event rate k_E are defined as:

$$k \equiv k_B - k_D,$$

$$k_E \equiv k_B + k_D. \quad (6)$$

Aside from the x-dependence of the coefficients, Eq. (5) represents the flow of a fluid of density n. The first term on the right is analogous to *convection*, the sweeping along of mass towards positive x by fluid motion with velocity kx (if $k > 0$); whereas the second term corresponds to *diffusion* of mass (with diffusion constant equal to $\frac{1}{2}k_E x$). A group of individuals (like molecules of a fluid) identified at size x, will on average be swept towards larger group size through growth if $k > 0$. However, through both birth and death, the size of the group at later times becomes less and less determined, just as with the position of diffusing fluid particles. Interestingly, the potential for diffusion is hidden in the asymmetry of the first-order differences in Eq. (2).

A major difference of the present system from a fluid is the fact that the total fluid mass is not conserved (unless the growth-coefficient $k=0$). The difference is contained in the x-dependence of the coefficients and may be quantified by expressing Eq. (5) in the form:

$$\frac{\partial n}{\partial t} + \frac{\partial}{\partial x}[kxn + \frac{1}{2}k_E(n - x\frac{\partial n}{\partial x})] = kn. \quad (7)$$

If the right-hand side were zero, then Eq. (7) is of a type known as a *continuity equation* conserving the total fluid mass, with the expression in brackets representing a convective, diffusive flux of mass towards positive x. With $k > 0$, however, Eq. (7) implies that the rate of increase of the total mass equals k multiplied by the total mass, which therefore increases as $\mathrm{Exp}[kt]$ (if no newly grouped individuals are introduced (at $x=1$)).

Eq. (7) indicates a boundary condition at $x=1$. An approximation of the total number $\Sigma_1^\infty n(x, t)$ of individuals as in trapezoidal integration is given as:

$$\sum_1^\infty n(x,\ t) \approx \frac{1}{2}\, n(1,\ t) + \int_1^\infty dx\, n(x,\ t). \quad (8)$$

Then it is possible to show that Eq. (7) implies:

$$n - \frac{\partial n}{\partial x} = \phi/k_B \text{ at } x = 1, \quad (9)$$

where ϕ again represents the rate of influx of new individuals (Eq. (3)).

3. Results

3.1. Restricted case, diffusion neglected

Insight into the solution of Eq. (5) as well as approximate solutions can be obtained by considering a restricted part of the equation. Suppose the effect of diffusion is ignored. Then Eq. (5) becomes:

$$\frac{\partial n}{\partial t} = -kx\frac{\partial n}{\partial x}, \quad (10)$$

which is very easily solved: any function of $x\ \mathrm{Exp}[-kt]$ solves this equation. Therefore, a general solution can be easily constructed. Suppose that at $t=0$, the distribution n is known and

is expressed as $n(x, 0)$ ($=0$ at $x<1$). Suppose, further, that the number of singleton groups at $x=1$ is specified for all time $t>0$ by $n(1, t)$ (through the introduction of new single-member groups at a rate $\phi = kn(1, t)$, understood to vanish at $t<0$). Then the solution $n(x, t)$ is given by:

$$n(x,\ t) = n(x\ \mathrm{Exp}[-kt],\ 0) + n(1,\ t - \ln[x]/k). \quad (11)$$

There are several interesting features of this solution. As indicated by the first term on the right-hand-side, the initial distribution $n(x, 0)$ is quickly swept away from values of x in the vicinity of 1 towards large x, through the excess of birth over death. Any group identified at size x at time t would be found at $x\ \mathrm{Exp}[k(t'-t)]$ at a later time t'. The second term indicates that, without the smoothing effect of diffusion, a sharp front may exist, advancing at $x = \mathrm{Exp}[kt]$, where n goes to zero with discontinuous slope. If values of x only close to 1 are of interest, then the first term represents only a transient effect, which may be ignored after a brief time interval.

3.2. Diffusion

Further insight into this system is obtained by isolating the diffusional motion of groups. The following transformation converts the problem into one with approximately *constant* diffusion constant. Define the variable z as:

$$z \equiv 2\sqrt{x} - 1 (\geq 0 \text{ at } x \geq 1). \quad (12)$$

Furthermore, define a new density function ρ by:

$$n \equiv x^{1/4}\rho \quad (13)$$

Eq. (5) then becomes:

$$\frac{\partial \rho}{\partial t} = \frac{1}{2}\ k_{\mathrm{E}} \frac{\partial^2 \rho}{\partial z^2} + pert, \quad (14)$$

where the perturbing function *pert* is:

$$pert = -kx^{3/4}\frac{\partial (x^{1/4}\rho)}{\partial x} - \frac{3}{32}\ k_{\mathrm{E}}\frac{\rho}{x} \quad (15)$$

