Surname distribution in population genetics and in statistical physics

Paolo Rossi

Dipartimento di Fisica dell'Università di Pisa and I.N.F.N., Sezione di Pisa,
Largo Bruno Pontecorvo 3, I-56127 Pisa, Italy

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

Surname distribution in population genetics and in statistical physics

Paolo Rossi

Dipartimento di Fisica dell'Università di Pisa and I.N.F.N., Sezione di Pisa,
Largo Bruno Pontecorvo 3, I-56127 Pisa, Italy

Abstract

Surnames tend to behave like neutral genes, and their distribution has attracted a growing attention from genetists and physicists. We review the century-long history of surname studies and discuss the most recent developments. Isonymy has been regarded as a tool for the measurement of consanguinity of individuals and populations and and has been applied to the analysis of migrations. The analogy between patrinileal surname transmission and the propagation of Y chromosomes has been exploited for the genetic characterization of families, communities and control groups. Surname distribution is the result of a stochastic dynamics, which has been studied either as a Yule process or as a branching phenomenon: both approaches predict the asymptotic power-law behavior which has been observed in many empirical researches. Models of neutral evolution based on the theory of disordered systems have suggested the application of field-theoretical techniques, and in particular the Renormalization Group, to describe the dynamics leading to scale-invariant distributions and to compute the related (critical) exponents.

Keywords: Surname distribution; Isonymy; Consanguinity; Population; Yule process; Branching process; Power law; Renormalization group; Y chromosome

1. Introduction

The frequency distribution of family names has been an interesting issue in human biology since the last quarter of the 19th 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 transmission rules of surnames are the same as those of the Y chromosome, that in preserved in the male descendants and is only affected by mutation (and by false paternity).

The very first statistical studies on surname frequency go back to George Darwin's analysis of marriage isonymy in England as a tool for the evaluation of inbreeding in the English society and to Galton and Watson's evaluation of the probability of surname extinction, which was the starting point for the mathematical study of branching processes.

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. Only in the Sixties of the 20th century a substantial revival occurred, triggered by a number of new and important results. In 1965 Crow and Mange proposed a model for consanguinity estimates based on marital isonymy, 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 which led genetists to a systematic study of surname distribution and extinction as an empirical model for neutral gene propagation. A further relevant development was started in the Eighties by the first studies linking surname distributions with migration rates of populations.

Among the fathers of the modern study of surnames in the context of human biology one must certainly mention the anthropologist G.W. Lasker, who extended the use of isonymy to the study of consanguinity between populations and observed (with W.R. Fox) that the empirical frequency distribution of surnames could be accurately described by means of a discrete Pareto (power law) distribution.

In the last 25 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, creating the conditions for a more systematic phenomenological approach to the issue.

Different parametrizations of surname frequency distributions, based on statistical models of reproductive behavior and population dynamics, have since then been proposed. Especially relevant appears to be the observation that surname dynamics can be seen as a Yule process, thus leading to a natural statistical explanation for the appearance of scaling in a proper limit.

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, often relating it to the concept of self-organized criticality. The presence of scaling in surname distributions did therefore attract the attention of scholars in the field of statistical physics.

Starting from the late Eighties theoretical physicists 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 exploiting some developments in the theory of disordered systems. Evolution of populations in flat fitness landscapes was studied, both in asexual and in sexual reproduction. Statistical properties of genealogical trees were explored from a theoretical point of view. In the mean time the phenomenological study of surname distributions, performed in Western countries especially by biologists, in Far Eastern countries (Japan, Korea, China) became the object of research activity by a small but significant community of physicists.

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.

Probably 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. This model encompasses many previous models and leads to many testable predictions, which have been verified by the authors in the case of Far Eastern countries, crossing phenomenological data with information coming also from historical and demographic records.

The behaviours predicted by solving the master equation can be independently reproduced by applying renormalization group techniques in order to describe the evolution dynamics of surnames. This result brings the issue of surname dynamics back into the context of studies concerning disordered systems, a field whose constant progress leads to the expectation that better and better analytical and numerical tools will keep being developed, allowing for a deeper understanding of the role played by different factors not only in surname evolution but also in other important evolutionary (and cultural) processes admitting an analogous description and representation.

2. The early history of surname studies in human biology (to 1985)

The first statistical studies on surname frequency go back to 1875 and to George Darwin's analysis of marriage isonymy in England as a tool for the evaluation of inbreeding in the English society [1]; his father Charles (the founder of modern evolutionary theory) had married a first cousin, and George was interested in the possible genetic effects of consanguinity in the families and in the population. He supposed that the number of marriages between persons of the same surname who were not first cousins should be proportional to the frequency of the surname in the population, and therefore frequent only for common surnames. Since the Registrar General had published in 1853 the frequency of the fifty most common surnames in England, Darwin could compute the sum of the squares of these frequencies (0.00092) and estimated that marriages between unrelated people of the same surname should not be more than one part per thousand: an exceeding percentage should therefore be ascribed to marriages between first cousins.

Taking into account the fact that same-name marriages are just a fraction of the total number of marriages between cousins, Darwin concluded that the rate of first cousin marriages was about 4.5 % among the aristocracy, 3.5 % among the middle classes and landed gentry, 2.25 % in the rural population and 2 % in the urban population.

A similar analysis was performed in 1908, more than thirty years later, by G.B.L. Arner, who published a study on cousin marriages in two American populations [2]. Taking into account more than ten thousand marriage licences in New York dated before 1784, he found that same-name marriages were about 2 %, against an expected ratio of 0.076 % based on the frequency of the 50 most common surnames; hence according to Darwin's method the estimate of cousin marriages in colonial New York should have been around 5,9 %. Considering this value to be too high, Arner studied a number of genealogies and came to the conclusion that only 2.76 % of marriages were between first cousins. In Ashtabula County, Ohio, considering about 134,300 marriages between 1811 and 1886 he concluded, by Darwin's method, that 1,12 % of the marriages were between first cousins.

The obvious bias of Darwin's and Arner's studies was the choice of considering only the marriage between first cousins and not taking into account other possible degrees of relationship; this led to a long term dismission of surname analysis as a tool for the study of consanguinity.

