Entropy **2015**, *17*, 425-437; doi:10.3390/e17010425

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Self-Similarity in Population Dynamics: Surname Distributions and Genealogical Trees

Paolo Rossi

Dipartimento di Fisica dell'Università di Pisa and I.N.F.N., Largo B. Pontecorvo 3, I-56127 Pisa, Italy; E-Mail: paolo.rossi@df.unipi.it; Tel.: +39-050-2214884

Academic Editor: Giorgio Kaniadakis

Received: 13 December 2014 / Accepted: 13 January 2015 / Published: 19 January 2015

Abstract: The frequency distribution of surnames turns out to be a relevant issue not only in historical demography but also in population biology, and especially in genetics, since surnames tend to behave like neutral genes and propagate like Y chromosomes. The stochastic dynamics leading to the observed scale-invariant distributions has been studied as a Yule process, as a branching phenomenon and also by field-theoretical renormalization group techniques. In the absence of mutations the theoretical models are in good agreement with empirical evidence, but when mutations are present a discrepancy between the theoretical and the experimental exponents is observed. Hints for the possible origin of the mismatch are discussed, with some emphasis on the difference between the asymptotic frequency distribution of a full population and the frequency distributions observed in its samples. A precise connection is established between surname distributions and the statistical properties of genealogical trees. Ancestors tables, being obviously self-similar, may be investigated theoretically by renormalization group techniques, but they can also be studied empirically by exploiting the large online genealogical databases concerning European nobility.

Keywords: surname distribution; isonymy; consanguinity; population; Yule process; branching process; renormalization group; Y chromosome; sampling; genealogical trees

PACS classifications: 87.23.-n; 89.75.Da; 05.10.-a

1. Introduction and Historical Background

The frequency distribution of family names has been an interesting issue in human biology since the last quarter of the nineteenth century. Surnames have a cultural origin, but their propagation usually follows definite rules linked to the reproductive behavior, exactly as it happens to genes, and more specifically to the so-called neutral genes, not affecting the phenotype and therefore not subject to selective pressure. In many human cultures, the propagation rules of surnames are the same as those of the Y chromosome, that is, preserved in the male descendants and is only affected by mutation (and by false paternity).

The very first statistical studies on surname frequency appeared in England around 1875, when George Darwin analyzed marriage isonymy (coincidence of surnames) as a tool for the evaluation of inbreeding in the English society [1], while Galton and Watson started the theory of branching processes by computing the probability of surname extinction, a phenomenon then perceived as a signal of physical decline in English aristocracy [2].

Theoretical weakness of the approaches and the lack of adequate statistical data led to a long period of latency in the mathematical study of surname distribution. A substantial revival occurred only until the 1960s, triggered by a number of new and important results. In 1965 Crow and Mange proposed a model for consanguinity (blood relationship) estimates based on marital isonymy [3], paving the way to a large number of theoretical and phenomenological studies. On the other side, Karlin and McGregor in 1967 produced a statistical theory of the behavior of neutral mutations in finite and constant populations [4], which led geneticists to a systematic study of surname distribution and extinction as an empirical model for neutral gene propagation. It is worth noticing that in a realistic limit Karlin and McGregor's results lead to the distribution introduced by Fisher *et al.* in 1947 in order to describe the relation between the number of species and the number of individuals in an ecosystem [5].

The anthropologist G.W. Lasker extended the use of isonymy to the study of consanguinity between populations [6] and observed (with W.R. Fox [7]) that the empirical frequency distribution of surnames could be accurately described by means of a discrete Pareto (power law) distribution.

A wide collection of results was presented in the 1982 Eugene conference on surnames as biological markers of inbreeding and migration [8] and in Lasker's book published in 1985 [9].

In the last 30 years, also thanks to the greatly increased availability of digital and online databases, a large amount of quantitative studies have been performed, concerning both European and American countries, thus creating the conditions for a more systematic phenomenological approach to the issue. The most significant references are included in a recent paper by Boattini *et al.* [10]. A very detailed discussion of the theoretical aspects and experimental results related to the connection between surnames and Y chromosome types is offered in a recent volume by Redmonds, King and Hey [11].

Different parametrizations of the observed frequency distributions were proposed, based on statistical models of reproductive behavior and population dynamics. Especially relevant was the observation that surname dynamics can be seen as a Yule process [12], thus leading to a statistical explanation for the appearance of scaling in a proper limit. A mathematical representation of a generic Yule process was offered in 1955 by the stochastic model introduced by Simon [13] and subsequently considered by many authors.