Eq. (14) may now be solved, for the moment neglecting *pert*. In this case, Eq. (14) expresses ordinary diffusion (see, e.g. Sommerfeld, 1964) with diffusion constant equal to $k_{\mathrm{E}}/2$. A group initially identified at x' becomes normally distributed (in the variable z, ignoring the boundary conditions at $x=1$ for the time being). The distribution broadens as time t increases, with $\rho(x, t)$ proportional to $G[x,\ x';\ t]$ given by:

$$G[x,\ x';\ t] = \frac{\mathrm{e}^{-2(\sqrt{x}-\sqrt{x'})^2/(k_{\mathrm{E}}t)}}{\sqrt{2\pi k_{\mathrm{E}} t}}. \quad (16)$$

Now if $n(1, t)$ is specified (through the influx of new individuals at $x=1$), then ρ may be solved by superposing pulses (Eq. (16)) launched continuously at $x=1$ (Sommerfeld, 1964). Details regarding this approach are given in Appendix A.

3.3. Specific exact solution through separation of variables

Eq. (5) is challenging to solve. However, exact solution is straightforward in a particularly important situation—when the temporal and spatial dependences in n factor, so that:

$$n(x,\ t) = \mathrm{Exp}[k't]\ n(x), \quad (17)$$

where k' is a constant, and $n(x)$ depends only on the group size x.

This type of solution is important, since general solutions can be constructed by the *Laplace transform method* through superposition over a range of complex k'. Solutions of the form of Eq. (17) are also directly significant. Suppose $n(x)$ is integrable over $x>1$, and suppose the influx of new singleton ($x=1$) groups is so rapid as to be approximately proportional to the total number of individuals. Bacterial mutation in the presence of a mutagen is an example, with focus on the smaller groups, singletons being provided from the pool of the largest colonies. Influx, together with growth, then results in a net growth rate k' given by:

$$k' = k + \lambda k_{\mathrm{E}}, \quad (18)$$

defining λ as the fraction of birth or death events at which a singleton group is entered at $x=1$. Eqs. (17) and (18) then exactly express the total number of individuals increasing as $\mathrm{Exp}[(k+\lambda k_{\mathrm{E}})t]$.

Even in the case that $n(x)$ is not integrable, as with lower singleton influx rates, solutions of the form of Eq. (17) may be relevant as small-x

asymptotic ($t \to \infty$) limits. As an extreme example, suppose the influx rate is adjusted so that $n(1, t)$ is simply held constant. Then at $t \to \infty$, $n(x, t) \to$ constant for all x (corresponding to $k' = 0$).

Ignoring transients from any initial distribution (at $t = 0$), solutions of the form of Eq. (17) are determined (up to a constant factor) by requiring $n \to 0$ at $x \to \infty$. Eq. (5) becomes:

$$k'n(x) = \frac{1}{2}\, k_E x \frac{d^2 n(x)}{dx^2} - kx \frac{dn(x)}{dx}. \tag{19}$$

Now Eq. (19) is a particular case of the thoroughly investigated *Kummer equation* (Abramowitz and Stegun, 1965) and, requiring finiteness at $x \to \infty$, is solved by the *Kummer function* U:

$$n(x) \propto U[k'/k,\ 0, \frac{2k}{k_E}\ x]. \tag{20}$$

The Kummer function with zero as the second argument was investigated by Bateman (1931). The Bateman function U is easily computed numerically. For example, *Mathematica*™ provides a standard routine, HypergeometricU. Alternatively, a simple interpolation formula is given below (Eq. (25)).

As is easily seen, U depends only on the group-size x, the influx parameter k'/k, and the growth ratio g defined by:

$$g \equiv k/k_E. \tag{21}$$

The function U decreases monotonically as x increases from 1 as indicated in Fig. 2. For illustration, U/x is plotted versus group-size x at $\lambda = 0.1$ and $g = 0.4$. Also shown, as an indication of the accuracy of the continuum approximation adopted here, are the results of a simulation directly implementing discrete birth, death, and influx rates.

As is seen, the major drop in $n(x)/x$ as x increases occurs near $x = 1$. The asymptotic limit for $x \gg 1$ is given (Abramowitz and Stegun, 1965) by:

$$U[k'/k,\ 0, \frac{2k}{k_E}\ x] \to \frac{1}{\left(\frac{2k}{k_E}\ x\right)^{k'/k}},\quad x \to \infty. \tag{22}$$

Therefore, for $k' > k$, integrals to $x = \infty$ (or sums over integral x) converge. If $k' < k$, i.e. when the influx at $x = 1$ is small, then the integrals do not converge, yet still may represent the finite-x $t \to \infty$ limit. The large-x limit of Eq. (22) is identical to the convective limit of Eq. (11).

Note: Eq. (22) is strictly valid only at $k \neq 0$. At $k = 0$, Eq. (20) simplifies to:

$$n(x) \propto x^{1/2} K_1\left[\sqrt{8k_{net}\frac{x}{k_E}}\right]$$

$$\to \text{constant} \times x^{1/4}\, e^{-\sqrt{8k_{net}(x/k_E)}},\quad x \to \infty, \tag{23}$$

where K_1 is a *modified Bessel function* (Abramowitz and Stegun, 1965). This expression reflects the x-dependence in Eq. (13).