In 1954 M. Kamizaki wrote an article (in Japanese), where he calculated the expected frequency of isonymous marriages for various degrees of relationship [3], anticipating results later derived by Crow, but his paper went completely unnoticed in the West. In 1960 R.F. Shaw underlined the opportunity offered by the Spanish surname system, attributing two surnames to each individual (the first surnames of the paternal and maternal grandfathers): the frequency of identical surnames would allow a faster statistical measure of consanguinity [4]. The very early history of the use of surnames for the study of human inbreeding was traced in 1967 by Yasuda and Morton [5] .

However the research on the relationship between isonymy in marriage and inbreeding was really revived only around 1964 by J.F. Crow, who did not refer to previous results but recalled a lecture in the 1940s by the Nobel Prize H.J. Muller, who suggested that surnames could be used in genetic models of inbreeding.

Crow noticed that in many relationships the ratio between the inbreeding coefficient (as defined by Wright [6]) for the offspring of a related couple and the chance that the two parents share the same surname is constant and equal to 1/4. There are also obvious counterexamples (first noticed by C. Cotterman) that can be found in marriages across generations, when one partner is a direct descendant of the other, or in some types of cumulative inbreeding. Nevertheless Crow observed that the cases satisfying the 1/4 ratio encompass most realistic instances in human mating.

Therefore, in collaboration with A. Mange (who had assembled records of the marriages of the Hutterites) in 1965 he wrote a fundamental paper on the subject of surname models of human inbreeding [7, 8]. Crow and Mange defined the total inbreeding (F_t) of a population and its random (F_r) and nonrandom (F_n) components, of which only the first can be explained by the random mating of individuals present in a population. So they introduced the restrictive (and historically unrealistic) assumptions that all the surnames are monophyletic (i.e. that sharing the last name means sharing the ancestor from which it derives), and that both sexes are equally represented in the migrant groups. Under these assumptions they observed that, for a large number of types of relationship, the isonymy in marriages I (marital isonymy) is indicative of the degree of the inbreeding in the population regardless of the degree of consanguinity in individual marriages, because the probability that two descendants from a common ancestor have the same last name varies, in most cases, to an extent proportional to the degree of consanguinity.

To be precise, the probability P that two individuals have the same last name is obviously linked to the degree of kinship (and it is one of the brethren, 1/2 for the uncle-nephew relationship, 1/4 for the cousins, etc.), but also the degree of inbreeding F among their children depends on the degree of kinship (and it is 1/4 of the children of brothers, 1/16 for the children of cousins etc), and it is quite easy to see that typically, the relation F = P/4, with rare exceptions related to complex and crossed family relationships whose statistical significance is however very poor. Consequently, for the more remote kinship the actual number of consanguineous marriages is higher than indicated by marital isonymy, and the ratio grows in inverse proportion to P, but in return, the consanguinity between the descendants of marriages between relatives is reduced in proportion to F, and the two effects offset each other in virtue of the relation $F/P = 1/4$.

Therefore the percentage of isonymic marriages is a direct measure of the degree of consanguinity F_t of the population as a whole, equal to 1/4 of marital isonymy: $F_t = 1/4\, I$.

To separate the random component of inbreeding from the non-random one, Crow and Mange noted that, calling $p(i)$ and $q(i)$ the fractions of the male and female population identified bythe i-th name, the frequency of marriages between random pairs with the same surname is $p(i)q(i)$ and the coefficient of random inbreeding F_r is then, as previously demonstrated, $F_r = 1/4 \sum_i p(i)q(i)$, while the coefficient of non random consanguinity F_n can easily be obtained by the relation $F_t = F_n + (1 - F_n)F_r$, where in turn F_t is measured by isonymy in marriages.

Theoretical developments came in 1971 when Holgate showed that isonymy tends to grow with time in models of monogamous populations with random mating [9], while Cavalli-Sforza and Bodmer suggested elimination of brother-sister pairs from estimates of random inbreeding [10].

Another relevant contribution came in 1977 from Lasker's article [11]: Lasker, noting that the coefficient of relationship for a couple of parents is exactly twice the inbreeding coefficient of their children, proposed an extension of the formula of Crow and Mange that would allow to use isonymy as a measure of the degree of relatedness between any two populations. The relationship measured by isonymy is $RI = 1/2 \sum_i p_1(i)p_2(i)$, where $p_1(i)$ and $p_2(i)$ are the frequencies with which the i-th surname appears in the two populations.

The renewed attention on the possible use of surnames as a tool in human genetics studies prompted empirical research, for instance in Switzerland by Morton *et al.* [12, 13, 14, 15] and by Ellis *et al.* [16, 17, 18], and in the Pyrenees by Bourgoin and Vu Tien Khang [19]. As a matter of fact, a poor correspondence was usually found between estimates of inbreeding from surnames and from pedigrees, which was partly explained by the inclusion of remote inbreeding granted only by the surname method, but should also be ascribed to the polyphyletic (i.e. independent) origin of many surnames and to the effects of migrations.

In March and April of 1982 a conference and two symposia, whose subject was surnames as biological markers of inbreeding and migration, were held in Eugene, Oregon. As many as 18 communications, grouped into three major areas (Surnames as population markers, Marital isonymy as a measure of population structure and Communality of surnames between groups as a measure of population structure), were later published in a single issue of the journal *Human Biology* [20]. Crow's final conclusions [21] and the large bibliography represent the state-of-the-art of the subject in 1983.

Shortly later Lasker and Kaplan [22] confirmed Devor's suggestion [23] that more insight on the genetic structure of a population might be obtained by the study of the full matrix of wife vs husband surnames, instead of just focusing on its diagonal values (corresponding to marital isonymy).

A wide collection of results was presented also in Lasker's review article [24] and in his book, published in 1985 [25].

3. The frequency distribution of surnames (I)

An independent line of research on surname frequency was originated at the end of the 19th century by Galton and Watson's study on the extinction probability of family names [26]. The study was triggered by the worried observation that several families belonging to English aristocracy were extinct or on the verge of extinction. The initial (and wrong) conclusion was that, after a sufficiently long time span, the extinction probability for any family name should in all cases be equal to one: actually this result holds only under the assumption that the average number of male descendants be less or equal to one, while in the opposite case the probability is not marginal, but it is definitely less than one. However Galton and Watson's paper gave way to the still flourishing mathematical study of branching processes [27, 28]. In the context of population theory the approach was revived first by Lotka in 1931 [29].

