On the distribution of family names

William J. Reed [a] and Barry D. Hughes [b]

[a] *Department of Mathematics and Statistics, University of Victoria, Victoria, British Columbia, Canada V8W 3P4* (reed@math.uvic.ca)

[b] *Department of Mathematics and Statistics, University of Melbourne, Parkville, Victoria 3010, Australia* (hughes@ms.unimelb.edu.au)

Abstract

We present a model for the distribution of family names that explains the power-law decay of the probability distribution for the number of people with a given family name. The model includes a description of the process of generation or importation of new names, and a description of the growth of the number of individuals with a name, and corresponds for a long-enduring culture to a Galton–Watson branching process killed at a random time. The exponent that characterizes the decay of the resulting distribution is determined by the characteristic rates for the creation of new names and for the growth of the population. The power-law decay is modulated by small-amplitude log-periodic oscillations. This is rigorously established for a particular form of the offspring distribution in the branching process, but arguments are presented to show that the phenomenon will occur under wide circumstances.

1 Introduction

It has been observed recently [1,2] that empirical frequency distributions of family names appear to exhibit power-law decay over several orders of magnitude. An explanation for this phenomenon, based on a model of Simon [3] originally used to explain Zipf's law for word frequencies, was proposed by Zanette and Manrubia [2]. In this article we offer another model to explain the power-law phenomenon, which is based on the *Galton–Watson branching process* originally proposed as a model for the evolution of family names [1]. We add to the Galton–Watson process assumptions concerning the introduction of new names. Specifically we assume that new names are created in a birth process with immigration *i.e.* that a new name can either be created from an

[1] Francis Galton formulated the model in a question on the proliferation and extinction of family names in the *Educational Times* in 1873. A partial solution was provided by H.W. Watson and a joint paper on the subject resulted [4].

existing one (*e.g.* through a change in spelling *etc.*) with a constant probability for all names at all times; or it can be introduced through the arrival of an immigrant, with arrivals occuring in a Poisson process. With this assumption we are able to show that the probability of there being m individuals with a given name is given for large m by

$$\Pr\{\text{exactly } m \text{ individuals have this name}\} \approx m^{-1-\kappa} \quad \text{as } m \to \infty \tag{1}$$

(in a sense made more precise below), or, equivalently,

$$\Pr\{\text{at least } m \text{ individuals have this name}\} \approx m^{-\kappa} \quad \text{as } m \to \infty. \tag{2}$$

The exponent κ depends on the mean number of offspring per individual in the Galton–Watson process and on the rate of creation of new names from old, but not on the rate of immigration.

In Section 2 we show how the assumptions of the model lead to the formulation of the process as a 'killed' Galton–Watson branching process. The predictions of the model are then extracted in three ways in Section 3. The first way, which is rather naive, produces the behaviour (1) for the specific case of a geometric offspring distribution. The second way, a more careful analysis for a geometric offspring distribution, shows that the true asymptotic form is

$$\Pr\{\text{exactly } m \text{ individuals have this name}\} \sim \frac{Q(m)}{m^{1+\kappa}} \quad \text{as } m \to \infty, \tag{3}$$

where $Q(m)$ is a bounded periodic function of $\log m$. The third analysis, for a more general offspring distribution, shows that the 'log-periodic' modulation of the asymptotic power law holds under quite general circumstances. A few comments on experimental data are made in Section 4.

We shall use the following notation. Angle brackets denote the expectation of a random variable. The probability generating function (PGF) of a random variable X is defined by

$$\sum_{m=0}^{\infty} \Pr\{X = m\} s^m = \langle s^X \rangle, \tag{4}$$

where $|s| \leq 1$. A bar over a random variable indicates that it is derived by killing a time-evolving random variable at a random time.