We believe that the power-law decrease in U as $x \to \infty$ indicated in Eq. (22) corresponds to a (convergent (when $k' > k$)) power law derived by Simon (1955). Asymptotically, the group distribution $f(x) = n(x)/x$ behaves at large x as $1/x^{\nu}$, where $\nu > 2$. Simon (1955) also mentioned a large-x exponential convergence (suggested heuristically), which differs from the power-law form found here.

What is interesting about the function U, as may be seen in Fig. 2, is that at smaller values of x (closer to 1), the decrease in U/x may also be approximated by a power law, yet with exponents less than 2, corresponding to a gentler fall-off than at $x \to \infty$. To investigate this explicitly, using the form of $n(x)$ in Eq. (20), the power-law exponent ν can be expressed in terms of the growth ratio g by computing derivatives at $x = 1$. The result is:

$$\upsilon = 2g\frac{k'}{k}\frac{U[\frac{k'}{k}+1,\ 1,\ 2g]}{U[\frac{k'}{k},\ 0,\ 2g]},, \tag{24}$$

where g is again the growth ratio k/k_E. Eq. (24) is presented in Fig. 3 showing contours of constant small-x power-law exponent ν over k'/k and g. The exponent is generally less than 2.

A useful interpolation function for computing the Bateman function $U[a, 0, z]$ is obtained by approximation as a *superhyperbola*:

$$U[a,\ 0,\ z]^{-3/(4a)} - z^{3/4} \approx \Gamma[1+a]^{3/(4a)}, \tag{25}$$

where Γ is the gamma function.

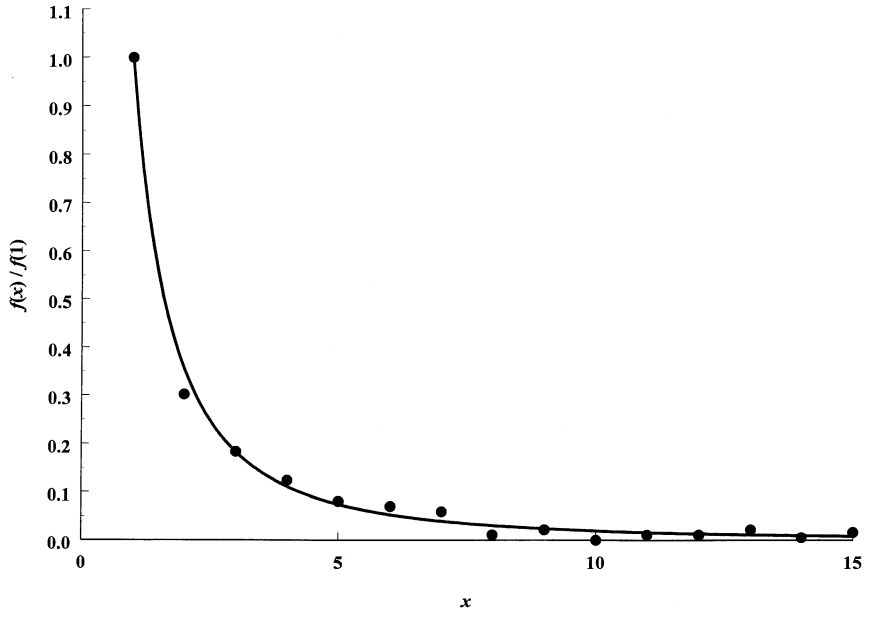

Fig. 2. Group distribution f vs. group-size x, solid curve: theoretical; solid dots: simulation. Growth = 15 %; singleton influx rate = 10 %.

The formula (25) is exact at $z = 0$ and at $z \to \infty$; at intermediate values, the error is (generally well) within $(-3\%, +6\%)$ for all $0 < a < 1.5$. This error is of the order of the thickness of the solid curves in the fits described below and drawn in Figs. 4–8. Accurate and rapid fits to data may therefore be made using Eq. (25) together with a standard non-linear curve fitting routine.