Nevertheless only the appearance in 1967 of an article by Karlin and McGregor [30], who formulated the theory of the behavior of neutral mutations in finite populations of constant size, in conjunction with the observation that surnames can be considered as alleles transmitted along the male line, attracted the attention of genetists towards the study of the distribution and extinction of surnames. Karlin and McGregor, starting from a model by Kimura and Crow [31], described the growth of different mutant gene types, assuming that r different mutant types may exist in a population consisting of N individuals, where r is much larger than N, and denoted by β the probability that an individual of a given type change into a specified different type.

They evaluated the expected number $N^*(k)$ of alleles represented k times in the population once an equilibrium state is attained. In the limit when $r \to \infty$ while the overall mutation probability (per generation) $\nu = r\beta$ is kept finite, they obtained

$$N^*(k) = \frac{1}{k}\frac{N\nu}{1-\nu}\frac{\binom{\frac{N}{1-\nu}k-1}{N-k}}{\binom{\frac{N}{1-\nu}-1}{N}}.$$

The expected total number N^* of alleles represented in the population is therefore

$$N^* = \frac{N\nu}{1-\nu}\Big(\frac{\Gamma'(\frac{N}{1-\nu})}{\Gamma(\frac{N}{1-\nu})} - \frac{\Gamma'(\frac{N\nu}{1-\nu})}{\Gamma(\frac{N\nu}{1-\nu})}\Big),$$

where Γ stands for the usual Gamma function. It is possible to compute the probability of homozygosity F, that is the probability that two randomly chosen genes are of the same type: the result is $F = 1/(N\nu+1-\nu)$ and it is possible to show that $N^*F > 1$.

For most practical purposes it is convenient to consider the large N limit of the above formulae, obtaining in this case Fisher's distribution [32] $N^*(k) = \alpha_F x^k/k$, where it was convenient to introduce the new variables $x = 1-\nu$ and $\alpha_F = N\nu(1-\nu)$, and in the same limit one obtains $N^* = \alpha_F \ln\frac{1}{\nu}$.

The theory of Karlin and McGregor was applied in 1974 by Yasuda *et al.* [33] to the surname data for the population of the Parma valley collected by Cavalli-Sforza and collaborators since 1954 [34]. In their language, $N^*(k)$ is the number of surnames represented by k males in a population of N male individuals, ν is the probability of a surname change in a given generation, $S \equiv N^*$ is the total number of surnames, and the expected value of the random isonymy is $I \equiv F$.

The comparison between the theoretical and the empirical distribution allows to verify the correctness of assumptions and to set the values of the parameters N and ν. Since real surname mutations are relatively rare, a high value of ν (such as that found in the case of the valley of Parma, greater than 0.2) seems to indicate the importance of the phenomenon of immigration in the distribution of surnames.

In the same article Yasuda and his coworkers confronted the issue of the statistical distribution of the offspring, focusing in particular on the extinction of surnames after a sufficiently high number of generations. The probability of extinction they calculated (based on a geometric model of decrease for the probability of birth of n sons) resulted in reasonable agreement with the available empirical data.

A key finding in the study of the frequency distribution of surnames was published by Fox and Lasker in 1983 [35] (but it had been already announced in the article by Lasker in 1980 [24]). Fox and Lasker showed that the empirical data relating to the frequency distribution of surnames of 4794 people living in the area of Reading (England) could be described with good accuracy by using a discrete Pareto distribution.

Another important contribution to the investigation on the frequency distribution of surnames resulted from a series of works published since 1983 by Zei *et al.* and specifically dedicated to the study of the distribution of surnames in Sardinia [36, 37, 38]. It was shown that the empirical distribution showed a good agreement with the distribution of neutral alleles of Karlin-McGregor, and also with the logarithmic distribution of Fisher, seen as a limit of the previous one. The spatial dependence of the distribution of surnames was also taken into consideration, noting that the spatial variation was much stronger than the temporal variation, and largely independent of the frequency of names (except in the case of very low frequency). These results are all consistent with the hypothesis that the surnames behave as neutral alleles.

Other representations of the frequency distribution of surnames, based on statistical models of the reproductive behavior and of the evolution of surnames, were proposed by Panaretos in 1989 [39, 40] and by Consul in 1991 [41, 42]. Assuming a Poisson distribution for the number of males with a given surname from a pool of size λ, an exponential distribution for the degree of commonality of names and a probability density function for the distribution of pool sizes proportional to $(1+\lambda)^{-(c+1)}$, with $c > 0$, Panaretos showed that the surname frequency distribution would be the zero-truncated Yule distribution

$$P(k) = \frac{c(k-1)!\,\Gamma(c+1)}{\Gamma(k+c+1)}, \qquad k > 0.$$

For large enough k this function would behave like the discrete Pareto distribution $k^{-(c+1)}$, thus providing the rationale for Fox and Lasker's results.

In order to offer further motivation to his result, Panaretos considered also a stochastic model originally proposed by Simon [43] to explain the occurrence of Zipf's law in linguistics [44]. Assuming the existence of a nonzero probability α for the appearance of a new surname whenever a unit is added to the population, and adopting the (simplifying) hypothesis that surname frequencies increase proportionally to the total number of surnames, the probabilities for the surname frequencies satisfy the recursive relationship

$$\frac{P(k)}{P(k-1)} = \frac{(1-\alpha)(k-1)}{1+(1-\alpha)k},$$

whose solution is again the Yule function with the identification $c = 1/(1-\alpha)$.

The aim of Consul's approach was to provide a derivation of the distribution function based on a dynamical model of surname evolution. To this purpose he considered a birth-death model and, in alternative, a branching process starting from a first generation described by a negative binomial distribution of surnames. In both cases he obtained a Geeta distribution characterized by two independent parameters, $\beta > 1$ and x:

$$P(k) = \frac{x^{k-1}(1-x)^{\beta k-k}}{k!}\frac{\Gamma(\beta k-1)}{\Gamma(\beta k-k)},$$

and the fit to the data turned out to be better than the one based on the Yule distribution. However no direct connection appeared between the parameters of the distribution and the actual dynamics of the process.

The ubiquity of power laws in the description of phenomena (both natural and cultural), whenever the phenomenon is not characterized by an intrinsic scale, may be explained in the context of the modern theory of complexity and is the reason which has led in recent years scholars of statistical physics to investigate also the distribution of surnames, among a number of other subjects lying outside the traditional field of their discipline. Since we are trying to follow, at least partially, the historical development of research on surnames, we shall come back only in Sections 7-10 to the more recent approaches aimed at identifying the dynamical origin of surname distributions and predicting on theoretical grounds their possible shape.

