Renormalization group evaluation of exponents in family-name distributions

Andrea De Luca, Paolo Rossi

December 15, 2008

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 growth and the rates of production of new family-names.

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, mimicks 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 a ground for application of many techniques proper of statistical mechanics [9] [10] [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 nonvanishing 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 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 behaviour can be obtained as solutions of recursive equations which would correspond to the renormalization group equations of an (equivalent) physical system. In Sec. 2 we introduce and motivate our models. In Sec 3 we represent the Galton-Watson branching process in a Hilbert space. In Sec. 4 we discuss the simpler case of a system characterized by pure immigration without mutations. Finally in Sec. 5 we discuss the general case with mutation.

2 The models

In the following sections, we will introduce two models, that take care of two different way of generating new family names in a population: immigrations of new foreign names, and mutations that can occur after reproduction. The importance of new family name production is pointed out in the works [13, 14, 16]. We will see that the analogy, between the recursive equations, we will obtain, with the ones typically derived from a renormalization-group-approach to a physical system, will allow us to evaluate the asymptotic behaviour of the family-name distribution, that is:

$$N(k) \equiv \#\{\text{family-names represented exactly by } k \text{ individuals}\} \tag{1}$$

Supporting this kind of approach, it is worth noting that statistical and experimental studies put in evidence the power-law trend of this distribution, which is very common in physical system near critical point. We must say that obviously in a typical real situation both immigration and mutation contribute to the dynamic of family-name distribution. But in our models we will concentrate first in a population in which only immigration occurs; then on one in which only mutation occurs. This simplification is justified by the fact that, as we will see in detail later, in a population that is growing exponentially, which is a good approximation usually called Malthusian law, the effect of immigration can be neglected, comparing to mutation, at least in order to study the asymptotic behaviour. However, in peculiar historical condition, mutations can be heavy depleted and so the study of a pure-immigrating society retains its value. Since we are interested in the family-name distribution we can limit our attention to the male individuals of the population, which is consistent with the actual law present in the most of real societies. Then, in the rest, we will use the term 'individual' referring just to the males. Moreover, we will suppose that the evolution of the population can be described using the Galton-Watson model. It means we will consider:

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

At least the last one maybe considered a very strong restriction if applied to a biological system, since, for example, the exhaustion of resources induces a collective behaviour, limiting the growing rate. But we can consider this hypothesis valid in the contest of exponential grow of a population. It is useful to fix some definitions in the use of the Galton-Watson process. We set:

$$p_n = \text{probability for an individual to have } n \text{ sons} \tag{2}$$

It is straightforward to introduce for practical calculation the generating function of the Galton-Watson process:

$$f(z) = \sum_{n=1}^{\infty} p_n z^n \tag{3}$$

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

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

We will exclude the trivial case: $p_n = \delta_{1n}$. Before going into the details of the two models, it is useful to show how the Galton-Watson process can be seen as the evolution associated to a particular kind of hamiltonian. We will see it in the next section.

3 The Galton-Watson process in an Hilbert space

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

$$\begin{aligned} [a_k, a_h] &= 0 && (4)\\ [a_k^\dagger, a_h^\dagger] &= 0 && (5)\\ [a_k, a_h^\dagger] &= \delta_{kh} && (6) \end{aligned}$$

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

$$|n,t\rangle = (a_t^\dagger)^n|0\rangle \qquad \langle n,t| = \langle 0|(a_t)^n$$

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

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

The evolution of the state is connected to the parameters p_n of the process, and so to Eq. (3). In fact, setting the hamiltonian as:

$$H(t) = f(a^\dagger_{t+1})a_t \tag{7}$$

where $f(x)$ is given by Eq. (3), we can write the time-evolution operator:

$$U(t) = \exp(H(t)) \tag{8}$$

which evolves the states at time t to time $t+1$[1]. In fact, we have:

$$U(t)|1,t\rangle = U(t)a^\dagger_t|0\rangle = f(a^\dagger_{t+1})|0\rangle$$

And in general, by linearity we know that given a state $|\phi\rangle(t)$ with a particular probability distribution we get the state at time $t+1$, correctly evolved according to Galton-Watson process. In the following, it will be useful to represent the hilbert space in $C^\infty[0,1]$:

$$|n,t\rangle = (a^\dagger_t)^n|0\rangle \quad \xrightarrow{\mathcal{P}} \quad z_t^n \tag{9}$$

In this way, $U(t)$ can take a simple integral form: $U(t) \to U(z_t,t) = \delta(z_t - f(z_{t+1}))$ such that:

$$\phi(z_{t+1}, t+1) = \mathcal{P}(U(t)|\phi\rangle(t)) = \int U(z,t)\phi(z,t)dz = \phi(f(z_{t+1}),t)$$

where $\phi(z,t) = \mathcal{P}(|\phi(t)\rangle)$. We can now investigate the models in detail.