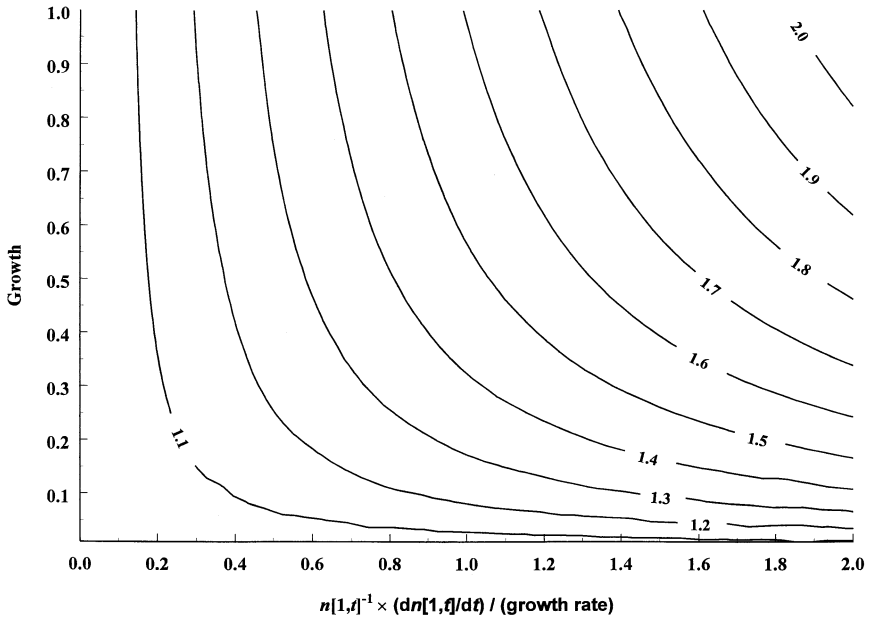

Fig. 3. Small-x power-law exponent in terms of singleton rate of increase and growth (birth−death) per event.

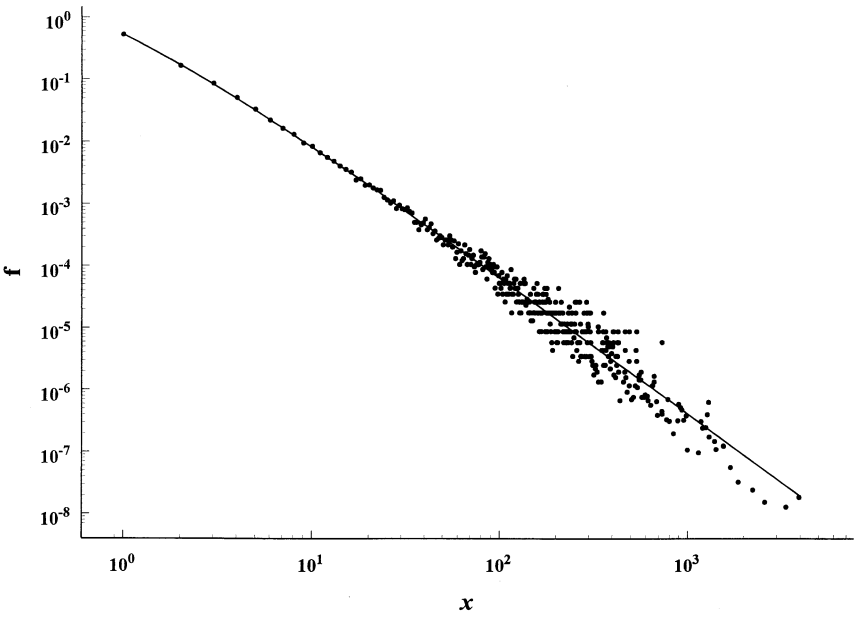

Fig. 4. Estimated surname distribution f vs. occurrence x: data from Hamburg. Solid curves are theoretical fits to data (solid circles).

4. Discussion

An interesting result of the above analysis is that the distributions may often be approximated by power-law functions—with two different exponents: one for smaller and one for larger group sizes x. With a sufficiently large singleton influx rate, the distribution is integrable over x. However, at lower influx rates, the small-size distribution breaks off in the large-time limit, resulting in

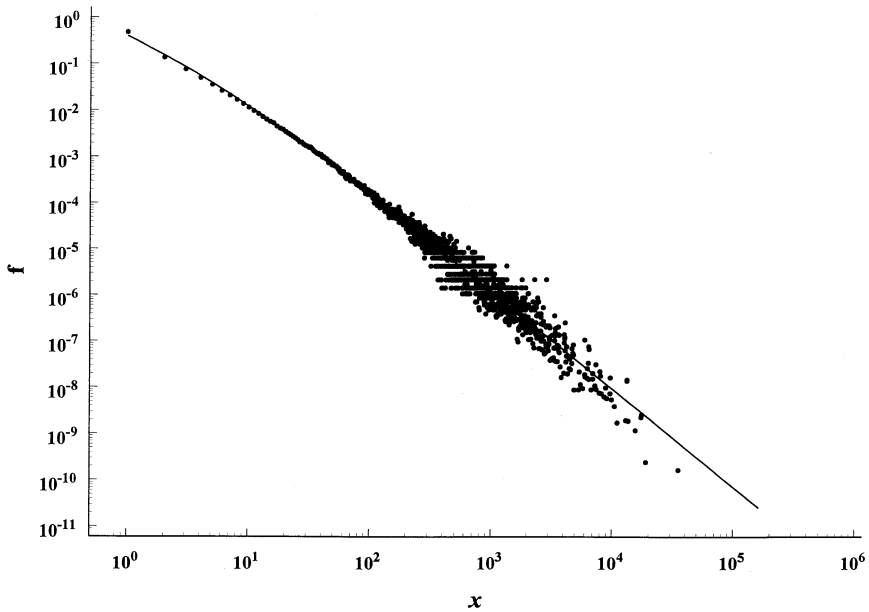

Fig. 5. Estimated surname distribution f vs. occurrence x: data from Germany. Solid curves are theoretical fits to data (solid circles).

