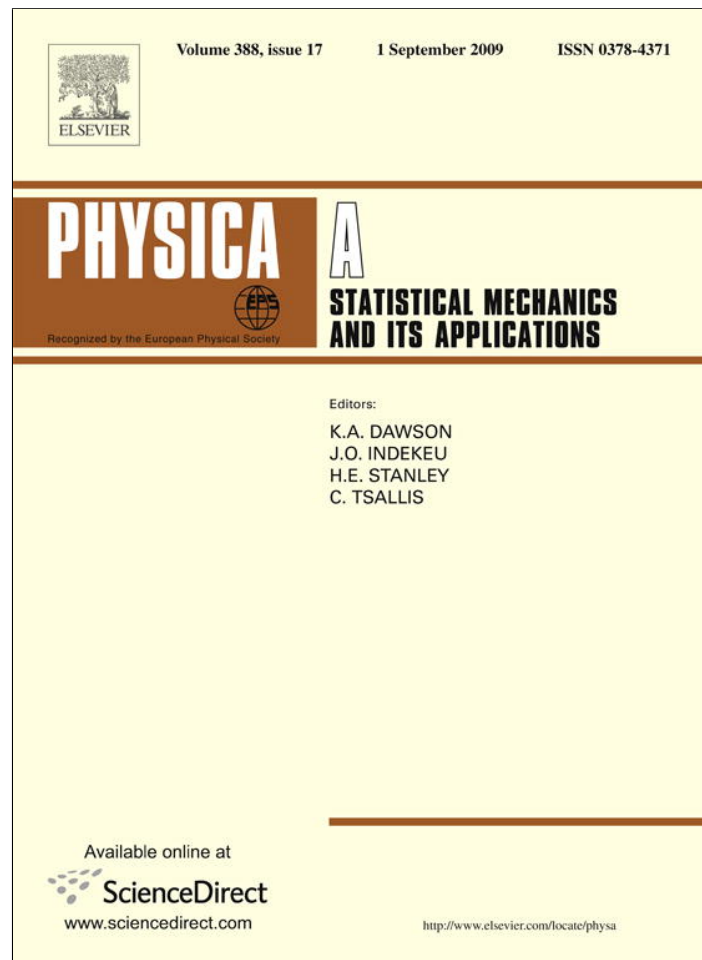


Provided for non-commercial research and education use.  
Not for reproduction, distribution or commercial use.



This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Contents lists available at ScienceDirect

Physica A

journal homepage: [www.elsevier.com/locate/physa](http://www.elsevier.com/locate/physa)

# Renormalization group evaluation of exponents in family name distributions

Andrea De Luca<sup>a,\*</sup>, Paolo Rossi<sup>b</sup><sup>a</sup> International School for Advanced Studies (Sissa), via Beirut 2-4, I-34014 Trieste, Italy<sup>b</sup> "E. Fermi" Physics Department, University of Pisa and I.N.F.N. Sezione di Pisa, Italy

## ARTICLE INFO

### Article history:

Received 23 December 2008

Received in revised form 22 March 2009

Available online 18 April 2009

### Keywords:

Branching process

Galton Watson

Renormalization group

Family name distribution

Power law

Population

## ABSTRACT

According to many phenomenological and theoretical studies the distribution of family name frequencies in a population can be asymptotically described by a power law. We show that the Galton–Watson process corresponding to the dynamics of a growing population can be represented in Hilbert space, and its time evolution may be analyzed by renormalization group techniques, thus explaining the origin of the power law and establishing the connection between its exponent and the ratio between the population growth and the name production rates.

© 2009 Elsevier B.V. All rights reserved.