4 Immigration

We want to analyse a population whose members increase through the Galton-Watson mechanism and furthermore a group of individuals comes from outside. Each son inherits is family-name from his father, while the new individuals coming from outside bring new family-names. We are interested in the asymptotic behaviour of $N(k,t)$, which as in Eq. (1), 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:

$$\sum_{k=1}^{\infty} N_0(k) = S_0 < \infty \qquad \sum_{k=1}^{\infty} kN_0(k) = N_0 < \infty$$

where S_0 is the initial number of family-names and N_0 is the initial number of individuals. Is it useful to interpret the $N(k,t)$ as a state in the hilbert state we introducted in the previous section:

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

[1] 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:

$$P_t|0\rangle = |0\rangle \qquad P_t(a^\dagger_t)^n|0\rangle = 0$$

$$\forall h \neq t \quad [P_t, a^\dagger_h] = 0$$

in this way we eliminate all the parts of the states that do not evolve correctly to time $t+1$.

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. In the same way we introduce a state:

$$|\theta(t)\rangle = \sum_k \theta(k)|n,t\rangle$$

Within this formalism, it is easy to write an equation connecting the t distribution to the $t+1$ one:

$$|n(t+1)\rangle = U(t)|n(t)\rangle + |\theta(t)\rangle$$

where clearly the already present individuals evolve according to the $U(t)$ operator of the Galton-Watson, and the state of the outside-people is added. Converting this equation with the representation of Eq. (9) we get:

$$n_{t+1}(z) = n_t(f(z)) + \theta(z) \tag{10}$$

A formal solution is given by:

$$n_t(z) = n_0\left(f^{(t)}(z)\right) + \sum_{k=0}^{t-1} \theta\left(f^{(k)}(z)\right)$$

We are interested in the limit $t \to \infty$ and in the asymptotic behaviour: $k \gg 1$. In order to achieve this aim, we notice that the equation (10) is formally analogous to the equations coming from the renormalization group approach, linking the system at two different degrees of magnification. So the system can be studied by using this analogy with the corresponding physical system. The role of the flow is carried out by the Galton-Watson generating function $f(z)$ and so the phases and the critical points corresponds to the fixed points of $f(z)$:

$$f(z) = z \tag{11}$$

From the fact that $f(z)$ is convex being positive all its derivatives and $f'(1) > 1$, we find that Eq. (11) has three solutions: $q, 1, \infty$. 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 was $f'(q) \geq 1$ one would have by convexity:

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

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

$$n(z) = \lim_{t\to\infty} n_t(z) \simeq (1-z)^\alpha$$

One can see that in this case we have that for t large:

$$N(k,t) \simeq k^{-1+\alpha} \tag{12}$$

To evaluate α we use the renormalization group equations; given

$$F_{n+1}(T) = g(T) + \frac{1}{\mu} F_n(\phi(T)) \tag{13}$$

we get near the critical point:

$$F(x) \simeq (T-T_c)^\alpha \qquad \alpha = \frac{\ln \mu}{\ln \phi'(T_c)} \tag{14}$$

In our case $\mu = 1$, $m \equiv f'(1) = \phi(T_c)$ 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:

$$\lim_{z\to 1} \frac{n'(z)}{\exp[\ln(m)n(z)]} = \lim_{z\to 1}\lim_{t\to\infty} \frac{n_t'(z)}{\exp[\ln(m)n_t(z)]} =$$
$$= \lim_{t\to\infty}\lim_{z\to 1} \frac{n_t'(z)}{\exp[\ln(m)n_t(z)]} = \lim_{t\to\infty} \frac{1}{(m-1)}\frac{m^{t+1}-m}{m^t} = \frac{m}{m-1} \tag{15}$$

So we get near 1

$$n'(x) \simeq \frac{m^{n(x)+1}}{m-1}$$

which can be solved giving

$$n(x) \simeq -\frac{\ln(1-x)}{m}$$

which implies for large k:

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

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

5 Mutation

The contest is analogous to the previous one but we do not have immigration anymore. Now, each son has a certain probability ρ that his family-name mutates in a new one, different from that of his father. We suppose that ρ does not depends on the family and we neglect the case in which two or more sons changes in the same way their family-name. It 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 family of size 1. It results in the equation:

$$n_{t+1}(z) = n_t\left(f\left(z^{1-\rho}\right)\right) + \rho m n_t'(1) z \tag{16}$$

where we used the fact that $n_t'(1)$ equals the total number of individual at generation t. Observe that mutations does not contribute to the total number of individual and so:

$$n_t'(1) = m^t$$

and indicating by $r(z) = f\left(z^{1-\rho}\right)$ and by $r_n(z)$ its iterate, we get the formal solution:

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

The last inequality shows that no limit can exist. However renormalizing we can obtain a limit:

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

Since $n_t(1) \simeq m^t$ in this way we are considering the distribution renormalized to the total number of families. So we can put Eq. (16) in the form:

$$\eta_{t+1}(z) = \rho z + \frac{\eta_t(r(z))}{m} \tag{17}$$

which is again of the form of Eq. (13). However the flow is slightly changed respect to the Galton-Watson generating function. We have $r'(1) = (1-\rho)m$ and we suppose ρ small enough that 1 is still a repulsive fixed point for the flow. In this case we must have a critical behaviour near 1, whose exponent can be evaluated using Eq. (14):

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

where:

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

Using Eq. (12) 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 ρ very small as it is true in the real situations (see [16]). Again the behaviour is completely independent of the initial condition and shows the typical features of a scale-free system.

6 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 for two possible models with different family-name production mechanisms: immigrations and mutations. 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 behaviour corresponds to the critical power law coming out 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 pure immigrating society and approximately as k^{-2} for society in which family name mutations occurs. Our method shows the robustness of this results, that are independent of the offspring distribution. Some important prosecutions remain to be investigated with this approach as the model in which the mutation rate depends on the dimension of the family.

References

[1] M. S. N. Yasuda, L.L. Cavalli-Sforza and A. Moroni, "The evolution of surnames: An analysis of their distribution and extinction," *Theor. Pop. Biol.*, vol. 5, p. 123, 1974.

[2] B. K. S.E. Colantonio, G.W. Lasker and V. Fuster, "Use of surname models in human population biology: A review of recent developments," *Human Biology*, vol. 75, pp. 785–807, 2003.

[3] A. C. C. Scapoli, E. Mamolini, A. Rodriguez-Larralde, and I. Barrai, "Surnames in western europe: A comparison of the subcontinental populations through isonymy," *Theor. Pop. Biol.*, vol. 71, pp. 37–48, 2007.

[4] B. Sykes and C. Irven, "Surnames and the y chromosome," *Am. J. Hum. Genet.*, vol. 66, pp. 1417–1419, 2000.

[5] S. Karlin and J. McGregor, "The number of mutant forms maintained in a population," in *Proc. 5th Berkeley Symp. Math. Stat. Prob.*, vol. 4, pp. 415–438, 1967.

[6] W. Fox and G. Lasker, "The distribution of surname frequencies," *Int. Stat. Review*, vol. 51, pp. 81–87, 1983.

[7] J. Panaretos, "On the evolution of surnames," *Int. Stat. Review*, vol. 57, p. 161, 1989.

[8] P. C. Consul, "Evolution of surnames," *Int. Stat. Review*, vol. 59, p. 271, 1991.

[9] B. Derrida and L. Peliti, "Evolution in a flat fitness landscape," *Bull. Math. Biol.*, vol. 53, p. 355, 1991.

[10] M. Serva and L. Peliti, "A statistical model of an evolving population with sexual reproduction," *J. Phys. A: Math. Gen.*, vol. 24, p. L705, 1991.

[11] S. M. B. Derrida and D. Zanette, "Statistical properties of genealogical trees," *Phys. Rev. Lett.*, vol. 82, p. 1987, 1999.

[12] T. N. S. Miyazima, Y. Lee and H. Miyajima, "Power law distribution of family names in japanese societies," *Physica A*, vol. 278, p. 282, 2000.

[13] D. H. Zanette and S. C. Manrubia, "Vertical transmission of culture and distribution of family names," *Physica A*, vol. 295, p. 1, 2001.

[14] S. C. Manrubia and D. H. Zanette, "At the boundary between biological and cultural evolution: the origin of surname distributions," *J. Theor. Biol.*, vol. 216, p. 461, 2002.

[15] W. Reed and B. Hughes, "On the distribution of family names," *Physica A*, vol. 319, p. 579, 2003.

[16] H. K. S.K. Baek and B. Kim, "Family name distributions: Master equation approach," *Phys. Rev. E*, vol. 76, p. 046113, 2007.

[17] B. Kim and S. M. Park, "Distribution of korean family names," *Physica A*, vol. 347, p. 683, 2005.