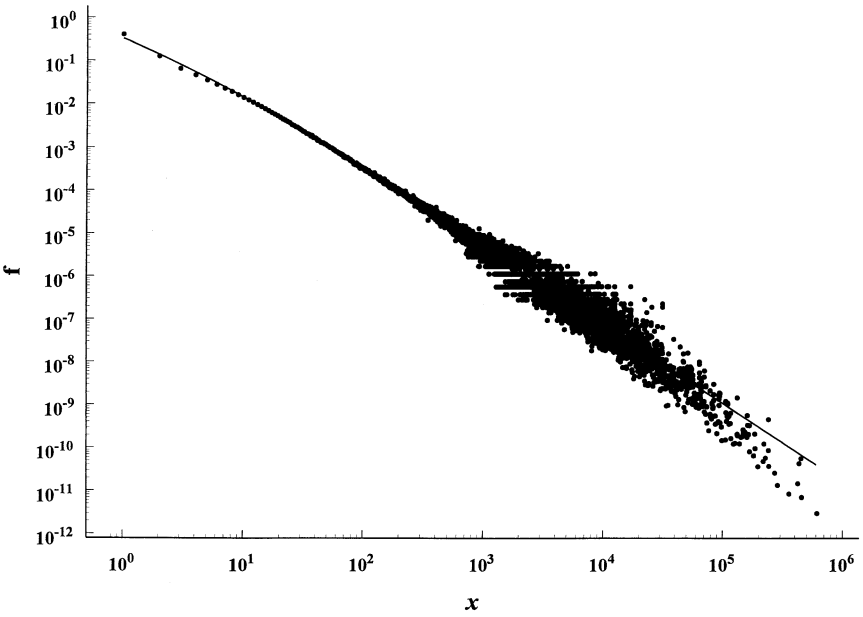

Fig. 6. Estimated surname distribution f vs. occurrence x: data from the USA. Solid curves are theoretical fits to data (solid circles).

an approximation, which is accurate at small, but not large sizes. In fact, the large-time limiting function is not integrable. In other words, in some case the $x \to \infty$ and time $t \to \infty$ limits are not interchangeable. Appreciation of these possibilities may unravel misunderstandings between Mandelbrot and Simon (Mandelbrot, 1959).

The solution given by Eq. (20) was applied through the fitting of surname distributions in Hamburg, Germany, and in Italy described by

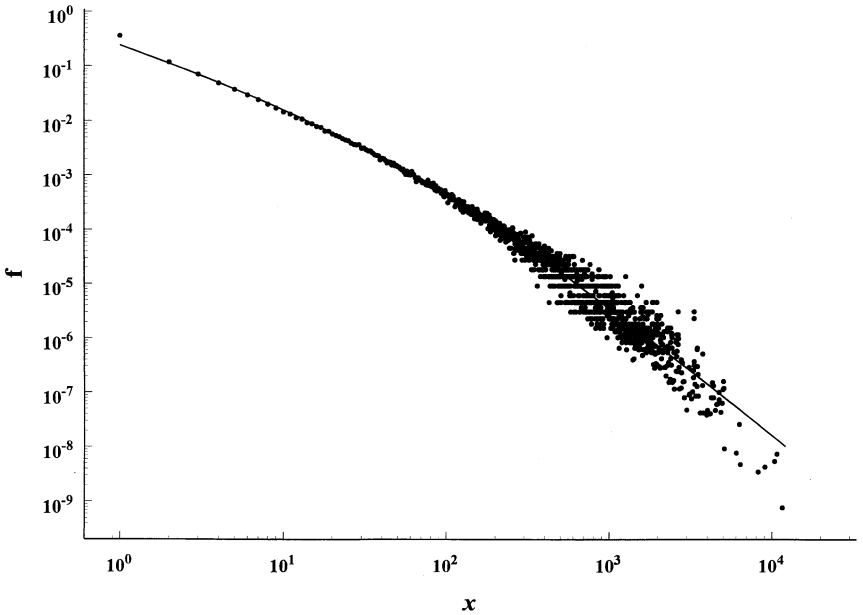

Fig. 7. Estimated surname distribution f vs. occurrence x: data from Italy. Solid curves are theoretical fits to data (solid circles).