2 The model

We assume that once a family name originates its evolution follows a Galton–Watson branching process (see *e.g.* [5]), *i.e.* its frequency X_n, n generations after origination can be written as

$$X_n = Z_1 + Z_2 + Z_3 + \ldots + Z_{X_{n\ 1}} \tag{5}$$

where $X_0 = 1$ and $\{Z_i\}$ is a set of independent, identically distributed (iid) random variables with support on $\{0, 1, 2, \ldots\}$ representing the number of offspring produced by individuals $i = 1, 2, \ldots, X_{n\ 1}$ of the previous generation. [2]

Different family names will have been present for different lengths of time. To account for this fact we will consider a model for the way in which family names originate. For convenience we do this in continuous time, and then convert the results to discrete time for incorporation with the (discrete time) Galton–Watson model. Specifically we assume that the number of names $N(t)$ evolves as a linear birth process with immigration. Thus conditional on $N(t) = n$, $N(t + h)$ will either be $n + 1$ with probability $(\lambda n + \rho)h + o(h)$; or n with probability $1 - (\lambda n + \rho)h + o(h)$; or some other value with probability $o(h)$. The parameter λ represents the rate at which new names develop from existing names (*e.g.* through a change in spelling) and the parameter ρ the rate at which new names enter through immigration (in a Poisson process). These assumptions lead to the set of differential equations

$$P'_n(t) = [\lambda(n-1) + \rho]P_{n\ 1}(t) - (\lambda n + \rho)P_n(t) \tag{6}$$

for $P_n(t) = \Pr\{N(t) = n\}$. An ordinary differential equation for the evolution of the mean number of names $\langle N(t)\rangle$ can be derived simply by multiplying Eq. (6) by n and summing over n. If $N(0) = n_0$, it is readily shown that

$$\langle N(t)\rangle = (\rho/\lambda)(e^{\lambda t} - 1) + n_0 e^{\lambda t}. \tag{7}$$

This is a known result for birth processes with immigration (cf. [6], p. 238).

Feigin [7] has shown that the birth process with immigration is one of only two homogeneous point processes $N(t)$ which have the property that, conditional

[2] Since family names are usually carried through the male line the model is, strictly speaking, restricted to the population of males, with the X_i's representing numbers of sons. However when we compare the model with data, we will assume a one-to-one sex ratio, so that the relative frequency distribution of names in the whole population will be assumed to be identical to that of the male population.

on $N(\tau) - N(0) = k$, the successive jump times are distributed in the interval $(0, \tau)$ as the order statistics of k independent, identically distributed random variables U_1, U_2,..., U_k, where

$$\Pr\{U_i \leq u\} = \frac{m(u) - m(0)}{m(\tau) - m(0)}, \quad 0 \leq u \leq \tau, \tag{8}$$

and $m(u) = \langle N(u) \rangle$. In our case this reduces to

$$\Pr\{U_i \leq u\} = \frac{e^{\lambda u} - 1}{e^{\lambda \tau} - 1}, \quad 0 \leq u \leq \tau. \tag{9}$$

If we let $T_i = \tau - U_i$ denote the time since the jump occurred, we find that

$$\Pr\{t < T_i \leq \tau\} = \frac{e^{\lambda t} - e^{\lambda \tau}}{1 - e^{\lambda \tau}}, \quad 0 \leq t \leq \tau, \tag{10}$$

corresponding to the probability density function

$$f(t) = \frac{\lambda e^{\lambda t}}{1 - e^{\lambda \tau}}, \quad 0 \leq t \leq \tau. \tag{11}$$

It may be observed that this density is independent of k.

If a name is selected at random from the list of all names at time τ, it will either be one of the n_0 'Ur' names in existence at time 0, or one of the new names that has come into existence at some time $t \in (0, \tau)$. The probabilities of these events are, respectively, $n_0/N(\tau)$ and $1 - n_0/N(\tau)$, and we find that the time that a name observed at time τ has been in existence has the density

$$\phi_\tau(t) = \delta_+(\tau - t)\Big\langle \frac{n_0}{N(\tau)} \Big\rangle + \frac{\lambda e^{\lambda t}}{1 - e^{\lambda \tau}} \Big\{ 1 - \Big\langle \frac{n_0}{N(\tau)} \Big\rangle \Big\}, \tag{12}$$

where $0 \leq t \leq \tau$ and δ_+ denotes the one-sided Dirac delta function. As $N(\tau) \to \infty$ with probability 1, it follows that

$$\Big\langle \frac{n_0}{N(\tau)} \Big\rangle \to 0 \quad \text{as } \tau \to \infty \tag{13}$$

and so for large τ we arrive at the exponential density

$$\phi(t) = \lambda e^{\lambda t} \tag{14}$$

for the time that a name has been in existence.