The ubiquity of power laws in the description of both natural and social phenomena, whenever there is no intrinsic characteristic scale, is a phenomenon that the modern theory of complexity aims to explain [14]. The presence of scaling in surname distributions did therefore attract the attention of statistical physicists.

Starting from the late 1980s they began to develop stochastic models trying to catch some aspects of evolutionary dynamics, especially in connection with the growing diffusion of the neutral theory of molecular evolution and with the contemporary developments in the theory of disordered systems. The dynamics of populations in flat fitness landscapes was studied [15], both in asexual [16] and in sexual [17] reproduction.

In more recent times the attention of physicists focused on the problem of identifying dynamical models depending on a minimal number of free parameters but still retaining sufficient descriptive and predictive power. Different approaches converged in identifying birth and mortality rates, and especially migration and mutation rates, as the crucial parameters needed for the parametrization of surname frequency distribution and evolution. An extended review of these approaches has been recently published [18].

The most important recent contribution to phenomenological description and model building was offered in 2007 by the master equation approach of Baek, Kiet and Kim, whose formal solution allows for the (statistical) prediction of the surname distribution as a function of time once the initial distribution is given and the four above mentioned rates are assigned [19]. This model is briefly described (without technical details) in Section 2. It encompasses many previous models and leads to many testable predictions, which have been verified by the authors in the case of Far Eastern countries.

The behaviors predicted by solving the master equation can be independently reproduced by applying renormalization group techniques, as well as other stochastic dynamics methods, to the description of the time evolution of surname distributions. A cursory review of the above results is presented (without proofs) in Section 3.

Section 4 is devoted to a synthetic presentation of some published results [20] concerning the effects of sampling on frequency distributions, which may be relevant for a correct parametrization and a proper analysis of empirical data.

Also the statistical properties of genealogical trees (which may be relevant to the issues of inbreeding and surname extinction) were explored from a theoretical point of view [21–23]. The main results obtained in this context are recalled (without proofs) in Section 5.

However no strict connection was yet established between the frequency distribution of ancestors and the distribution of surnames in the relevant population. One of the main aims of the present paper is to establish such connection. In Section 6 we show that simple assumptions lead to the determination of a functional relationship between the surname distributions found in different generations of ancestors of people belonging to a definite social group and the surname distribution characterizing the full population of the group at the corresponding time in the past. This relationship may allow for the extraction of nontrivial properties of the population that cannot be directly extracted from available data.

Finally in Section 7 we present some preliminary results obtained from the analysis of a wide set of empirical data extracted from (almost) complete ancestors tables, spanning all the modern age, for several members of the European nobility.

2. The Evolution of Surnames as a Yule Process and the Master Equation Approach

Baek *et al.* showed that population dynamics can be expressed in terms of a master equation, whose variables are the probabilities $P_{j,s}(k,t)$ for a family to have k members at time t if the number of members at time s was j. The time evolution of $P_{j,s}(k,t)$ is governed by the differential equation [19]

$$\frac{dP_{j,s}(k,t)}{dt} = \lambda(k-1)P_{j,s}(k-1,t)+[\mu(k+1)+\beta(k+1)]P_{j,s}(k+1,t)-[\lambda(k)+\mu(k)+\beta(k)]P_{j,s}(k,t),$$

where time is treated as a continuous variable.

The parametric dependence on the birth, death and surname change probabilities at time t is expressed by the (time-dependent) functions $\lambda(k)$ (birth rate), $\mu(k)$ (death rate) and $\beta(k)$ (surname creation rate).

With the simplifying assumption that all these probabilities are proportional to k and have the same time dependence, the master equation can be formally solved for assigned values of j and s.

Therefore, the overall distribution of surnames in the population $P(k,t)$ can be obtained once the number of surnames appeared at each time s is known. The above described model is quite general, and it encompasses in particular the Simon model.

It is possible to represent the case of an exponentially growing population in which both mutation and migration are present, assuming that the number of surnames appearing at each time has two components, of which one is constant in order to take into account immigration, while the other is proportional to N, in order to take into account mutations.

In the absence of mutations the probability for a family to have k members at time t is given by a (time-dependent) Fisher distribution. These results seem to describe quite accurately the case of Korea [24] and China, as they are known on historical grounds.