## 1. Introduction

The frequency distribution of family names in local communities, regions and whole countries has been the object of a sustained interest by geneticists and statisticians for more than thirty years, starting from the seminal paper by Yasuda et al. [1]. For a recent review of the relevant literature we refer to Colantonio et al. [2], while Scapoli et al. [3] have recently collected and synthesized their results on the major countries of continental Western Europe. The main motivation for these researches resides in the deep analogy existing between surname distributions and the frequency of neutral alleles in a population: both distributions are generated by an evolutionary branching process subject to mutation and migration but not conditioned by natural selection. In particular it has been observed that the dynamics of family names, in countries with an European family name system, mimics that of the Y chromosome [4]. Models for such processes have been advanced in the genetic and statistical literature, starting from the Karlin–McGregor [5] statistical theory of neutral mutations. A significant theoretical evolution occurred in particular after Lasker's empirical observation [6] that a power law could offer a good fit of the observed surname distributions. As a consequence Panaretos [7] suggested the use of the Yule–Simon distribution, while Consul [8] proposed to employ the Geeta distribution with motivations coming from a branching process modelization. Evolutionary processes have attracted also the attention of physicists, who have found that neutral evolution might be grounds for the application of many techniques typical of statistical mechanics [9–11]. In particular Miyazima et al. [12], studying family name distributions in Japanese towns, found the systematic emergence of scaling laws, and further theoretical studies [13,14] justified the appearance of power laws of the Yule–Simon type in the case of growing populations with a non-vanishing probability for mutations. A different explanation was offered by Reed and Hughes [15] who considered a branching process with mutation and migration and found that the asymptotic form of the distributions should follow a power law. The most recent and comprehensive result is due to the Korean group of Baek et al. [16,17], who wrote down a master equation for the frequency distribution of family names and its time evolution in the presence of birth, death, mutation and migration, and found the possibility of different power laws with exponents depending on the

\* Corresponding author. Tel.: +39 040 3787 593.

E-mail address: [deluca@sissa.it](mailto:deluca@sissa.it) (A. De Luca).

mutation and migration parameters. In the present paper we reconsider the models of family name evolution in the context of a Hilbert space representation of branching processes, and show that distributions characterized by an asymptotic power law behavior can be obtained as solutions of recursive equations which correspond to the renormalization group equations of an (equivalent) physical system. In Section 2 we introduce and motivate our models. In Section 3 we discuss the simpler case of a system characterized by pure immigration without mutations. Finally in Section 4 we discuss the case with mutation. In Appendix we represent the Galton–Watson branching process in a Hilbert space.

## 2. The models

In the following sections we shall introduce two models, that take care of two different ways of generating new family names in a population: immigration from abroad and mutation occurring after reproduction. The importance of the appearance of new family names was pointed out in Refs. [13,14,16]. The analogy of the recursive equations we shall obtain with those typically derived by a renormalization–group approach to a physical system will allow us to evaluate the asymptotic behavior of the family name distribution  $N(k)$ , where  $N$  is the number of family names represented by exactly  $k$  individuals. Obviously in a typical real situation both immigration and mutation contribute to the dynamics of the family name distribution. But in our models we shall first focus on a population in which only immigration occurs, and then on one in which only mutation occurs. This simplification is justified by the fact that in an exponentially growing population (an approximation usually called Malthusian law) the effect of immigration can be neglected in comparison to mutation, at least in order to study the asymptotic behavior. However, in peculiar historical conditions, mutations can be heavily depleted and as a consequence the study of a society where name change is only due to immigration retains its value. Since we are interested in the family name distribution we can limit our attention to the male individuals, which is consistent with the legislation on names present in most real societies. In the following we shall use the term “individual” referring just to males. Moreover we shall suppose that the evolution of the population can be described by the Galton–Watson model. This means we shall consider:

- time as discrete, moving from one generation to the next;
- the system as completely markovian;
- each individual as independent of all others.