To check on the rate of decay to zero of $\langle n_0/N(\tau)\rangle$, we form the generating function

$$\langle s^{N(t)}\rangle = \sum_{n=0}^{\infty} P_n(t)s^n = P(t,s), \quad \text{say} \tag{15}$$

and deduce from Eq. (6) the partial differential equation

$$\frac{\partial P}{\partial t} + \lambda s(1-s)\frac{\partial P}{\partial s} + \rho(1-s)P = 0, \tag{16}$$

to be solved subject to the initial condition $P(0,s) = s^{n_0}$. Solving this in the usual manner by the method of characteristics [8] we find that

$$\langle s^{N(t)}\rangle = P(t,s) = s^{n_0}[s + (1-s)e^{\lambda t}]^{\ n_0\ \ \rho/\lambda}. \tag{17}$$

Thus

$$\Big\langle \frac{n_0}{N(\tau)} \Big\rangle = n_0 \int_0^1 \langle s^{N(\tau)\ \ 1}\rangle ds = n_0 \int_0^1 \frac{s^{n_0\ \ 1} ds}{[s + (1-s)e^{\lambda\tau}]^{n_0+\rho/\lambda}} \tag{18}$$

$$\leq n_0 \int_0^1 \frac{ds}{[s + (1-s)e^{\lambda\tau}]^{n_0+\rho/\lambda}} \tag{19}$$

$$= \frac{n_0}{(n_0 - 1 + \rho/\lambda)(e^{\lambda\tau} - 1)}\{1 - e^{\lambda\tau(1\ \ n_0\ \ \rho/\lambda)}\} = O(e^{\ \lambda\tau}). \tag{20}$$

Since family names have been around for a very long time, we assume that $e^{\ \lambda\tau} \ll 1$, and this implies that time that a name, randomly selected from those currently in existence, has been in existence is distributed exponentially; or equivalently in discrete time that it is distributed geometrically.

Coupling this with the Galton–Watson process model for the evolution of names after their introduction, we see that the distribution of the frequency of any name, should be that of a Galton–Watson process after a geometrically distributed number of generations, where the parameter p of the geometric distribution is related to λ above by $p = 1 - e^{\ \lambda\Delta}$, where Δ is the length of one generation.

3 Analysis of the model

The PGF for the number X_n of individuals in the nth generation of a Galton–Watson branching process $X_{n+1} = Z_1 + Z_2 + \ldots + Z_{X_n}$, started with one

individual for the zeroth generation, is given [5] by

$$G_n(s) = G_{n\ \ 1}(g(s)). \tag{21}$$

Here $g(s) = G_1(s)$ is the pgf for the number of offspring of an individual.

The PGF $G(s) = E(s^{\bar{X}})$ of the state $\bar{X}$ of the branching process killed on the production of the Nth generation according to the geometric distribution

$$\Pr\{N = n\} = p(1-p)^{n\ \ 1}, \quad n = 1, 2, 3, \ldots \tag{22}$$

is given by

$$G(s) = \sum_{n=1}^{\infty} G_n(s) p(1-p)^{n\ \ 1}. \tag{23}$$

Splitting the first term off from the sum, we obtain a functional equation for $G(s)$:

$$G(s) = pg(s) + (1-p)G(g(s)). \tag{24}$$

For a general offspring PGF $g(s)$, there seems to be little that one can do except to attack this functional equation directly. However we first look at a particular case where more elementary means are at our disposal.

Harris ([5], p. 9) has derived the explicit form for $G_n(s)$ in the case where the offspring distribution is

$$P(\text{individual has } k \text{ offspring}) = \begin{cases} 1 - b/(1-c), & k = 0, \\ bc^{k\ \ 1}, & k = 1, 2, \ldots \end{cases} \tag{25}$$

In the special case in which each individual has at least one offspring (so that $b = 1 - c$), to which we now restrict our attention, Harris' solution gives the relatively simple PGF

$$G_n(s) = \frac{b^n s}{1 - (1-b^n)s} = \sum_{m=1}^{\infty} b^n (1-b^n)^{m\ \ 1} s^m \tag{26}$$

and so

$$G(s) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} p(1-p)^{n\ \ 1} b^n (1-b^n)^{m\ \ 1} s^m. \tag{27}$$