When admitting the possibility of mutations, it is possible to show that the asymptotic behavior of $P(k,t)$ is proportional to $k^{-\gamma}$ where

$$\gamma = 2 + \frac{\beta}{\delta - \beta}.$$

The exponent of the distribution stays near to 2, and in any case it depends on the growth ratio $\delta \equiv \lambda - \mu$ and on the mutation coefficient β.

Comparison with previous results by Manrubia and Zanette [25,26] may be obtained recalling the fact that all rates are referred to different time scales.

3. Alternative Approaches

The treatment of critical phenomena based on the renormalization group techniques may very well apply to the description of evolutionary models, despite the still controversial status of self-organized criticality intended as a universal explanation for evolutionary dynamics [27–30].

The applicability of field theoretical methods to classical evolution processes can be traced back to the Fock space formalism for classical objects first introduced by Doi [31] and later reformulated by several authors [32–34].

A strong motivation for this approach comes from the evidence that many evolution processes lead to some sort of scale invariance, and as a consequence to the appearance of power law behavior.

The property of scale invariance has been quite successfully described and explained in the context of field theory by the methods based on the Renormalization Group (RG).

The RG approach was applied directly to the study of family name distribution in an article by De Luca and Rossi [35]. The starting point was a representation of the Galton-Watson branching process in a Hilbert space. Reproduction governed by chance is seen as a decay process described by a non-Hermitian Hamiltonian and by the corresponding evolution operator.

Again the result is completely independent of the initial conditions and of the offspring distribution. In both cases (presence and absence of mutations), one may check that the exponents are consistent with those obtained by the master equation approach and by previous authors.

Another approach was proposed in 2002 by Reed and Hughes [36,37]. They showed that, when interrupting (or observing) randomly a stochastic process characterized by exponential growth, the distribution of the observed state follows (at least asymptotically) a power law. Therefore, by considering the evolution of surname distribution as a Galton–Watson branching process and adding a finite probability for the appearance of new surnames (by mutation or immigration), it is possible to show that the distribution of family sizes follows a power law, with an exponent $2 + \beta/\delta$ depending on the probability β of appearance of new surnames by mutation and δ by the growth ratio of the population, but not dependent on the rate of immigration.

A disturbing aspect of this theoretical behavior (implying that the exponent should always be higher than 2) is the fact that the empirical values of the exponent are always lower than 2.

Bartley *et al.* analyzed the mechanisms leading asymptotically to a power law behavior [38]. They considered a model admitting birth and death and the creation of new kinds, and approximated it with a continuum equation for the distribution of surname frequencies $P(x,t)$, where the continuous variable x replaces the number k of individuals belonging to a family.

The resulting equation is of the Fokker–Planck type:

$$\frac{\partial P}{\partial t} = -k\frac{\partial(xP)}{\partial x} + \frac{1}{2}k_E\frac{\partial^2(xP)}{\partial x^2},$$

where $k \equiv \lambda - \mu$ and $k_E = \lambda + \mu$, λ and μ representing (constant) birth and death rates, and the mutation rate β appears in a boundary condition at $x = 1$.

The asymptotic limit for large family sizes shows a power law behavior, with an exponent depending on birth, death and mutation rates and going to 2 from above when the mutation rate goes to zero. However, the authors showed that the solution could be approximated by a power law also for smaller values of the family size, yet with exponents less than 2, thus bypassing the difficulty generated by empirical data.

4. The Effects of Sampling on Discrete Frequency Distributions

A subtle criticism to all the approaches refers to the absence of any discussion on the effects of sampling. In fact, there are good reasons to presume that the distribution parameters will in general depend on the absolute and relative size of the sample, and the deduction of quantitative properties referring to the whole population may not be straightforward [20,39].

We are considering a set of N objects belonging to S different kinds (“families”). A sample is a set of n objects, randomly extracted from the original set, and belonging to $S' \leq S$ different kinds.