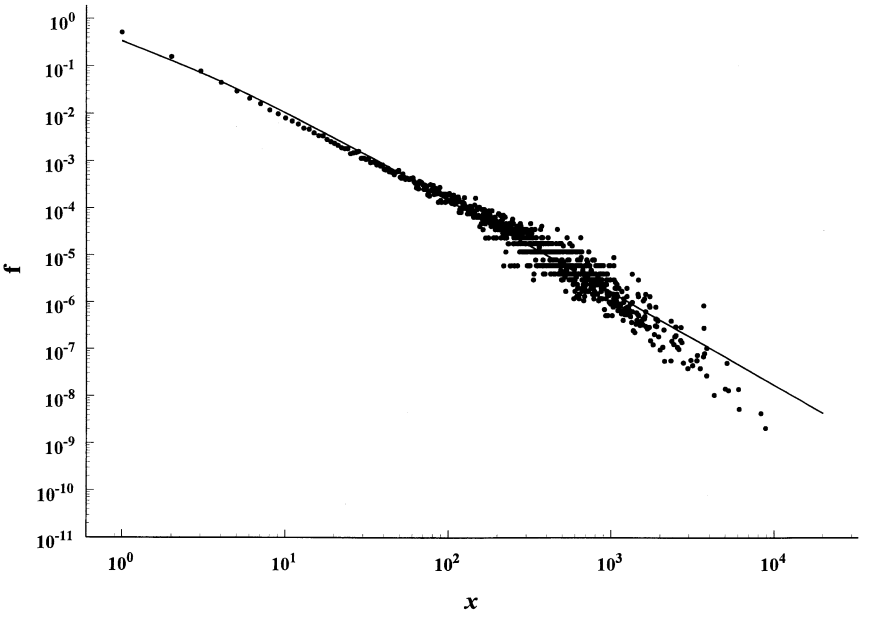

Fig. 8. Estimated surname distribution f vs. occurrence x: data from Switzerland. Solid curves are theoretical fits to data (solid circles).

Barrai et al. (1996, 1999). Surname distributions within the USA (Tucker, 2001; Hanks and Tucker, 2000) and Switzerland were also investigated. The (un-normalized) surname distributions f were approximated by dividing known family group numbers $f_{\mathrm{raw}}[x]$ by their separation in group size x, important at large sizes, where family groups are separated in size by ranges greater than unity. The positive bias induced by this division was fine-tuned by further dividing by $1+\mathrm{e}^{-(3/2)f_{\mathrm{raw}}}$, which was deduced empirically through simulation (Appendix B). Random variation in the large-x family group separation results in broadening of the estimated distributions. The distributions were normalized, dividing by the total number of families. Least-square fits were made by varying the ratio k'/k, the growth factor g (k/k_{E}) and multiplicative constant using a nonlinear fitting routine of *Mathematica*™.

The specific assumptions behind such a program may be summarized as follows:

Birth and death rates are assumed constant over time and family. The family birth and death rates crudely accommodate inward and outward migration of families as well as family propagation.

The number of individual families $(x=1)$ is assumed to increase rapidly, in fact, at a rate proportional to their number. It is simple to show that such a rapid rate is required for a non-trivial (i.e. non-constant) limit of the ratio $n(x,t)/n(1,t)$ to exist in a neighborhood of $x=1$ at $t\to\infty$. In the case of integrability, the singleton family creation rate is proportional to the total number of families.

The results are shown in Figs. 4–8. As can be seen, the fitted curves approximate the data remarkably closely despite the simplifying assumptions behind this theory. Resulting R^2 values exceeded 0.97 (calculated asymptotically by means of the *Mathematica*™ nonlinear curve fitting routine). Inaccuracy visible at small x is slight, considering the many decades in x over which the fits were made. As this range in x translates to time (of many generations), temporal shifts in the rate constants are a possible explanation.

Information to be drawn from the coefficients determined by the fits is limited, as the fitted

values of k'/k and g are strongly correlated. Nevertheless, the statements contained in Table 1 characterize the parameters. Note how Hamburg's is distinguished from the other surname groups.

Many possibilities exist for future research. We have not tried to relate the surname study to information from other sources on family birth, death and singleton influx rates in various countries. Of course, relaxation of the assumption of constancy of these rates in time and across families may be necessary to account for details in surname and other frequency distributions. Also, family migration may fruitfully be separated from propagation. Another major advance of the theory in general would be the introduction of non-linear effects. Probably the relation of such systems to physical phenomena, as in the present study, will help in their understanding as well.

Acknowledgements

We thank Dr. I. Barrai and Dr. K. Tucker for generously giving us access to the original data in their referenced works.

Appendix A: Approximate general solution

An approximate general solution of Eq. (5) is presented here for time t large enough and x small enough that any initial distribution has been convected away, given the singleton group value $n(1, t)$ defined for time t. Using procedures described by Sommerfeld (1964), the solution ρ_0 of Eq. (14) when the perturbing function pert is neglected is:

Table 1
Parameters resulting in fit of the Bateman function to surname distributions

Location	Influx λ (%)	Growth g (%)	Small-x exponent u
Hamburg	10	30	1.6
Italy	1	1	1.1
USA	<0.3	3	1.2
Germany	3	10	1.3
Switzerland	<3	20	1.4

$$\rho_0(z,\ t) = z \times \int_0^t dt' \frac{1}{t-t'}\ G[x,\ 1;\ t - t']n(1,\ t'), \tag{A1}$$

where G is the broadening pulse function of Eq. (16). Eq. (A1) is simple to evaluate numerically, and results have been found to agree with simulations of the discrete system.