It is easy to verify directly that the series (27) satisfies the functional equation (24). Extracting the coefficient of s^m from $G(s)$ we see that

$$P(\bar{X} = m) = \sum_{n=1}^{\infty} p(1-p)^n \ {}^1 b^n (1-b^n)^m \ {}^1. \tag{28}$$

3.1 The simplest model: naive analysis

To determine the large-m asymptotic form of $P(\bar{X} = m)$, we first use a naive argument that reveals part of the behaviour, but misses a subtle point. We write $\approx$ to indicate that no claim is made to the status of the result (that is, whether it is truly an asymptotic representation, or a numerical approximation of any quality). Approximating the sum (28) by an integral, we obtain

$$P(\bar{X} = m) \approx \frac{p}{1-p} \int_0^{\infty} [b(1-p)]^x (1-b^x)^m \ {}^1 dx. \tag{29}$$

If we write $t = b^x$ we obtain

$$P(\bar{X} = m) \approx \frac{p}{(1-p)\log(1/b)} \int_0^1 t^{\kappa} (1-t)^m \ {}^1 dt \tag{30}$$

$$= \frac{p}{(1-p)\log(1/b)} \frac{\Gamma(\kappa+1)\Gamma(m)}{\Gamma(\kappa+m+1)}, \tag{31}$$

where we have written

$$\kappa = \frac{\log[1/(1-p)]}{\log(1/b)} = \frac{\log(1-p)}{\log b}. \tag{32}$$

From the known asymptotic behaviour of the gamma function, we arrive at the approximation

$$P(\bar{X} = m) \approx \frac{p\Gamma(\kappa+1)}{(1-p)\log(1/b)} \, m^{\ 1 \ \kappa}. \tag{33}$$

The expression for $P(\bar{X} = m)$ as a series has been computed numerically using the Mathematica package for $p = b = 0.5$. This case corresponds to the predicted asymptotic form $P(\bar{X} = m) \sim (m^2 \log 2)^{\ 1}$. For $m \geq 2$ the values obtained by truncating the series at 100 terms and at 1000 terms agree to 6 significant figures, and the 1000-term estimates have been used in place of the exact result and compared with the predicted asymptotic form. We

find that $(m^2 \log 2)P(\bar{X} = m)$ is strictly increasing from 0.528112 at $m = 2$ to 0.999067 at $m = 1000$, giving no evidence of oscillation in the dominant behaviour. However, the subdominant behaviour is more interesting, even for these modest values of m. We write

$$P(\bar{X} = m) = \frac{1}{m^2 \log 2}[1 - \frac{A(m)}{m}], \tag{34}$$

and

$$\epsilon_m = 1 + \frac{\log[1 - (m^2 \log 2)P(\bar{X} = m)]}{\log m}. \tag{35}$$

The naively expected asymptotic behaviour would have $A(m)$ converging to a positive constant, and consequently $\epsilon_m \to 0$ as $m \to \infty$, while if $A(m) \sim cm^{\epsilon}$, we would obtain $\epsilon_m \to \epsilon$. We have computed $A(m)$ and ϵ_m for $m \leq 10000$. We observe that ϵ_m changes sign many times (see Table 1). The gap in $\log m$ between two successive sign changes of the same type is shown in the table, and that the gap converges to $\log 2 = 0.693147181$ is easily believed.

We also see oscillations by a direct plot of $A(m)$ against $\log m$ (Fig. 1). The growth of these oscillations is actually due to some small-amplitude oscillations in the coefficient of the dominant term which are not revealed in a superficial analysis. The sequence $(m^2 \log 2)P(\bar{X} = m)$ increases for $n \leq 1250$, attaining the local maximum value $0.999\,231\,7\cdots$ at $n = 1250$ and then decreases until $n = 1374$, where is attains the local minimum value $0.999\,226\,9\cdots$, and the alternation of increasing and decreasing behaviour persists.