A frequency distribution is a set of values $\{N_k\}$, where N_k is the number of kinds such that for each of them there are k objects in the original set. It is in principle possible to compute the probability of any sample distribution $\{n_l\}$ as a function of a given set $\{N_k\}$. A very important limit of the above result may be obtained when considering the (rather typical) case $k, l \ll N, n$. In this limit, setting $\rho \equiv n/N$,

$$\langle n_l \rangle = \sum_{k=1}^{N} N_k P_{kl}, \qquad P_{kl} \equiv \binom{k}{l} \rho^l (1-\rho)^{k-l}.$$

In the same limit we may derive a very important relationship between the generating function of the original frequency distribution and the generating function of the expectation values of its samples. Let us define:

$$G(z) \equiv \sum_{k=1}^{N} N_k (1 - \frac{z}{N})^k, \qquad g(z) \equiv \sum_{l=0}^{n} n_l (1 - \frac{z}{n})^l.$$

By replacing in $\langle g(z) \rangle$ the expression found for $\langle n_l \rangle$ and exchanging the order of summations, we easily obtain [20]

$$\langle g(z) \rangle = G(z).$$

By expanding $g(z)$ and $G(z))$ in powers of z, it is then possible to identify a set of moments whose expectation values are independent of the size of the sample, and therefore reflect very directly the properties of the original frequency distribution. The expression of the first nontrivial invariant moment (related to the standard isonymy parameter α_F) is:

$$\frac{1}{2\,n^2} \sum_{l=0}^{n} l(l-1) \langle n_l \rangle = \frac{1}{2\,N^2} \sum_{k=1}^{N} k(k-1) N_k \equiv \frac{1}{2\,\alpha_F}$$

.

Even without taking any special limit, it is possible to prove in full generality that the following equations hold for all $p \leq n$:

$$\langle m_p^{(n)} \rangle \equiv \langle \frac{(n-p)!}{n!} \sum_{l=p}^{n} n_l \binom{l}{p} \rangle = \frac{(N-p)!}{N!} \sum_{k=p}^{N} N_k \binom{k}{p} \equiv M_p$$

.

The expectation values of the above defined "moments" $m_p^{(n)}$ evaluated for samples of arbitrary size n coincide with the "moments" M_p of the original frequency distribution, as long as $p \leq n$.

One may also define more general sets of expected values, which can be proven to be independent of the sample size [20].

An important test of randomness in sampling is offered by the measure of the (standard) correlation C between different samples. Let us consider two random samples, characterized by their size n and m. The correlation between the two samples can be expressed in terms of the parameter α_F [20]:

$$\langle C \rangle = \frac{\frac{1}{\alpha_F} + \frac{1}{N}}{\sqrt{\frac{1}{\alpha_F} + \frac{1}{n}} \sqrt{\frac{1}{\alpha_F} + \frac{1}{m}}}$$

.

5. Evolution of Populations and the Dynamics of Disordered Systems

The attention of theoretical physicists towards the creation and the study of stochastic models appropriate to the description of some dynamical aspects of biological evolution dates back to the 1980s, when the neutralist theory of evolution, stating that the largest part of individual variability has no relevant effects on fitness, became quite popular. Such a theory lends itself very easily to representations typical of the systems studied in statistical mechanics. Models of neutral evolution belong quite naturally to the domain described by the theory of disordered systems, which was having notable developments in the same period.

The statistical properties of genealogical trees in a neutral model of a closed population with sexual reproduction, random mating and non-overlapping generations were quantitatively studied in 1999–2000, by theoretical and numerical methods, in the papers by Derrida, Manrubia and Zanette [21–23].

They measured the probability $H(r, G)$ for r repetitions of an ancestor appearing in the past generation G, assuming to be N the number of individuals in the same generation. After a sufficient number of generations, the distribution of the repetitions of ancestors takes a universal form, collapsing to a single curve in the plot of $P(w) \equiv 2^G H(r, G)/N$ as a function of $w \equiv rN/2^G$, and the left tail of $P(w)$, for small values of r, is a power law with a positive exponent $\beta \approx 0.3$. The fraction $\sigma(G)$ of the total population, which is expected to be absent from a given genealogical tree, can also be estimated.

The numerical results are confirmed by analytical calculations based on the assumption that the probability for an individual belonging to a given genealogical tree to have k children belonging to the tree becomes, for large N, a Poisson distribution. Then the generating function $h_G(z) \equiv \langle \exp[z\, w(G)] \rangle$ for the weights $w(G)$ satisfies a recursion equation having the form of a renormalization group transformation:

$$h_{G+1}(z) = \exp[m\, h_G(z/m) - m],$$

where m is the average number of descendants of a couple [23] .