An approximate solution of Eq. (14) accounting for pert can now be constructed. The solution ρ is expressed in terms of a perturbation ρ_1 from ρ_0:

$$\rho = \rho_0 + \rho_1. \tag{A2}$$

Using methods of perturbation theory (see, e.g. Schiff, 1955), ρ_1 may be expressed in terms of a *retarded Green's function* G_{ret}:

$$G_{\text{ret}} = \{G[z - z';\ t - t'] - G[z + z';\ t - t']\}\Theta[t - t'], \tag{A3}$$

where the function G is expressed here symbolically in terms of the transformed z-variables. Because of the step-function Θ, G_{ret} solves:

$$\begin{aligned} &\frac{\partial G_{\text{ret}}}{\partial t} - \frac{1}{2}\ k_{\text{E}} \frac{\partial^2 G_{\text{ret}}}{\partial z^2} = \delta[t - t'], \\ &G_{\text{ret}} = 0,\ \ z = 0, \\ &G_{\text{ret}} \to \delta[z - z'],\ \ t \to t'_+, \end{aligned} \tag{A4}$$

where δ is the Dirac delta function. Multiplying G_{ret} by *pert* (given by Eq. (15) expressed at the zeroth-order solution ρ_0) and integrating over positive t' and z' results in ρ_1 given by:

$$\rho_1(z,\ t) = \int_0^\infty dz' \int_0^t dt' \{G[z - z';\ t - t'] - G[z + z';\ t - t']\} pert(z',\ t'). \tag{A5}$$

The integrals of Eq. (A5) are simple (though time-consuming) to carry out numerically. Comparison to simulations implementing discrete birth and death rates indicates accuracy generally within 5% with slowly increasing (e.g. linear) functions $n(1, t)$ of time, if $k/k_{\text{E}} < 30\%$. Fig. 9 shows such a solution for various values of k/k_{E}.

Appendix B: Correction of bias in estimates of the distribution $f(x)$ when measured groups are separated

Within the text we described how distributions f may be approximated by dividing measured group numbers f_{raw} by their separation in group size x. At small sizes where the group numbers are nearly always >1, no change between group numbers f_{raw} and f is effected. However, at large sizes, where groups are separated in size, the transformation becomes significant, and yet there exists a positive bias in the resulting estimate when the group numbers drop to unity, and the groups become widely separated in x. One approach to eliminate this bias is to average over several separated groups (Ogden, 2000).

Alternatively, the bias may be minimized pointwise through correction. In the limit of infinite separation, the bias can be estimated through the following continuum approximation: Suppose that

$$X_1, \ldots, X_n \sim \text{Uniform}(0, L). \tag{B1}$$

Arrange the X_is in ascending order: $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. For

$$D_k = 1/[(X_{(k+1)} - X_{(k-1)})/2], \tag{B2}$$

we have

$$D_k \sim \frac{1}{L}\,(2 + (n-1)F(2(n-1),\ 4)), \tag{B3}$$

and therefore the expectation value is

$$E(D_k) = 2(n/L), \tag{B4}$$

where $F(u, v)$ is a random variable with an F-distribution. Therefore, the bias equals 100% when groups are extremely separated.

B.1. Proof

The above result can be derived from the following lemmas, defining

$$Y_i = X_i/L, \tag{B5}$$

and therefore

$$Y_{(i)} = X_{(i)}/L,\ D_k = (2/L)/(Y_{(k+1)} - Y_{(k-1)}). \tag{B6}$$

Lemma 1. *The random variables* $V_1 = Y_{(i)}/Y_{(j)}$ *and* $V_2 = Y_{(j)}$ *are statistically independent, with* V_1

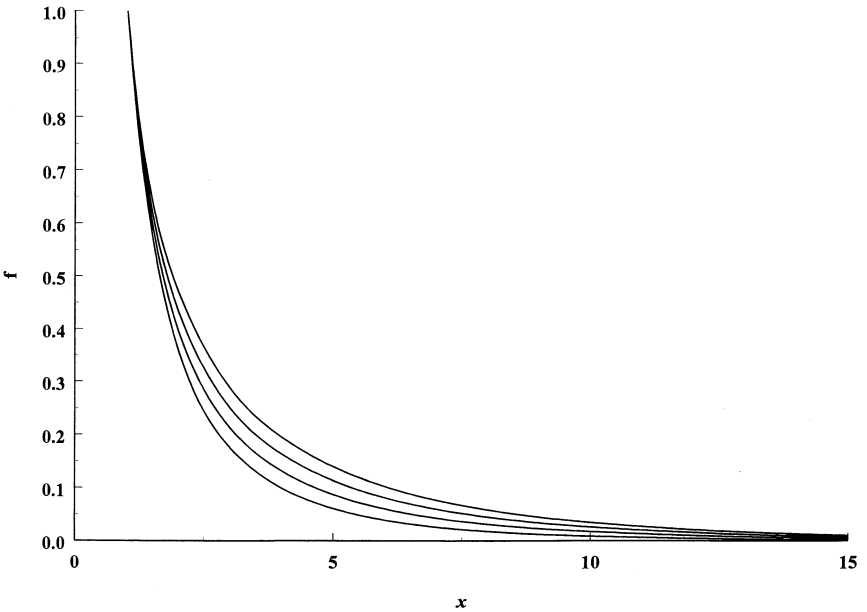

Fig. 9. Group distributions f vs. size x with growth ratio $k/k_{\text{E}} = 0$, 10, 20, 30% (left to right) with singleton ($x = 1$) influx rate equal to 10% k_{E}. Curves were computed using the Green's function approach at a time t equal to 1000 $\times$ the initial event time.