The last hypothesis may be considered a very strong restriction if applied to a biological system, since, for example, the exhaustion of resources induces a collective behavior, limiting the growing rate. But we can consider this hypothesis to be valid in the context of exponential grow of a population. It is useful to fix some definitions in the use of the Galton–Watson process. We set:

$$p_n = \text{probability for an individual to have } n \text{ sons} \tag{1}$$

It is straightforward to introduce the generating function of the Galton–Watson process:

$$f(z) = \sum_{n=1}^{\infty} p_n z^n. \tag{2}$$

Our hypothesis of growing population forces us to take  $p_n$  such that the mean number of sons is greater than one:

$$\sum_{n=1}^{\infty} n p_n = f'(1) \equiv m > 1.$$

We will exclude the trivial case:  $p_n = \delta_{1n}$ . We omit the explicit derivation of the recursive equations, which can be found with details in Appendix. However, their meaning will be somehow intuitive.

## 3. Immigration

We want to analyze a population whose members increase in number by the Galton–Watson mechanism and furthermore a group of individuals comes from outside. Each son inherits his family name from his father, while the new individuals coming from outside bring new family names. We are interested in the asymptotic behavior of  $N(k, t)$ , which corresponds to the number of family names represented by  $k$  individuals at time  $t$ . The values  $N(k, 0) = N_0(k)$  are assigned as initial conditions of the problem, with:

$$\sum_{k=1}^{\infty} N_0(k) = S_0 < \infty \quad \sum_{k=1}^{\infty} k N_0(k) = N_0 < \infty \tag{3}$$

where  $S_0$  is the initial number of family names and  $N_0$  is the initial number of individuals. We introduce the generating function:

$$n_t(z) = \sum_{k=0}^{\infty} N(k, t) z^k.$$

Now we suppose that the individuals from outside come always distributed in the same manner:  $\theta(k)$  is the number of new family names represented by  $k$  individuals among them. We suppose the number of individuals  $\theta_0$  and the number of new

family names  $G_0$  to be finite:

$$\begin{aligned} \theta_0 &= \sum_k \theta(k) \\ G_0 &= \sum_k k\theta(k). \end{aligned} \tag{4}$$

As before we introduce the generating function:

$$\theta(z) = \sum_k \theta(k)z^k.$$

We can obtain a recursive equation for  $n_t(z)$  involving  $\theta(z)$ . The explicit derivation is given in [Appendix](#):

$$n_{t+1}(z) = n_t(f(z)) + \theta(z) \tag{5}$$

A formal solution is given by:

$$n_t(z) = n_0(f_t(z)) + \sum_{k=0}^{t-1} \theta(f_k(z)).$$

where  $f_k(z)$  indicates the function  $f(z)$  iterated  $k$ -times. From this expression it is easy to compute the mean number of individuals  $N_t$  and the mean number of family names  $S_t$  at time  $t$ :

$$N_t \equiv n'_t(1) = n'_0(1)[f'(1)]^t \sum_{k=0}^{t-1} \theta'(1)[f'(1)]^k = N_0 m^t + G_0 \sum_{k=0}^{t-1} m^k = \left(N_0 + \frac{G_0}{m-1}\right) m^t - \frac{G_0}{m-1} \tag{6}$$

$$S_t \equiv n_t(1) = S_0 + t\theta_0. \tag{7}$$

We are interested in the limit  $t \rightarrow \infty$  and in the asymptotic behavior:  $k \gg 1$ . In order to achieve this goal, we notice that Eq. (5) is formally analogous to the equations coming from the renormalization group approach, linking the system at two different degrees of magnification. Therefore the system can be studied by using this analogy with the corresponding physical system. More explicitly, suppose  $\Phi_n(T)$  is the free energy of a hierarchical model, at scale  $n$  and temperature  $T$ . With standard renormalization group methods, we can obtain the recursive equation linking two different scales (see [18]):

$$\Phi_{n+1}(T) = g(T) + \frac{1}{\mu} \Phi_n(\phi(T)) \tag{8}$$

where  $g(T)$  is a regular function that comes up after summing the degree of freedom of the smaller scale and  $\phi(T)$  is the RG flow. Then near the critical point, for large  $n$ :

$$\Phi(T) \simeq (T - T_c)^\alpha \quad \alpha = \frac{\ln \mu}{\ln \phi'(T_c)}. \tag{9}$$