In the large G limit the generating function converges to $h(z)$, the solution of

$$h(z) = \exp[m\, h(z/m) - m].$$

In the case of a fixed size population, corresponding to $m = 2$, the fraction of individuals with no descendants is $\sigma^* = h(-\infty)$ and therefore it solves $\sigma^* = \exp(2\sigma^* - 2)$; numerically, $\sigma^* = 0.203$. In this case, one may also extract the exponent β, finding $\beta = -\ln \sigma^* / \ln 2 = 0.299$, in excellent agreement with the results of the simulation [21].

A strictly related issue concerns the possibility of estimating the time to a common ancestor of all present-day individuals, the so-called most recent common ancestor (MRCA). The first statistical model was introduced and studied by Chang, who showed that, assuming complete randomness, the number of generations to the MRCA has a distribution peaked around $\ln(n)$, where n is the (constant) number of individuals in the population; moreover, at about $1.77 \ln(n)$ generations before the present, all individuals who have descendants are ancestors of all present-day individuals (identical ancestors point) [40]. Computer simulations of a model including substantial population substructure indicate that the MRCA may have lived a few thousand years ago and the identical ancestors point occurred just a few thousand years earlier [41].

It is worth recalling that the notion of MRCA has limited relevance for the genetics of a population with bisexual reproduction, since gene dilution implies that the genetic contribution of a single ancestor to an individual genome can be flushed out completely in roughly 1000 years.

6. Evaluating the Structure of Populations from Individual Genealogies

The ancestors of any individual (at any given generation in the past) may be viewed as a special sample of the social group (population) within which marriages may have occurred.

Let us briefly recall that an ancestors table includes in principle 2^{G-1} couples and 2^G individuals in the G-th generation, but in practice due to consanguinity, the number of different couples and individuals may be sensibly reduced, especially for large values of G, leading to important repetitions of individuals and surnames that may be described by appropriate frequency distributions, which become highly nontrivial when 2^G is comparable with (or larger than) the dimension of the community.

By considering many different ancestors tables for individuals belonging to the same generation and the same community, it is then possible to establish a statistical relationship linking the (average) frequency distribution of the ancestors (and more specifically the frequency distribution of their surnames) to the (statistical) properties of the community, in particular to its dimensions, and to the frequency distribution of the surnames appearing in the community itself. In this context, we call “probability" the ratio between the average frequency of a specific occurrence and the total number of possible occurrences.

In order to obtain closed form results, we need a number of simplifying assumptions that will not weaken substantially the value of our conclusions. We assume the Western pattern of surname propagation: all first generation descendants carry their father’s surname. We ignore generation mixing (even if in the long run it is certainly a relevant phenomenon), since it should not affect the pattern of surname distributions. We shall need the following definitions:

$m_G(k)$: probability of k repetitions of an individual in the G-th generation of ancestors (m_G is related to $H(r, G)$ and may be found by solving the recursive equation described in the previous section);

$M_G(k)$: probability that k (different) males belonging to the same family (*i.e.*, carrying the same surname) may appear in the G-th generation of ancestors;

$F_G(k)$: probability that k (different) females belonging to the same family may appear in the full population in the G-th generation ($F_G(k)$ should reflect the surname distribution of the population $P(k)$, defined in Section 2, at the time corresponding to G);

$R_G(k)$: probability that a surname may appear k times in the G-th generation (surname distribution of the genealogical tree, including repetitions, trivially coincident with the surname distribution of males in the (G+1)-th generation);

$D_G(k)$: probability that k females (including repetitions) belonging to the same family may appear in the G-th generation of ancestors (surname distribution of females in the G-th generation);

C_G: number of different surnames appearing in the G-th generation in the genealogical tree;

C^*: number of different surnames in the population;

S_G: number of different males (or couples) in the G-th generation in the genealogical tree;

N: number of different females in the full population (consisting of $2N$ individuals).

We may then establish the functional relationship existing between $R_{G-1}(k)$, $M_G(k)$ and $m_G(k)$ by noticing that under the reasonable assumption $2^G >> S_G > C_G >> 1$ (holding for a sufficiently large value of G), all probabilities may be treated as independent and therefore

$$R_{G-1}(y) = \sum_k \sum_{\{x_i\}} M_G(k) \prod_{i=1}^{k} m_G(x_i),$$

where $\{x_i\}$ are all the sets of k integers such that $\sum_{i=1}^{k} x_i = y$.