4. Surnames and migratory phenomena

The development of surname studies in the late Eighties and in the Nineties was marked by a special interest on the potential use of data referring to the spatial distribution of surnames in order to trace the migratory dynamics of some human populations. The first significant contribution to this issue came from Chen and Cavalli-Sforza who showed that an unexpected correlation between surnames in the opposite sides of Taiwan could be explained on historical grounds because both sites had been colonized by immigrants coming from the same area of the mainland [45].

In 1984 the seminal paper by Wijsman *et al.* [46] included the computation of migration matrices for nine different geographic areas of Sardinia, based on a kinship matrix defined as a generalization of Lasker's formula for the relationship between populations. The matrices were evaluated with reference to different years, in the time span going from 1850 to 1970. A strong and non linear dependence of immigration coefficients on the distance between the areas was found.

The use of surnames for the study of migratory dynamics was further developed in subsequent papers. In particular Piazza *et al.* in 1987 compared their estimates of migration coefficients in Italy based on the analysis of surname distributions extracted from phone directories with the corresponding estimates obtained from official demographic data sources, showing that an evaluation of the mutation parameter ν (as defined by Karlin and McGregor) could offer a plausible estimate of the migration coefficients [47, 48].

The study of migrations by the analysis of surname distributions was later performed also by Darlu and Ruffié for French internal fluxes [49, 50, 51], and by Degioanni *et al.* for migrations from Italy to France [52].

In turn Zei *et al.* [53] explored Italian migration dynamics, analyzing the geographical distribution of surnames, especially in the presence of strong regional differences, confronting their results with the existing geographical and linguistic barriers and comparing their data to those resulting from the genetic analysis of the population. Genetic and surname frontiers turned out to be significantly correlated with each other and especially with physical (but also with language) barriers.

The relationship between genetic (vertical) and cultural (horizontal) transmission of information was investigated first by Guglielmino *et al.* who analyzed and compared the surname distribution in different areas of Sicily [54, 55]. The complex anthropological history of Sicily attracted also the attention of other research groups [56, 57, 58]. Significant early contributions to the analysis of populations by their surname structure came also by Lucchetti and collaborators [59, 60].

5. Surnames and the Y chromosome

In many cultures, and in particular in European cultures, the hereditary transmission of the surname is patrilineal, in perfect analogy with the behavior of the Y chromosome, possessed only by males and transmitted without recombination from father to sons. Therefore one might expect a direct link to exist between the two,

The idea of comparing surnames to genetic markers was first put forward in the Thirties, but the first evidence that this link could be made in practice emerged in 1972 by the study of a French-Canadian family with a ten generation pedigree. The peculiarity of the case resided in the presence of a "satellited Y chromosome, which could be detected simply by looking at the chromosomes down a microscope. Only with the advent of the PCR method in 1985 and the subsequent discovery of polymorphic DNA markers it became possible to examine differences between Y chromosomes at the DNA level.

In 1997 Jobling suggested that it might be possible to use a Y chromosome haplotype to predict a surname. A small set of studies followed soon, and a strong correlation was usually found, indicating a strict link between genetic and cultural data.

In the case of of Cohanim Jews whose surname (Cohen and its variants) should indicate, according to Jewish tradition, a direct patrilineal descent from Aaron, brother of Moses (XIII century b.C), the analysis of the Y chromosome showed the presence of a common modal haplotype with a frequency above 60 % be compared with values below 15 % in the control group [61, 62].

Another interesting result concerned the relationship between the haplotypes of a group of 221 Irish males and their surnames, grouped according to the (historically known) geographic origin of the Gaelic surnames or to the foreign (Scottish, Norse, English) source of the others [63] .

A description of the first researches on the Y chromosome and on their relationship with the study of the origin and distribution of surnames can be found in a review paper written by Jobling in 2001 [64].

This line of research on the origin and biological evolution of surnames was highlighted in the year 2000 by Sykes and Irven [65]. They analyzed the Y-chromosome haplotypes of a large random sample of male subjects carrying the surname Sykes. Almost 50% of the sample shared a peculiar haplotype, not observed in the control groups. Sykes and Irven inferred that, by admitting a very limited number of nonpaternity cases (on average 1.3 % per generation) during the 700 years of documented history, the surname Sykes should be considered monophyletic, against some evidence of poliphyletism coming from written sources.

Later research was devoted to the correlation between a few specific Y chromosome haplotypes and the surname distribution in the different geographic and cultural areas of Sardinia [66] .

The method for the analysis of the spatial distribution of surnames was further refined in a paper by Manni *et al.* [67], in which the SOM (self-organizing maps) grouping technique was applied to Dutch surnames.

The authors suggested that the Y-chromosome sampling of an area exposed to migration phenomena should be performed after a preliminary selection based on surnames, when the purpose of research is to find out the genetic traits that were typical of the population living in the area before the most recent migration waves.

A few recent references concern the relationship between surnames, genetic and historical data in the British Isles [68, 69, 70, 71]. In particular, by comparing the Y chromosomes of randomly chosen couples of male individuals sharing the same surname, and assuming a maximum acceptable number of mutations consistent with the hypothesis that surnames became hereditary in Britain around 1300, it was found that only the rarer surnames (with less than 5000 bearers) had a serious chance of indicating a common ancestor, while most of the commonest family names are very probably polyphyletic. A review of recent results is presented in the 2009 article by King and Jobling [72].

A very detailed discussion of the theoretical aspects and experimental results related to the connection between surnames and Y chromosome types is offered in Chapters 7 and 8 of a recent volume by Redmonds, King and Hey [73].

6. Collection and analysis of empirical data

Empirical data on surname distribution were systematically collected and analyzed by several groups of researchers. Instead of quoting a very long list of references, we found it convenient to recall here the main review articles that have been devoted to the subject in the last decade, specifying their aim and their peculiarities.

The 2003 article by Colantonio *et al.* [74] included a very large reference list, organized by geographical areas, while the 2004 paper by Darlu [75] had the explicit and specific purpose of communicating research made by genetists to scholars devoted to demographic history, who are often interested in similar issues (population dynamics, migration phenomena), but did not seem to consider the study of surname distribution as a potential tool for their research.