Eq. (5) is formally analogous to Eq. (8) and in our case the role of the flow is carried out by the Galton–Watson generating function  $f(z)$  and so the phases and the critical points correspond to the fixed points of  $f(z)$ :

$$f(z) = z. \tag{10}$$

From the fact that  $f(z)$  is convex and  $f'(1) > 1$ , we find that Eq. (10) has three solutions:  $q, 1, \infty$ .<sup>1</sup> From the Galton–Watson theory we know that  $q \in [0, 1)$  is the extinction probability. Moreover it is easy to see that  $f'(q) < 1$ . In fact if it were  $f'(q) \geq 1$  one would have by convexity:

$$f(1) > f(q) + f'(q)(1 - q) \geq 1.$$

So we have that  $q, \infty$  are attractive, while 1 is a repulsive fixed point which separates the two stable phases. We get a critical behavior near 1:

$$n(z) \equiv \lim_{t \rightarrow \infty} n_t(z) \simeq (1 - z)^\alpha.$$

One can see that in this case we have that for  $t \gg 1$ :

$$N(k, t) \simeq k^{-1+\alpha}. \tag{11}$$

To compute  $\alpha$ , we take  $\mu = 1, T_c = 1, m \equiv f'(1) = \phi(T_c)$  in Eq. (9) and we notice we are in an atypical situation in which  $\alpha = 0$ . It means that the function is diverging more slowly than any power and it is easy to see that it is logarithmic. In fact using Eqs. (6) and (7):

$$A \equiv \lim_{z \rightarrow 1} n'(z) m^{-\frac{n(z)}{\theta_0}} = \lim_{z \rightarrow 1} \lim_{t \rightarrow \infty} n'_t(z) m^{-\frac{n_t(z)}{\theta_0}} = \lim_{t \rightarrow \infty} \lim_{z \rightarrow 1} n'_t(z) m^{-\frac{n_t(z)}{\theta_0}} = \left(N_0 + \frac{G_0}{m-1}\right) m^{-\frac{S_0}{\theta_0}}. \tag{12}$$

<sup>1</sup> More precisely these are the possible outcomes of:  $\lim_{n \rightarrow \infty} f_n(z_0)$  for different values of  $z_0$ .

So we get near 1

$$n'(z) \simeq \left( N_0 + \frac{G_0}{m-1} \right) m^{\frac{n(z)-S_0}{\theta_0}} = Ae^{bn(z)}$$

where we set  $b = \frac{\log m}{\theta_0}$ . It can be solved<sup>2</sup> giving

$$n(z) \simeq -\frac{1}{b} (\log(Ab) + \log(1-z))$$

which ensures us the logarithmic divergence and implies for large  $k$ :

$$N(k) = \lim_{t \rightarrow \infty} N(k, t) = \frac{C}{k} (1 + o(1)).$$

So for immigrations we find a power-law behavior with exponent  $-1$ . Notice that this behavior is completely independent of the initial condition and of the distribution of the immigrating family names at each generation.

#### 4. Mutation

The context is analogous to the previous one but we no longer have immigration. We use again the initial condition in Eq. (3). Now, each son has a certain probability  $\rho$  that his family name mutates into a new one, different from his father's. We suppose that  $\rho$  does not depend on the family and we neglect the case in which two or more sons take the same new family name. This means the Galton–Watson contribution is modified since only a part proportional to  $1 - \rho$  of the offspring holds the same family name and the remaining part is added to the families of size 1. This implies the equation:

$$n_{t+1}(z) = n_t(f(z^{1-\rho})) + \rho m n'_t(1)z \tag{13}$$

where we used the fact that  $n'_t(1)$  equals the total number of individuals at generation  $t$ . Observe that mutations do not contribute to the total number of individuals and so:

$$n'_t(1) = N_0 m^t$$

as it can be shown directly via Eq. (13). The recursive equation can now be solved, at least formally. Defining  $r(z) = f(z^{1-\rho})$  and indicating by  $r_k(z)$  the function  $r(z)$  iterated  $k$ -times, we get the solution:

$$n_t(z) = n_0(r_t(z)) + \rho N_0 \sum_{n=0}^{t-1} m^{t-n} r_n(z) > \rho N_0 m^t r_0(z).$$

The last inequality shows that no limit in  $t$  can exist. However we can obtain a limit for the function:

$$\eta_t(z) \equiv n_t(z) m^{-t}$$

Since for large  $t$ :  $n_t(1) \propto m^t$ , as one can check by putting  $z = 1$  in Eq. (13), we are basically considering the distribution normalized to the total number of families. So we can put Eq. (13) in the form:

$$\eta_{t+1}(z) = \rho N_0 z + \frac{\eta_t(r(z))}{m} \tag{14}$$

which is again in the form of Eq. (8). However the flow is slightly changed with respect to the Galton–Watson generating function. We have  $r'(1) = (1 - \rho)m$  and we suppose  $\rho$  small enough for 1 to be a repulsive fixed point for the flow. In this case we must have a critical behavior near 1, whose exponent can be evaluated using Eq. (9):

$$\eta(z) = \lim_{t \rightarrow \infty} \eta_t(z) \simeq (1-z)^\alpha$$

where the exponent can be obtained using Eq. (9):

$$\alpha = \frac{\ln(m)}{\ln(r'(1))} = \frac{\ln(m)}{\ln(m) + \ln(1-\rho)}.$$

Using Eq. (11) we get the exponent of the family name power-law distribution:

$$\gamma \equiv \alpha + 1 = 2 - \frac{\ln(1-\rho)}{\ln(m) + \ln(1-\rho)} \simeq 2 - \frac{\rho}{\ln(m)}$$

where we considered  $\rho$  very small as it is true in the real situations (see [16]). Again the behavior is completely independent of the initial condition and shows the typical features of a scale-free system.

#### 5. Conclusion

In this paper we represented the Galton–Watson process as a quantum evolution defining the Hilbert space and the time evolution operator corresponding to the Galton–Watson probabilities. In this way we obtained two recursive equations

<sup>2</sup> The arbitrary constant can be fixed by imposing the solution diverges in 1.

for two possible models with different family name production mechanism: immigration and mutation. The structure of the branching allowed us to interpret these equations as the ones that connect different scales of a physical system and, in particular, the asymptotic behavior corresponds to the power law emerging near the critical point. The exponents are consistent with those evaluated in [16] with a master equation approach:  $N(k)$  goes as  $k^{-1}$  for a society where name change is only due to immigration and, approximately, as  $k^{-2}$  for a society where family name mutation occurs. Our method shows the robustness of these results, which are independent of the offspring distribution. Possible extensions of the model remain to be investigated and will be the object of future studies.

### Acknowledgment

A.D.L. thanks the MIUR grant “*Fisica Statistica dei Sistemi Fortemente Correlati all'Equilibrio e Fuori Equilibrio: Risultati Esatti e Metodi di Teoria dei Campi*” - 2007JHLPEZ, for partially supporting this work.