It is convenient to define generating functions for all probability distributions according to the general formula $\tilde{f}(z) = \sum_{k=1}^{\infty} f(k)\, z^k$. It is then straightforward to recognize that

$$\tilde{R}_{G-1}(z) = \tilde{M}_G\big(\tilde{m}_G(z)\big).$$

We may also establish the functional relationship between $D_G(k)$, $F_G(k)$ and $m_G(k)$ under the assumption that wives are chosen at random within the social group. To this purpose we need some expressions derived from the theory of sampling within a frequency distribution.

Let $P_{N,S}(r,k)$ be the probability that k (different) females belonging to the same family (characterized by the presence of r females in the full population) appear (by random mating) among the spouses of S different males present in the same generation in the genealogical tree:

$$P_{N,S}(r,k) = \frac{\binom{N-r}{S-k}\binom{r}{k}}{\binom{N}{S}}.$$

Under the previously specified assumptions, since we know that males appear in the genealogical tree with repetitions described by the function $m(k)$ and females of different families are distributed according to $P_{N,S}(r,k)$, we can now compute the probability $\Pi_{N,S}(r,y)$ of y individual repetitions in the genealogical tree for females belonging to a family characterized by a female frequency r in the full population:

$$\Pi_{N,S_G}(r,y) = \sum_k \sum_{\{x_i\}} P_{N,S_G}(r,k) \prod_{i=1}^{k} m_G(x_i).$$

Hence we obtain

$$\tilde{\Pi}_{N,S_G}(r,z) = \tilde{P}_{N,S_G}\big(r, \tilde{m}_G(z)\big).$$

It is fully reasonable to consider the limit $k, r << S, N$, in which case we can easily find that

$$P_{N,S} \to \left(\frac{S}{N}\right)^k \left(1 - \frac{S}{N}\right)^{r-k} \binom{r}{k},$$

and as a consequence

$$\tilde{\Pi}_{N,S_G}(r,z) \to \left(1 - \frac{S_G}{N} + \frac{S_G}{N}\tilde{m}_G(z)\right)^r.$$

By repeating the previous arguments, we can now compute the probability distribution for the frequency of female surnames in the genealogical tree:

$$D_G(y) = \sum_r F_G(r) \Pi_{N,S_G}(r,y),$$

implying

$$\tilde{D}_G(z) = \sum_r F_G(r)\tilde{\Pi}_{N,S_G}(r,z) = \tilde{F}\Big(1 - \frac{S_G}{N} + \frac{S_G}{N}\tilde{m}_G(z)\Big).$$

Collecting all results and keeping in mind that the full distribution is the convolution of the male and female distributions, we then obtain the desired relationship:

$$\tilde{R}_G(z) = \tilde{D}_G(z)\tilde{R}_{G-1}(z),$$

implying the master equation

$$\tilde{F}\Big(1 - \frac{S_G}{N} + \frac{S_G}{N}\tilde{m}_G(z)\Big) = \frac{\tilde{R}_G(z)}{\tilde{R}_{G-1}(z)} = \frac{\tilde{M}_{G+1}\big(\tilde{m}_{G+1}(z)\big)}{\tilde{M}_G\big(\tilde{m}_G(z)\big)}.$$

Hence, the (expected) surname distribution of the population at time G in the past can be extracted from the comparison of two subsequent surname distributions in the available genealogical trees.

A number of trivial properties follow, concerning the first derivatives of the generating functions:

$$\tilde{m}'_G(1) = \frac{2^{G-1}}{S_G}, \qquad \tilde{M}'_G(1) = \frac{S_G}{C_{G-1}}, \qquad \tilde{F}'_G(1) = \frac{N}{C^*}, \qquad \tilde{R}'_G(1) = \frac{2^G}{C_G}, \qquad \tilde{D}'_G(1) = \frac{2^{G-1}}{C^*}.$$

Moreover, as a really nontrivial consequence of the master equation, we obtain the prediction:

$$C^* = \frac{C_G\, C_{G+1}}{2\, C_G - C_{G+1}}.$$

7. Empirical Studies of Ancestors Tables

European nobility (and especially German higher nobility, or *Hochadel*) is characterized by the existence of records of kinship that go back to the middle ages and involve a significant number of different families. We have explored these records and obtained some interesting results. Work is in progress, and we list here only a short resumé of the results: a complete presentation will be given in a future publication.

(1) The ancestors tables for about 100 individuals (almost all relevant subjects in the Western European high nobility, obviously excluding brothers/sisters) living around the year 1800 have been reconstructed up to the eleventh generation, with a limited number of missing entries.