The 2007 article by Scapoli *et al.* [76] includes a summary and a new analysis of the data collected in twenty years of activity by a wide group of researchers whose most relevant exponent is I. Barrai.

The first researches by the Barrai group, until 1994, focused on some local Italian communities (Ferrara, Perugia, Sicily), and explored both the levels of consanguinity and phenomena due to migration and to isolation by distance.

Later on, when digitalized versions of phone directories became available for many Western countries, the group extended its investigations to all Western Europe, including a very large number of subjects, and covering on average 9 % of the total population of Switzerland, Germany, Italy, Austria, Netherland, Belgium, Spain and France. Some research was devoted also to Venezuela, Argentina and the United States.

The methodology adopted by the Barrai group is clearly exposed in the 1997 paper by Scapoli *et al.* [77] . The main parameters studied and empirically evaluated in all their papers are the isonymy I, intended as a measure of the random component of consanguinity, the α_F coefficient (as defined by Fisher [32]), measuring the surname abundance within a group, the ν index defined by Karlin and Mc Gregor and proportional to the immigration coefficient, and also the entropy of the surname distribution, expressed by the formula $H = \sum p(i) \ln p(i)$. In the large N limit, when Fisher's theory holds, one may derive also the relationship $I = 1/N\nu = 1/\alpha_F + 1/N$. Isolation by distance is computed measuring the correlation between the degree of isonymy and the geographic distance between two populations.

A common feature of all the papers by this group of researchers (clearly inspired by Fox and Lasker's result) is the representation of surname distributions by log-log graphs, turning power laws into linear relationships, and the estimate of the exponent (slope) for each country. Even in the case when all European data are assembled together the power law representation appears to give a reasonable description of the general trend of the distribution.

Very recent contributions, employing and comparing different methodologies in the study of the relationship between population structure and the geographic distribution of surnames, are due to Cheshire *et al.* [78] and to Boattini *et al.* [79].

7. 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 goes back to the Eighties, when the neutralist theory of evolution, stating that the largest part of individual variability has no relevant effects on fitness, became quite popular [80]. 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, that was having notable developments in the same period.

Among the papers anticipating the issues discussed in the present context it is worth recalling the 1991 article by Derrida and Peliti [81], concerning the evolution of a model for a population of asexually reproducing individuals in a flat fitness landscape (that is in the absence of natural selection).

The basic assumptions of the model are the following: the population has a fixed number N of individuals, with a genome characterized by a sequence of binary units, and the probability for an individual to leave behind m offspring (with its same genome) is a binomial, becoming a Poisson distribution in the large N limit; point mutations are allowed with a constant probability for each generation. The genealogy statistics of the model can be represented by the Annealed Random Map model introduced by Derrida and Bessis [82] in order to describe a dynamical system with stochastic dynamics. For large N it is then possible to calculate exactly the statistics of genealogies, that is the set of probabilities $X_{n_1,..,n_k}(t)$, where n_i is the number of individuals who share the i-th ancestor belonging to a set of k individuals who lived at a time t in the past.

This is the starting point for the computation of the average consanguinity $\bar{Y}(t)$, that is the probability that two individuals chosen at random had a common ancestor at the time t. Fluctuations of $Y(t)$ may also be computed. Assuming two individuals to belong to the same family if their last common ancestors lived more recently than t, it is also possible to compute the probability $Z_k(t)$ that there are exactly k families and the average number of families $\bar{\Phi}(t)$. The distribution of sizes of families $f(p_1,..,p_k)$, where p_i is the percentage of individuals belonging to the i-th family in a set of k families, is related by a simple integral transform to $X_{n_1,..,n_k}$, and it can therefore be computed starting from $Z_k(t)$. The probability distribution $\pi(Y)$ of $Y(t)$ is also obtained from $Z_k(t)$, noticing that $Y(t) = \sum_i p_i^2$, where the sum runs over all families. The mutation probability plays a role in the evaluation of the average genetic variability of a population and of its fluctuations, that do not vanish even in the limit of an infinite population. Many features of the model bear a strong resemblance to similar properties found in the study of disordered systems, like spin glasses.

In the immediately subsequent paper by Serva and Peliti [83] a model of sexual reproduction was considered, and it was shown that in this case fluctuations in the genetic distance among individuals tend to vanish in the limit of an infinite population.

The statistical properties of genealogical trees in a neutral model of a closed population with sexual reproduction and nonoverlapping generations were quantitatively studied in 1999-2000, by theoretical and numerical methods, in the papers by Derrida, Manrubia and Zanette [84, 85, 86] .

Starting from an observation by Ohno [87] regarding the repetition of ancestors, Derrida *et al.* defined $M(r)$ as the number of ancestors appearing r times in a given genealogical tree (more precisely in an *Ahnentafel* or ancestor table) and the function $F(r) = M(r)/N_A$, where N_A is the total number of different ancestors. They computed by a numerical simulation of the model the distribution of repetitions $F(r)$ as a function of N_A and of the number of generations G and showed that the model could reproduce the empirical values for the repetition of ancestors in the genealogical tree of Edward III (1312-1377), king of England (In passing we notice that it would be quite interesting to extract form the large genealogical data bases now available online a statistically significant collection of values of $F(r)$).

They also measured the probability of repetitions $H(r, n_g)$ at every past generation $n_g \leq G$, making reference to the whole population N at generation n_g. 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 $P(w) \equiv 2^{n_g} H(r, n_g)/N$ as a function of $w \equiv rN/2^{n_g}$, 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 $S(n_g)$ of the total population in the oldest generation which is absent from a given genealogical tree can also be estimated, and asymptotically, for large values of n_g, the proportion of individuals without descendants reaches a fixed value $S^* \approx 0.203$.

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 $g_{n_g}(z)$ for the weights $w(n_g)$ satisfies a recursion equation that has the form of a renormalization group transformation. In the large n_g limit the generating function converges to $g(z)$, the solution of

$$g(z) = \exp[2g(z/2) - 2].$$

The fraction of individuals with no descendants is $S^* = g(-\infty)$ and therefore it solves $S^* = \exp(2S^* - 2)$; numerically $S^* = 0.203$. One may also extract the exponent β finding $\beta = -\frac{\ln S^*}{\ln 2} = 0.299$, in excellent agreement with the results of the simulation.

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 [88], 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). 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 [89].

It is worth recalling that the notion of MRCA has a 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 may be flushed out completely in roughly 1000 years.