3.2 The simplest model: proper analysis

We shall prove that as $m \to \infty$,

$$P(\bar{X} = m) = \frac{pQ(m)}{(1-p)\log(1/b)} m^{-1-\kappa} + O(m^{-2-\kappa}) \tag{36}$$

where $Q(x)$ is a function that is periodic in $\log x$ with period $\log(1/b)$, that is, $Q(x) \equiv Q(bx)$ for all $x > 0$. We shall write

$$P(\bar{X} = m) = \frac{p}{1-p} \left\{\phi(m-1) - \psi(m-1)\right\}, \tag{37}$$

where

$$\phi(x) = \sum_{n=1}^{\infty} [b(1-p)]^n e^{-b^n x} \tag{38}$$

and

$$\psi(x) = \sum_{n=1}^{\infty} [b(1-p)]^n \left\{ e^{-b^n x} - (1-b^n)^x \right\}. \tag{39}$$

Our analysis will reveal the dominant form of $\phi(x)$ and show that $\psi(x)$ is of lesser order. It is based on Mellin transform methods (cf. [9], Appendix 2).

The Mellin transform $\tilde{f}$ of a function f is defined and inverted by the formulae

$$\tilde{f}(z) = \int_0^{\infty} x^{z-1} f(x) dx \quad \text{and} \quad f(x) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} x^{-z} \tilde{f}(z) dz, \tag{40}$$

with the first integral restricted to those (complex) values of z for which the integral converges. The second integral is a contour integral along the vertical contour $\mathrm{Re}\{z\} = c$, placed in a strip in the z-plane in which the first integral converges. It is easily verified that the integral defining the Mellin transform of ϕ converges for $\mathrm{Re}\{s\} < 1 + \kappa$, with κ as defined above, and that

$$\tilde{\phi}(z) = \frac{\Gamma(s) b^{1-s} (1-p)}{1 - b^{1-s}(1-p)}. \tag{41}$$

The analytic continuation of $\tilde{\phi}(z)$ has simple poles to the right of the inversion contour at the points $z = \kappa + 1 + 2\pi k i / \log(1/b)$, where k is any integer. Translating the integration contour to the right and summing the residues from the poles that are crossed, we deduce that

$$\phi(x) = \frac{\Gamma(\kappa+1)\Phi(\log x / \log(1/b))}{\log(1/b) x^{1+\kappa}} + \frac{1}{2\pi i} \int_{c'-i\infty}^{c'+i\infty} x^{-z} f(z) dz \tag{42}$$

where $c' > \kappa + 1$ and

$$\Phi(\omega) = \sum_{k=-\infty}^{\infty} \frac{\Gamma(1 + \kappa + 2\pi k i / \log(1/b))}{\Gamma(1+\kappa)} \exp(-2\pi k \omega i). \tag{43}$$

The absence of any singularities further right than $\mathrm{Re}\{s\} = 1 + \kappa$ ensures that the integral on the right of Eq. (42) decays faster than any fixed power of x.

The rapid decay of $\Gamma(\sigma + i\tau)$, with

$$|\Gamma(\sigma + i\tau)| \sim (2\pi)^{1/2}|\tau|^{\sigma\ \ 1/2}\exp(-\pi|\tau|/2) \tag{44}$$

as the real parameter $\tau \to \pm\infty$ for fixed real σ, ensures the convergence of the doubly-infinite series, and as $\Gamma(\sigma - i\tau)$ is the complex conjugate of $\Gamma(\sigma + i\tau)$ we have

$$\Phi(\omega) = 1 + \sum_{k=1}^{\infty} 2\mathrm{Re}\Big\{\frac{\Gamma(1 + \kappa + 2\pi ki/\log(1/b))}{\Gamma(1+\kappa)}\ \exp(-2\pi k\omega i)\Big\}. \tag{45}$$