### Appendix. The Galton–Watson process in an Hilbert space

The structure of the branching process that characterizes the Galton–Watson process allows us to consider the reproduction governed by chance as a decay process whose interaction is given by a hamiltonian, which as we will see, is not hermitian. We first introduce the creation and destruction operators at each time with the usual commutation rules:

$$[a_k, a_h] = 0 \tag{15}$$

$$[a_k^\dagger, a_h^\dagger] = 0 \tag{16}$$

$$[a_k, a_h^\dagger] = \delta_{kh} \tag{17}$$

where, respectively,  $a_k^\dagger$  creates and  $a_k$  destroys an individual at time  $k$ . The Hilbert space is obtained in the usual way, acting on the vacuum Fock state with polynomials in  $a_t^\dagger$  for all possible values of  $t$ . A basis for the space is then given by the following set:

$$|n, t\rangle = (a_t^\dagger)^n |0\rangle \quad \langle n, t| = \langle 0|(a_t)^n. \tag{18}$$

Then, at each time  $t$ , the state of the system, which is determined by the probability  $b_k(t)$  that exactly  $k$  individuals are present, can be written:

$$|\phi(t)\rangle = \sum_k b_k(t) |n, t\rangle.$$

It must be possible to connect the dynamics to the parameters  $p_n$  introduced in Eq. (1) and so to the generating function  $f(z)$  of Eq. (2). This can be done setting the hamiltonian as:

$$H(t) = f(a_{t+1}^\dagger) a_t \tag{19}$$

we can write the time-evolution operator:

$$U(t) \equiv \exp(H(t)) \tag{20}$$

which is the one-time-step evolution operator: it evolves the states at time  $t$  to time  $t + 1$ ,<sup>3</sup> giving the correct probabilities according to the Galton–Watson process. In fact, one can easily check that:

$$U(t)|1, t\rangle = U(t)a_t^\dagger|0\rangle = f(a_{t+1}^\dagger)|0\rangle = \sum_n p_n (a_t^\dagger)^n |0\rangle. \tag{21}$$

And in general, by linearity we know that given a state  $|\phi(t)\rangle$  with a particular probability distribution we get the state at time  $t + 1$  correctly evolved. Starting from a state  $|\phi_0\rangle$  at time 0 we can obtain the state at time  $T$  by:

$$|\phi_T\rangle = \mathcal{U}(T)|\phi_0\rangle \equiv U(T - 1)U(T - 2) \cdots U(0)|\phi_0\rangle \tag{22}$$

<sup>3</sup> It should be observed that the correct expression for  $U(t)$  should be:

$$U(t) = P_t e^{H(t)}$$

where  $P_t$  destroys all the states at time  $t$ :  $P_t|0\rangle = |0\rangle$ ,  $P_t(a_t^\dagger)^n|0\rangle = 0$  and  $\forall h \neq t$   $[P_t, a_h^\dagger] = 0$ . In this way we eliminate all the parts of the states that do not evolve to time  $t + 1$ . E.g.:

$$e^{H(t)}|2, t\rangle = \left( f(a_{t+1}^\dagger) a_t + \frac{f(a_{t+1}^\dagger)^2 (a_t)^2}{2} \right) (a_t^\dagger)^2 |0\rangle = \left( 2f(a_{t+1}^\dagger) a_t^\dagger + f(a_{t+1}^\dagger)^2 \right) |0\rangle$$

and the operator  $P_t$  then eliminates the first term in the parenthesis giving the correct result.

and this equation defines the full time-evolution operator. We see now how to derive Eq. (5). We want to write the equation of evolution for:

$$|n(t)\rangle = \sum_{k=0}^{\infty} N(k, t)|k, t\rangle.$$

With the notation of Section 3, we define the state:

$$|\theta(t)\rangle = \sum_{k=0}^{\infty} \theta(k)|k, t\rangle.$$

In the absence of immigration the evolution would be simply given by Eq. (22). But here at each step, the number of family names represented by  $k$  individuals grows due to individuals coming from outside. So we get the equation:

$$|n(t+1)\rangle = U(t)|n(t)\rangle + |\theta(t)\rangle \tag{23}$$

Now we use the map  $\mathcal{W}$  from the Hilbert space to  $C^\infty[0, 1]$  defined on the basis of Eq. (18) as:

$$|n, t\rangle = (a_t^\dagger)^n |0\rangle \xrightarrow{\mathcal{W}} z_t^n. \tag{24}$$