8. The frequency distribution of surnames (II)

After the publication of the results by Miyazima *et al.* [90] concerning the surname distribution in Japanese towns, in which the authors found that the number of different family names scales as a power law of the population and also the frequency distribution of Japanese family names is described by a power law, Zanette and Manrubia [91] addressed the problem of the frequency distribution of surnames, starting from the stochastic model proposed by Simon [43] in 1955.

It is worth recalling that Simon's model is a mathematical representation of a generic Yule process, as defined by Yule [94] in 1925 for the purpose of studying the statistics of biological taxa.

In the original language of species and genera, the process is described as follows: species are added to genera by speciation, which is assumed to happen at some stochastically constant rate, and as a consequence a genus with k species will gain new species at a rate proportional to k. However it may happen (with probability α) that the new species produced is sufficiently different from the others in its genus as to be considered the founder of a new genus.

In the context of surname evolution the model considers a growing population, a probability α (different from zero) for the attribution of a new surname to a newborn and a probability $1-\alpha$ for the newborn to belong to one of the existing families, with a chance proportional to the family size; under these assumptions the evolution equation for the average number of families $N_k(s)$ with exactly k individuals at step s is (according to Simon):

$$N_k(s+1) = N_k(s) + \beta(s)[(k-1)N_{k-1}(s) - kN_k(s)], \qquad N_1(s+1) = N_1(s) + \alpha - \beta(s)N_1(s),$$

where $\beta(s) \equiv (1-\alpha)/N(s)$ and $N(s) = N(0) + s$ is the total size of the population.

This equation was considered by Panaretos in the approximation $N_i(s+1) - N_i(s) \approx N_i(s)/N(s)$. However it is possible to solve the equation completely for arbitrary initial conditions, since one easily finds

$$N_1(s) = \frac{\alpha}{2-\alpha} N(s) + \Big(N_1(0) - \frac{\alpha}{2-\alpha} N(0) \frac{\Gamma(N(0))\,\Gamma(N(s)-1+\alpha)}{\Gamma(N(s))\,\Gamma(N(0)-1+\alpha)} \Big),$$

and all $N_k(s)$ with $i > 1$ can be recursively obtained.

The term containing all information about the initial condition is in general decreasing as $s^{\alpha-1}$, and as a consequence the recursion may be analyzed asymptotically by the Ansatz $N_k(s) \approx P(k)N(s)$, leading to the (Panaretos) equation $P(k) = (1-\alpha)[(k-1)P(k-1) - kP(k)]$, solved by the Beta function $B(k, 1+1/(1-\alpha))$, which Simon called the Yule function.

Therefore the distribution evolves in time towards a (negative) power law, with an exponent $1+1/(1-\alpha)$ going to 2 from above when α goes to 0. Since the time interval between steps is inversely proportional to the total population, the evolution equations imply that the total population and the number of surnames grow exponentially in time.

When a mortality rate $\hat{\mu}$ is introduced in the model, the exponent of the power law is modified to $1 + (1-\hat{\mu})/(1-\hat{\mu}-\alpha)$, but the $\alpha \to 0$ limit is not affected. A disturbing aspect of this theoretical behavior is the fact that the empirical values of the exponent are always lower than 2.

In the same period also Bartley *et al.* analyzed the mechanisms leading asymptotically to a power law behavior [95]. They considered a model admitting birth and death and the creation of new kinds, and a discrete evolution equation similar to the one examined by Manrubia and Zanette, but they managed to approximate it with a continuum equation for the distribution of surname frequencies $n(x)$, where the variable x replaces the number k of individuals belonging to a family. The resulting equation is of the Fokker-Planck type:

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

where $k \equiv k_B - k_D$ and $k_E = k_B + k_D$, k_B and k_D representing (constant) birth and death rates, and the mutation rate appears in a boundary condition at $x = 1$:

Assuming exponential growth in time, a solution of this equation can be found and expressed in terms of the Kummer function $U(k'/k, 0, 2kx/k_E)$, where $k' = k + \lambda k_E$ and λ is related to the singleton influx rate. 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.

Quite recently Maruvka *et al.* [96, 97] reconsidered the problem of the surmame frequency distribution in the continuum limit, and for small birth and death rates they obtained an equation showing the same structure as the one derived by Bartley *et al.*, but for the explicit presence of the mutation rate in the coefficient of the first term in the r.h.s. and for the appearance of the coefficient σ^2 (standard deviation of the offspring distribution) instead of k_E in the last term. In this approximation the resulting large x power law behavior is the same as found by Manrubia and Zanette. In the same paper the authors considered also the problem of family size statistics in a given size subsample of the population.

A quite different approach to the problem of explaining the appearance of power laws was proposed in 2002 by Reed and Hughes [98]. 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 by the growth ratio δ of the population, but not dependent on the rate of immigration [99].

A systematic description of the ubiquitous presence of power laws in the description of natural phenomena, of their properties, and of the possible dynamical schemes leading to their appearance was offered in 2005 by M.E.J. Newman [100].

Newman emphasized that one of the most convincing and widely applicable mechanisms for generating power laws is the above described Yule process. An alternative view was originated by the study of critical phenomena (second order phase transitions) in statistical mechanics, and led to the notion of self-organized criticality: as first proposed by Bak, Tang and Wiesenfeld [101], some dynamical systems might possess stable critical points, and therefore they would tend to evolve towards criticality, independent of the initial conditions, and to be characterized by power law behavior. A simple model of self-organized biological evolution was proposed by Bak and Sneppen [102, 103] and solved by de Boer *et al.* [104]

The most effective approach to the treatment of critical phenomena is based on the renormalization group techniques, which may very well apply to the description of evolutionary models, as discussed in Section 10, despite the still controversial status of self-organized criticality intended as a universal explanation for evolutionary dynamics.

A phenomenological approach to the fitting of family name distributions, with special reference to the Japanese family names studied by Miyazima et al., was recently advocated by Yamada and Iguchi who stressed the advantages of employing the q-exponential function [105].

9. The master equation approach

Recent studies originated by the diachronic and synchronic analysis of Korean surnames [106, 107] led to the formulation of a model for population dynamics expressed in terms of a master equation [108], whose variables are the probabilities $P_{j,s}(k,t)$ for a a family to have k members at time t if the number of members at time s was j (with the obvious initial condition $P_{j,s}(k,s) = \delta_{jk}$. The time evolution of $P_{j,s}(k,t)$ is governed by the differential equation

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

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 functions $\lambda(k,t)$ (birth rate), $\mu(k,t)$ (death rate) and $\beta(k,t)$ (surname creation rate), with the assumption that all these probabilities are proportional to k and have the same time dependence $\phi(t)$.