Having established what we need for $\phi(x)$, we turn to $\psi(x)$. We shall simply show that $\tilde{\psi}(s)$ has no poles for $0 < \mathrm{Re}\{s\} < 2+\kappa$, so that $\psi(x) = O(x^{\ \ 2\ \ \kappa+\delta})$ for all positive δ. The stronger assertion that $\psi(x) = O(x^{\ \ 2\ \ \kappa})$ requires a more detailed analysis that we shall not pursue. We use the inequality

$$\begin{aligned}
0 &< e^{\ b^n x} - (1-b^n)^x = \int_{b^n x}^{\log[(1\ \ b^n)^{\ \ 1}]x} e^{\ t} dt \\
&< xe^{\ b^n x}\{\log[(1-b^n)^{\ \ 1}] - b^n\} = xe^{\ b^n x}\sum_{k=2}^{\infty}\frac{b^{kn}}{k} \\
&< \frac{xe^{\ b^n x}b^{2n}}{2}\sum_{k=2}^{\infty}(b^n)^{k\ \ 2} = \frac{xe^{\ b^n x}b^{2n}}{2(1-b^n)} < \frac{xe^{\ b^n x}b^{2n}}{2(1-b)},
\end{aligned}$$

from which the required convergence of the Mellin transform for $0 < \mathrm{Re}\{s\} < 2 + \kappa$ follows easily. This completes the analysis. We have actually established that the dominant power law decay in $P(\bar{X} = n)$ is periodic in $\log(n-1)$ rather than $\log n$, but for large n this distinction may be dropped.

Concerning the magnitude of the non-constant terms in Eq. (45), which we already know to decrease rapidly with k, we observe that for $p = b = 0.5$ and $\kappa = 1$, the expansion becomes

$$\Phi(\omega) = 1 + \{0.0000749933\cos(2\pi\omega) + 0.0000500182\sin(2\pi\omega)\} + \cdots, \tag{46}$$

so the oscillations are of very small amplitude.

3.3 The general case

In the absence of explicit solutions for $G_n(s)$ for a general offspring PGF $g(s)$, our analysis has to rely on the functional equation (24). Functional equations

of this kind were first encountered in stochastic processes over 20 years ago by Hughes *et al.* [10] and by Shlesinger and Hughes [11], who observed their close analogy with real-space renormalization methods in statistical mechanics, and their antecedents in the classical analysis of Hardy and Littlewood, in the theory of nondifferentiable functions, and noncontinuable analytic functions. More recently, Gluzman and Sornette [12,13] have reviewed the existence of log-periodic oscillations mirroring underlying scale hierarchies in several areas of physics, and have developed classifications of such things in terms of complex dimensions and the decay of the coefficients in the expansion of the periodic coefficient function. Grabner and Woess [14] (see also [15], §16) have given an elegant discussion of random walks on the Sierpiński lattice where similar phenomena are encountered.

If $G^\dagger(s)$ is any solution of the functional equation (24), then the most general solution has the form

$$G(s) = G^\dagger(s) + H(s), \tag{47}$$

where $H(s)$ satisfies the homogeneous functional equation

$$H(s) = (1-p)H(g(s)). \tag{48}$$

If $G^\dagger(s)$ is chosen to be holomorphic at $s = 1$, then all critical behaviour of $G(s)$ at $s = 1$ resides in $H(s)$. Since the behaviour of $P(\bar{X} = m)$ for large m is reflected in the singularity structure of $G(s)$ at $s = 1$, we can attempt to extract the dominant behaviour of $P(\bar{X} = m)$ by examining the implications of Eq. (48) near $s = 1$. We write $s = 1 - \epsilon$, so that if the mean offspring per individual is $\mu < \infty$, we have $g(s) = 1 - \mu\epsilon + o(\epsilon)$. If we guess the asymptotic form

$$H(1-\epsilon) \sim \epsilon^\kappa \mathcal{Q}(\epsilon), \tag{49}$$

we find that

$$\epsilon^\kappa \mathcal{Q}(\epsilon) \sim (1-p)(\mu\epsilon)^\kappa \mathcal{Q}(\mu\epsilon) \tag{50}$$

and consistency is obtained by requiring

$$\mathcal{Q}(\epsilon) \sim \mathcal{Q}(\mu\epsilon) \quad \text{and} \quad (1-p)\mu^\kappa = 1, \tag{51}$$

so that we have a coefficient with log-periodic oscillations, and we recover a critical exponent $\kappa = \log[1/(1-p)]/\log\mu$. The standard identification of branch point behaviour $(1-s)^\kappa$ in a generating function with a contribution