In principle about 200,000 ancestors might be involved, but due to repetitions and consanguinity the number of independent individuals is less than 27,000 (among them only 11,000 fully identified)

We obtained the (normalized) frequency distribution of ancestors, both cumulative and by generation, ignoring generation mixing, which is significant but should not affect our analysis.

We found evidence of universality, but still no onset of the specific scaling predicted by Derrida *et al.*, which is expected to appear after $O(\ln N)$ generations, where N is the dimension of the population.

We identified the MRCA of our 100 individuals (and of their relatives), in the person of Wilhelm I, Graf von Nassau-Dillenburg (1487–1559), living about 300 years before the group under scrutiny.

It is also possible to study the family correlations and build a taxonomy of German noble families.

(2) The ancestors tables have been reconstructed for Henri de Bourbon-Orléans, comte de Paris (1908–1999) and for Otto von Habsburg-Lothringen (1912–2011), involving 16 almost complete generations and essentially all European nobility back to the year 1400.

We found the first hints of Derrida scaling, and after fixing the parameters by best fit, we found some evidence of a decreasing population starting from less than 1000 individuals in the year 1400, thus confirming qualitative observations by historians about the marriage policies of the European nobility.

(3) The ancestors tables for 48 individuals living around the year 1900 (including almost all the European sovereigns, and again excluding brothers/sisters) have been reconstructed up to the twelfth generation, with about 5% missing entries.

We found that more than 70% of the effective ancestors in the oldest generation (including replications) belong to a very restricted number of families (about 60), some of which presently extinct, thus confirming on a European scale the long term social (and genetic) closeness of higher nobility. The MRCA (actually more than one in the same time lapse) have been found to live in the first half of the 1600s, three centuries before the above defined group of their descendants.

We studied the surname distribution and found some kind of universality, but the required scaling is nontrivial and there is no evidence of a power law behavior. More work is certainly needed before the implications of the data analysis are fully understood, and a real comparison with the theory presented here has not yet been performed.

Conflicts of Interest

The author declares no conflict of interest.

References

1. Darwin, G.H. Marriages between first cousins in England and their effects. *J. Stat. Soc.* **1875**, *38*, 153–184.
2. Galton, F.; Watson, H.W. On the Probability of the Extinction of Families. *J. Anthropol. Inst. Great Brit. Ireland* **1874**, *4*, 138–144.
3. Crow, J.F.; Mange, A.P. Measurement of inbreeding from the frequency of marriages between persons of the same surname. *Eugenics Quarterly* **1965**, *12*, 199–203.
4. Karlin, S.; McGregor, J. The number of mutant forms maintained in a population. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*; University of California Press: Oakland, CA, USA, 1967; Volume 4, pp. 415–438.
5. Fisher, R.A.; Corbet, A.S.; Williams, C.B. The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. *J. Anim. Ecol.* **1943**, *12*, 42–58.
6. Lasker, G.W. A coefficient of relationship by isonymy: A Method for Estimating the Genetic Relationship between Populations. *Hum. Biol.* **1977**, *49*, 489–493.
7. Fox, W.R.; Lasker, G.W. The Distribution of Surname Frequencies. *Int. Stat. Rev.* **1983**, *51*, 81–87.
8. Gottlieb, K., Ed. Surnames as markers of inbreeding and migration. *Hum. Bio.* **1983**, *55*, 209–408.
9. Lasker, G.W. *Surnames and Genetic Structure*; Cambridge University Press: Cambridge, UK, 1985.