And in general:  $\mathcal{W}(|\phi(t)\rangle) = \phi(z_t) \in C^\infty[0, 1]$ . The action of an operator becomes an integral transformation. For  $U(t)$  we have a simple kernel of the form:  $U(t) \rightarrow U(z_t, z_{t+1}) = \delta(z_t - f(z_{t+1}))$  as can be deduced from Eq. (21). Then:

$$\phi_{t+1}(z_{t+1}) = \mathcal{W}(U(t)|\phi(t)\rangle) = \int U(z_t, z_t + 1)\phi(z_t, t) dz = \phi_t(f(z_{t+1}))$$

where  $\phi_t(z) = \mathcal{W}(|\phi(t)\rangle)$ . Acting on the Eq. (23) with the map  $\mathcal{W}$  we get Eq. (5).

## References

- [1] N. Yasuda, L.L. Cavalli-Sforza, M. Skolnick, A. Moroni, The evolution of surnames: An analysis of their distribution and extinction, *Theor. Popul. Biol.* 5 (1974) 123.
- [2] S.E. Colantonio, G.W. Lasker, B.A. Kaplan, V. Fuster, Use of surname models in human population biology: A review of recent developments, *Human Biology* 75 (2003) 785–807.
- [3] C. Scapoli, E. Mamolini, A. Carrieri, A. Rodriguez-Larralde, I. Barraï, Surnames in Western Europe: A comparison of the subcontinental populations through isonymy, *Theor. Popul. Biol.* 71 (2007) 37–48.
- [4] B. Sykes, C. Irven, Surnames and the y chromosome, *Am. J. Hum. Genet.* 66 (2000) 1417–1419.
- [5] S. Karlin, J. McGregor, The number of mutant forms maintained in a population, in: *Proc. 5th Berkeley Symp. Math. Stat. Prob.*, vol. 4, 1967, pp. 415–438.
- [6] W.R. Fox, G.W. Lasker, The distribution of surname frequencies, *Int. Stat. Review* 51 (1983) 81–87.
- [7] J. Panaretos, On the evolution of surnames, *Int. Stat. Review* 57 (1989) 161.
- [8] P.C. Consul, Evolution of surnames, *Int. Stat. Review* 59 (1991) 271.
- [9] B. Derrida, L. Peliti, Evolution in a flat fitness landscape, *Bull. Math. Biol.* 53 (1991) 355.
- [10] M. Serva, L. Peliti, A statistical model of an evolving population with sexual reproduction, *J. Phys. A: Math. Gen.* 24 (1991) L705.
- [11] B. Derrida, S.C. Manrubia, D.H. Zanette, Statistical properties of genealogical trees, *Phys. Rev. Lett.* 82 (1999) 1987.
- [12] S. Miyazima, Y. Lee, T. Nagamine, H. Miyajima, Power law distribution of family names in Japanese societies, *Physica A* 278 (2000) 282.
- [13] D.H. Zanette, S.C. Manrubia, Vertical transmission of culture and distribution of family names, *Physica A* 295 (1–2) (2001) 1–8.
- [14] S.C. Manrubia, D.H. Zanette, At the boundary between biological and cultural evolution: The origin of surname distributions, *J. Theor. Biol.* 216 (2002) 461.
- [15] W.J. Reed, B.D. Hughes, On the distribution of family names, *Physica A* 319 (2003) 579.
- [16] S.K. Baek, H.A.T. Kiet, B.J. Kim, Family name distributions: Master equation approach, *Phys. Rev. E* 76 (2007) 046113.
- [17] B.J. Kim, S.M. Park, Distribution of Korean family names, *Physica A* 347 (2005) 683.
- [18] B. Derrida, L. de Sze, C. Itzykson, Fractal structure of zeros in hierarchical models, *J. Stat. Phys.* 33 (1983) 559–569.