proportional to $m^{-1-\kappa}$ suggests that $m^{-1-\kappa}$ behaviour dominates in the large-m expansion of $P(\bar{X} = m)$. By analogy with Tauberian Theorems, where slowly-varying coefficients $L((1-s)^{-1})$ indicate a factor of $L(n)$, one may guess that the log-periodicity of $\mathcal{Q}(1-s)$ is mirrored by a periodic function of $\log n$ of period $\log \mu$. Of course the precise conditions needed for the application of Tauberian Theorems or standard complex variable techniques do not hold in these cases, although a theorem of Odlyzko [16] establishes log-periodic behaviour rigorously for the coefficients f_n of solutions $f(z) = \sum_{n=0}^{\infty} f_n z^n$ of the functional equation $f(z) = P(z) + f(Q(z))$, where $P(z)$ and $Q(z)$ are polynomials, subject to the restrictions that $P(0) = Q(0) = Q'(0) = 0$. Our general analysis is therefore necessarily heuristic, but it is fully consistent with the rigorously analysed specific example above.

4 Discussion

Our model of the evolution of family names is constructed using a birth process with immigration to describe the creation of family names (in the limit of a long-enduring society), and a killed Galton–Watson process to describe the growth in the number of persons with a given family name, taking into account the time for which that name has existed. For a specific model of the offspring distribution for the Galton–Watson process we have rigorously proved that the probability that there are m individuals with a given family name decays for large m as $Q(m)m^{-1-\kappa}$, where $Q(m)$ is a bounded, log-periodic function and κ is determined by parameters associated with the birth process with immigration and the offspring distribution. We have argued that the same behaviour persists for more general offspring distributions, and that

$$\kappa = \frac{\log[1/(1-p)]}{\log \mu}, \tag{52}$$

where μ is the mean number of offspring per individual, $p = 1 - e^{-\lambda\Delta}$, Δ is the length of one generation, and λ describes the rate at which new surnames are created. We can rewrite κ in the form

$$\kappa = \frac{\lambda\Delta}{(\log \mu)}. \tag{53}$$

Since for a Galton–Watson process $E(Z_n) = \mu^n = e^{n \log \mu}$, we write $\mu = e^{\Delta\delta}$, where δ is the average rate of growth per unit time of a family. This gives

$$\kappa = \frac{\lambda}{\delta}. \tag{54}$$

The exponent κ therefore measures the ratio of the rate of surname generation to the rate of growth in size of surname families.

The fact that the empirical power-law exponent (corresponding to $-1-\kappa = -1-\lambda/\delta$ in our model) for names in USA and in Berlin [2] was found to be close to -2 (so that $\lambda \approx \delta$) suggests that the two growth rates are very similar in those populations. The empirical exponent for Japanese names [1] is close to -1.75; for Taiwanese names [17] around -1.9; and for Isle of Man names in 1881 [17] close to -1.5. This suggests that the average rate of growth in size of surname families exceeds the rate at which new names evolve from existing ones in all these three cases (by about 33% in Japan; 11% in Taiwan and by a factor of two in the Isle of Man).

References

[1] S. Miyazima, Y. Lee, T. Nagamine and H. Miyajima, *Physica A* **278** (2000) 282–288.

[2] D.H. Zanette and S.C. Manrubia, *Physica A* **295** (2001) 1–8.

[3] H.A. Simon, *Models of Man* (Wiley, New York, 1957).

[4] H.W. Watson and F. Galton, *J. Anthropol. Inst. Great Britain and Ireland* **4** (1874) 138–144.

[5] T.E. Harris, *The Theory of Branching Processes* (Springer, Berlin, 1963).

[6] G.R. Grimmett and D.R. Stirzaker, *Probability and Random Processes*, 2nd edition (Oxford University Press, 1992).

[7] P.D. Feigin, *J. Appl. Prob.* **16** (1979) 297–304.

[8] I.N. Sneddon, *Elements of Partial Differential Equations* (McGraw-Hill, New York, 1957).

[9] B.D. Hughes, *Random Walks and Random Environments*, Volume 1 (Oxford University Press, 1995).

[10] B.D. Hughes, M.F. Shlesinger and E.W. Montroll, *Proc. Nat. Acad. Sci. U.S.A.* **78** (1981) 3287–3291.