10. Boattini, A.: Lisa, A.; Fiorani, O.; Zei, G.; Pettener, D.; Manni, F. General method to unravel ancient population structures through surnames, final validation on Italian data. *Hum. Bio.* **2012**, *84*, 235–270.
11. Redmonds, G.; King, T.; Hey, D. *Surnames, DNA, and Family History*; Oxford University Press: Oxford, UK, 2011.
12. Yule, G.U. A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis. *Phil. Trans. R. Soc. Lond. B Biol. Sci.* **1925**, *213*, 21–87.
13. Simon, H.A. On a class of skew distribution functions. *Biometrika* **1955**, *42*, 425–440.
14. Newman, M.E.J. Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.* **2005**, *46*, 323–351.
15. Derrida, B.; Peliti, L. Evolution in a flat fitness landscape. *(Bull. Math. Biol.* **1991**, *53*, 355–382.
16. Derrida, B.; Bessis, D. Statistical properties of valleys in the annealed random map model. *J. Phys. A* **1988**, *21*, L509–L515.
17. Serva, M.; Peliti, L. A statistical model of an evolving population with sexual reproduction. *J. Phys. A* **1991**, *24*, L705–L709.
18. Rossi, P. Surname distribution in population genetics and in statistical physics. *Phys. Life Rev.* **2013**, *10*, 395–415.
19. Baek, S.K.; Kiet, H.A.T.; Kim, B.J. Family name distributions: Master equation approach. *Phys. Rev. E* **2007**, *76*, 046113:1-046113:7.
20. Rossi, P. Invariant expectation values in the sampling of discrete frequency distributions. *Physica A* **2014**, *394*, 177–186.
21. Derrida, B.; Manrubia, S.C.; Zanette, D.H. Statistical Properties of Genealogical Trees. *Phys. Rev. Lett.* **1999**, *82*, 1987–1990.
22. Derrida, B.; Manrubia, S.C.; Zanette, D.H. Distribution of repetitions of ancestors in genealogical trees. *Physica A* **2000**, *281*, 1–16.
23. Derrida, B.; Manrubia, S.C.; Zanette, D.H. On the genealogy of a population of biparental individuals. *J.Theor. Bio.* **2000**, *203*, 303–315.
24. Kim, B.J.; Park, S.M. Distribution of Korean Family Names. *Physica A* **2005**, *347*, 683–694 .
25. Zanette, D.H.; Manrubia, S.C. Vertical transmission of culture and distribution of family names. *Physica A* **2001**, *295*, 1–8.
26. Manrubia, S.C.; Zanette, D.H. At the Boundary between Biological and Cultural Evolution: the Origin of Surname Distributions. *J.Theor. Bio.* **2002**, *216*, 461-477.
27. Bak, P.; Tang, C.; Wiesenfeld, K. Self-organized criticality: An explanation of the 1/f noise. *Phys. Rev. Lett.* **1987**, *59* , 381–384.
28. Bak, P.; Sneppen, K. Punctuated equilibrium and criticality in a simple model of evolution. *Phys. Rev. Let.* **1993**, *71*, 4083–4086.
29. Flyvbjerg, H.; Bak, P.; Sneppen, K. Mean field theory for a simple model of evolution. *Phys. Rev. Lett.* **1993**, *71*, 4087–4090.
30. de Boer, J.; Derrida, B.; Flyvbjerg, H.; Jackson, A.D.; Wettig, T. Simple Model of Self-Organized Biological Evolution. *Phys. Rev. Lett.* **1994**, *73*, 906–909.

31. Doi, M. Second quantization representation for classical many-particle system. *J. Phys. A* **1976**, *9*, 1465–1478.
32. Goldenfeld, N. Kinetics of a model for nucleation-controlled polymer crystal growth. *J. Phys. A* **1984**, *17*, 2807–2821.
33. Peliti, L. Path integral approach to birth-death processes on a lattice. *J. De Phys.* **1985**, *46*, 1469–1483.
34. Jarvis, P.D.; Bashford, J.D.; Sumner, J.G. Path integral formulation and Feynman rules for phylogenetic branching models. *J. Phys. A* **2005**, *38*, 9621–9647.
35. De Luca, A.; Rossi, P. Renormalization group evaluation of exponents in family name distributions. *Physica A* **2009**, *388*, 3609–3614.
36. Reed, W.J.; Hughes, B.D. From gene families to incomes and internet file sizes: Why power laws are so common in nature. *Phys. Rev. E* **2002**, *66*, 067103:1–067103:4.
37. Reed, W.J.; Hughes, B.D. On the distribution of family names. *Physica A* **2003**, *319*, 579–590.
38. Bartley, D.L.; Ogden, T.; Song, R. Frequency distributions from birth, death and creation processes. *BioSystems* **2002**, *66*, 179–191.
39. Maruvka, Y.E.; Shnerb, N.M.; Kessler, D.A. Universal features of surname distribution in a subsample of a growing population. *J. Theor. Bio.* **2010**, *262*, 245–256 .
40. Chang, J.T. Recent common ancestors of all present-day individuals. *Adv. App. Prob.* **1999**, *31*, 1002–1026.
41. Rohde, D.L.T.; Olson, S.; Chang, J.T. Modelling the recent common ancestry of all living humans. *Nature* **2004**, *431*, 562–566.