The master equation can be formally solved for assigned values of j and s by the introduction of a generating function

$$\Psi_{j,s}(z,t) = \sum_{k=0}^{\infty} P_{j,s}(k,t)z^k,$$

satisfying the partial differential equation

$$\frac{1}{\lambda\phi}\frac{\partial\Psi}{\partial t} = (z-1)(z-\bar{\mu})\frac{\partial\Psi}{\partial z},$$

where $\bar{\mu} = (\mu+\beta)/\lambda$.

In particular, introducing a new "time" variable τ defined by $d\tau \equiv \lambda\phi\, dt$, one can solve explicitly the case $j = 1$ (assuming the family originated at time s), finding

$$P_{1,s}(0,t) = \bar{\mu}\eta \qquad\qquad P_{1,s}(k,t) = \eta^{k-1}(1-\eta)(1-\bar{\mu}\eta),$$

where $\eta = 1 - R/1 - \bar{\mu}R$, with $R \equiv \exp[-(1-\bar{\mu})(\tau-\sigma)]$; notice that the expression for $P_{1,s}(0,t)$ corrects a serious misprint in the original paper.

Therefore the overall distribution of surnames in the population can be obtained once the number of surnames appeared at each time s is known. $P(k,t)$ is therefore fully determined given the function $\Pi(s)$ representing the rate at which families are introduced. The number of family names is $N_f(t) = \int_0^t \Pi(s)\, ds$, and the resulting population distribution at time t is given by

$$N_f(t)P(k,t) = \int_0^t P_{1,s}(k,t)\,\Pi(s)\, ds.$$

Defining $\bar{k}_s(t) = \sum_k k\, P_{1,s}(k,t)$ (expected family size at time t), the total population is

$$N(t) = \sum_k k\, N_f(t)P(k,t) = \int_0^t \bar{k}_s(t)\,\Pi(s)\, ds.$$

The above described model is quite general, and it encompasses in particular the widely used Simon model, which can be seen as the special case when $\Pi(s)$ is constant and $\phi(t)$ is proportional to $1/N(t)$, where $N(t)$ is the total population at time t.

It is possible to represent the case of a 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: $\Pi(s) = M + \beta N(s)$.

An exponentially growing population can be described by the assumption that $\phi(t) = 1$, implying that $(1 - \bar{\mu})(\tau - \sigma) = (\lambda - \mu - \beta)(t - s)$. It is then possible to show that $\bar{k}_s(t) = \exp[(\lambda - \mu - \beta)(t - s)]$ and

$$N(t) = N(0)\exp[(\lambda - \mu)t] + \frac{M}{\lambda - \mu}(\exp[(\lambda - \mu)t] - 1).$$

As expected the result is independent of the coefficient β, related to surname change and therefore not affecting the total population.

In the absence of mutations the probability for a family to have k members at time t is given by the (time dependent) Fisher distribution:

$$P(k,t) = \frac{1}{\lambda t k}\eta(t)^k, \qquad \eta(t) = \frac{1 - \exp[-(\lambda - \mu)t]}{1 - \bar{\mu}\exp[-(\lambda - \mu)t]} \equiv \frac{N(t)}{N(t) + \frac{M}{\lambda}}, \qquad N_f(t) = Mt,$$

and in the long run, since $\eta \to 1$, $P(k,t)$ becomes proportional to $1/k$, while the number of surnames $N_f(t)$ grows in time like $\ln N(t)$. These results seem to describe quite accurately the case of Korea and China, as they are known on historical grounds.

When admitting the possibility of mutations, since $N(t)$ grows exponentially, in the long run also $\Pi(s)$ and the number of surnames grow proportionally to N. It is possible to show that the asymptotic behavior of $P(k,t)$ is proportional to $k^{-\gamma}$ where

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

The exponent of the distribution is near to 2, and in any case it depends on the growth ratio of the population and on the mutation coefficient, in agreement with the result of Reed and Hughes. In the limit $\beta << \lambda - \mu$ the number of family names is proportional to the total population: $N_f/N \to \beta/(\lambda - \mu)$.

However one must keep in mind that the limit $\beta \to 0$ is singular, as implied by the previous results.

Comparison with previous results by Manrubia and Zanette may be obtained keeping in mind the following correspondences: $\beta = \alpha\lambda$, $\mu = \hat{\mu}\lambda$ and $\gamma = 1 + c$.

A subtle criticism to all 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. Because of the relevance of this issue and since no systematic treatment is available in the literature, we devoted Appendix A to a discussion of the effects of sampling on generic discrete frequency distributions and Appendix B to the description of a possible parametrization for distributions originated by sampling of large populations.

10. The renormalization group approach

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 [109] and later reformulated by several authors.

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 first instances of a concrete application of these methods to birth-death processes can be found in the papers by Goldenfeld [110] and by Peliti [111], who cast the Fock space formalism in a path integral form and applied it to general birth-death processes on a lattice, recovering existing field theories of such processes in the continuum limit.

Further significant developments are contained in the paper by Jarvis *et al.* [112], aimed at a path integral formulation for branching processes, including the Feynman rules for a concrete evaluation of probabilities in the perturbative regime.

The RG approach was applied directly to the study of family name distribution in an article by De Luca and Rossi [113] . 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

$$H(t) = f(a_{t+1}^{\dagger})a_t, \qquad U(t) = exp(H(t)),$$

where creation and destruction operators are introduced with the usual commutation rules, the Hilbert space is obtained acting on the vacuum Fock state with polynomials in $a_t^{\dagger}$ and

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

p_n being the probability for an individual to have n sons, with the conditions$f(1) = 1$ and $f'(1) > 1$ (because of the assumption of a growing population).

The evolution equation for the frequency distribution of surnames is obtained from the evolution of the state $|n(t)>=\sum_{k=0}^{\infty} N(k,t)|k,t>$, where $N(k,t)$ is the number of family names represented by k individuals at time t and $|k,t>=(a_t^{\dagger})^k|0>$.