and V_2 having Beta(i, j−i) and Beta(j, n−j+1) distributions, respectively (Arnold et al., 1992).

Lemma 2. *If $V \sim Beta(p, q)$, then $1-V \sim Beta(q, p)$* (Johnson et al., 1995).

Lemma 3. *Suppose that $V_1 \sim Beta(p_1, q_1)$ and $V_2 \sim Beta(p_2, q_2)$. If $p_1 = p_2 + q_2$, then $Y = V_1 V_2$ has a beta distribution $Y \sim Beta(p_2, q_1 + q_2)$* (Johnson et al., 1995).

Lemma 4. *If $Y \sim Beta(p, q)$, then there exist two independent random chi-square variables $X_1 \sim \chi^2(2p)$ and $X_2 \sim \chi^2(2q)$ such that $Y = X_1/(X_1 + X_2)$* (Johnson et al., 1995).

Note that

$$D_k = \frac{2/L}{Y_{(k+1)} - Y_{(k-1)}} = \frac{2/L}{Y_{(k+1)}(1 - Y_{(k-1)}/Y_{(k+1)})}. \quad \text{(B7)}$$

According to Lemma 1 and Lemma 2,

$$D_k \sim \frac{2/L}{Beta(k+1,\ n-k) \times (1 - Beta(k-1,\ 2)} \sim \frac{2/L}{Beta(k+1,\ n-k) \times Beta(2,\ k-1)}. \quad \text{(B8)}$$

Then according to Lemma 3 and Lemma 4,

$$D_k \sim \frac{2/L}{Beta(2,\ n-1)} \sim \frac{2}{L}\frac{\chi^2(4) + \chi^2(2(n-1))}{\chi^2(4)} \sim \frac{2}{L}\left(1 + \frac{n-1}{2}\ F(2(n-1),\ 4)\right). \quad \text{(B9)}$$

B.2. Interpolation

The above indicates that the bias in the distribution estimate f ranges from 0 to 100%, as the group number f_{raw} falls from $\gg 1$ to 1. Through extensive simulation, the following interpolation formula has been found accurate at intermediate values to within ±3%:

$$f[x_i] = \frac{f_{raw}[x_i]}{\frac{1}{2}(x_{i+1} - x_{i-1})} \frac{1}{(1 + e^{-(3/2)f_{raw}})}. \quad \text{(B10)}$$

References

Abramowitz, M., Stegun, I.A., 1965. Handbook of Mathematical Functions. Dover Publications, New York.

Arnold, B.C., Balakrishnan, N., Nagaraja, H.N., 1992. A First Course in Order Statistics. Wiley, New York.

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. Annals of Human Biology 23, 431–435.

Barrai, I., Rodriguez-Larralde, A., Mamolini, E., Scapoli, C., 1999. Isonomy and isolation by distance in Italy. Human Biology 71, 947–961.

Bateman, H., 1931. The k-function, a particular case of the confluent hypergeometric function. Transactions of the American Mathematical Society 33, 817–831.

Fox, W.R., Lasker, G.W., 1983. The distributions of surname frequencies. International Statistical Review 51, 81–87.

Hanks, P., Tucker, D.K., 2000. A diagnostic database of American personal names. Names 48, 58–69.

Johnson, N.L., Kotz, S., Balakrishnan, N., 1995. Continuous Univariate Distributions. Wiley, New York.

Lasker, G.W., 1983. The frequencies of surnames in England and Wales. Human Biology 55, 331–340.

Mandelbrot, B.B., 1959. A note on a class of skew distribution functions: analysis and critique of a paper by H.A. Simon. Information and Control 2, 90–99.

Ogden, T., 2000, private communication.

Rodriguez-Larralde, A., Barrai, I., Nesti, C., Mamolini, E., Scapoli, C., 1998. Isonomy and isolation by distance in Germany. Human Biology 70, 1041–1056.

Schiff, L.I., 1955. Quantum Mechanics. McGraw-Hill Book Company, New York.

Simon, H.A., 1955. On a class of skew distribution functions. Biometrika 42, 425–440.

Sommerfeld, A., 1964. Partial Differential Equations in Physics. Academic Press, New York.

Tucker, D.K., 2001. Distribution of forenames, surnames and forename–surname pairs in the USA. Names 49, 66–96.