[11] M.F. Shlesinger and B.D. Hughes, *Physica A* **109** (1981) 597–608

[12] D. Sornette, *Physics Reports* **297** (1998), 239–270.

[13] S. Gluzman and D. Sornette, *Phys. Rev. E* **65**, 036142 (2002).

[14] P.J. Grabner and W. Woess, *Stochastic Proc. Appl.* **69** (1997) 127–138.

[15] W. Woess, *Random Walks on Infinite Graphs and Groups* (Cambridge University Press, 2000).

[16] A.M. Odlyzko, *Adv. Math.* **44** (1982) 180–205.

[17] W.J. Reed and B.D. Hughes, submitted to *Proc. Nat. Acad. Sci. USA.*

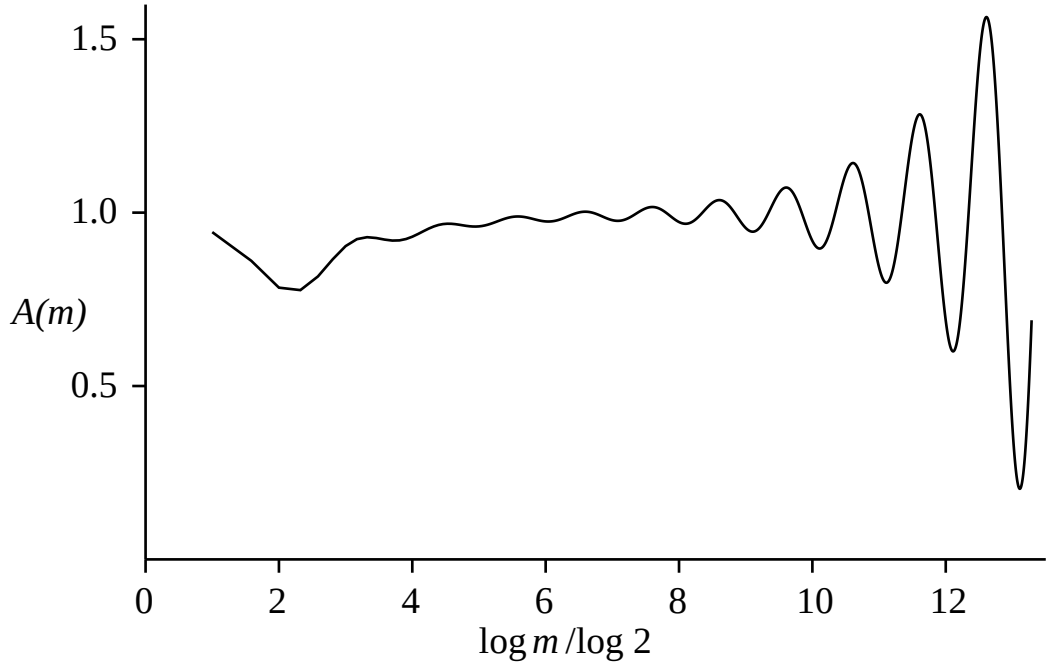

Fig. 1. Oscillations in the amplitude $A(m)$ are found when an attempt is made to analyse numerical values of $P(\bar{X} = m)$ for $2 \leq m \leq 10000$ using the naively anticipated form (34).

Table 1
Oscillations in the sequence ϵ_m defined by Eq. (35) when the naive asymptotic form (33) is used. The table shows values of m at which sign changes of the sequence occur. The gap in $\log m$ between two successive sign changes of the same type is well approximated by $\log 2 = 0.693147181$.

ϵ_m next becomes positive			ϵ_m next becomes negative		
m	$\log m$	gap in $\log m$	m	$\log m$	gap in $\log m$
90	4.499809670		104	4.644390899	
167	5.117993812	0.618184142	224	5.411646052	0.767255153
327	5.789960171	0.671966358	456	6.122492810	0.710846758
650	6.476972363	0.687012192	917	6.821107472	0.698614663
1299	7.169350017	0.692377654	1836	7.515344571	0.694237099
2598	7.862497197	0.693147181	3674	8.209036266	0.693691695
5197	8.555836815	0.693339618	7349	8.902319529	0.693283263