In the case of presence of immigration and absence of mutation, defining $|\theta(t)>=\sum_{k=0}^{\infty}\theta(k)|k,t>$, where $\theta(k)$ is the number of new family names represented by k individuals appearing in the system at each time step, the evolution equation is $|n(t+1)>=U(t)|n(t)>+|\theta(t)>$.

By applying the map $|k,t>\rightarrow z_t^k$ from the Hilbert space to $C^{\infty}[0,1]$ the equation can be turned into

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

where $n_t(z) = \sum_{k=0}^{\infty} N(k,t)z^k$ and $\theta(z) = \sum_{k=0}^{\infty}\theta(k)z^k$, and the equation is formally solved 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.

The average number of individuals and surnames at time t can be easily found to be

$$N_t = (N_0 + \frac{\Theta_0}{m-1})m^t - \frac{\Theta_0}{m-1}, \qquad S_t = S_0 + tT_0,$$

where we have defined $m \equiv f'(1)$ and

$$N_t = n_t'(1) = \sum kN(k,t), \quad S_t = n_t(1) = \sum N(k,t), \quad \Theta_0 = \theta'(1) = \sum k\theta(k), \quad T_0 = \theta(1) = \sum \theta(k)$$

.

Notice that the evolution equation is formally analogous to the equations coming from the RG approach, and can therefore be studied by methods usually applied in the context of critical phenomena. Criticality corresponds to the proximity of the fixed point of $f(z)$, solving $f(z^*) = z^*$. Since $f(z)$ is convex and $m > 1$ there are three solutions: $q < 1$, 1, ∞ and q is attractive, while 1 is repulsive.

Expanding around $z = 1$ one finds that $n(z)$ has a logarithmic singularity $n(z) \approx \ln(1-z)$, implying that, for large k and large t, $N(k,t)$ behaves like $1/k$, and this behavior is totally independent of the initial conditions and of the distribution of the immigrating families.

When immigration is suppressed and mutations are allowed, the Galton-Watson process must be modified since only a fraction $1-\alpha$ of the offspring holds the same family name and the remaining part is added to the families of size 1. The modified evolution equation is

$$n_{t+1}(z) = N_t(f(z^{1-\alpha})) + \alpha m N_t z,$$

and $N_t = N_0 m^t$, since mutations do not contribute to the total number of individuals.

Defining $r(z) = f(z^{1-\alpha})$ one can write down the formal solution of the equation:

$$n_t(z) = n_0(r_t(z)) + \alpha N_0 \sum_{k=0}^{t-1} m^{t-k} r_k(z).$$

No limit in t can exist for $n_t(z)$, but it is possible to obtain a limit for the function $\eta_t(z) \equiv n_t(z)m^{-t}$, basically corresponding to a distribution normalized to the total number of families, and satisfying

$$\eta_{t+1}(z) = \frac{1}{m}\eta_t(r(z)) + \alpha N_0 z.$$

Assuming that in the $t \to \infty$ limit $\eta(z) \to (1-z)^{\gamma-1}$ and applying standard RG arguments one may find that

$$\gamma = 1 + \frac{\ln(m)}{\ln(r(1))} = 1 + \frac{\ln(m)}{\ln(m) + \ln(1-\alpha)} \approx 2 + \frac{\alpha}{\ln(m)}.$$

Again the result is completely independent of the initial conditions and of the offspring distribution. In both cases the exponents are consistent with those obtained by the master equation approach and by previous authors.

11. Conclusions and perspectives

A large amount of theoretical and experimental effort has been spent in the study of surname distributions and in their connection with the genetic properties of human populations. Nevertheless there is certainly much space left for more thorough and systematic comparisons between surname data and Y chromosome haplotypes, notwithstanding some evidence that such comparisons may really make sense only for sufficiently rare surnames. We cannot expect to learn much from DNA studies applied to obviously polyphyletic surnames, apart from the possibility of employing them for the selection of ethnically characterized control groups. The dominant presence of surnames showing multiple origin questions seriously also the possibility of adopting isonymy as a reliable quantitative indicator of inbreeding. In turn the relative abundance of the most common surnames, together with their linguistic and ethnic characterization, makes them quite useful for the study of massive migrations, whose statistical relevance may be estimated only in the presence of a substantial number of recognizable subjects.

The convergent results of a few different approaches in identifying the functional dependence of the exponents appearing in the surname frequency distributions from a few typical parameters (birth, death, mutation and migration rates) is certainly encouraging, especially since comparison with historical data, when it turned out to be possible, seemed to confirm systematically the theoretical predictions. However one cannot forget that all results were derived under the assumption of exponentially growing populations; it would be important to explore different scenarios, possibly with the help of a more systematic use of numerical simulations.

More generally, past and present experience seems to suggest that the study of surnames would greatly benefit from an interdisciplinary approach going beyond the, albeit certainly fruitful, interaction between biologists and physicists, and opening itself systematically to the contributions that may come from experts in linguistics, in historical demography, in family history and in sociology. Such collaborations could help to shed clearer light on some of the above mentioned problems. We only mention here the real dynamics of population growth, the empirical data on migrations, the fixation time of surnames and their single or multiple origin, the experimental relationship between surnames and pedigrees, including the phenomenological estimates of the surname mutation rate, the historical data on the repetition of ancestors, and last but not least the conceptual problem of the real neutrality of surname propagation, which has always assumed without a compelling evidence, especially regards in the light of past social (and ethnic) conventions and practices regards to marriage choices and policies. A few interdisciplinary initiatives have already taken place, and some interesting results have emerged from the interaction [114, 115, 116, 117].

Appendix A. The effects of sampling on discrete frequency distributions

We are considering a set of N objects ("individuals") belonging to S different kinds ("species").

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. According to the definition, the following conditions must be satisfied:

$$\sum_{k=1}^{N} N_k = S, \qquad\qquad \sum_{k=1}^{N} k\, N_k = N.$$

The frequency distribution of a sample is a set of values $\{n_l\}$, satisfying the conditions

$$\sum_{l=0}^{n} n_l = S, \qquad\qquad \sum_{l=1}^{n} l\, n_l = n.$$

Notice that the frequency distribution of a sample includes the value n_0, corresponding to the number of kinds, present in the original set, which are not represented in the sample.

It is in principle possible to compute the probability of any sample distribution $\{n_l\}$ as a function of a given set $\{N_k\}$. To this purpose it is convenient to define the intermediate variables N_{kl}, representing the (random) number of kinds characterized by k objects in the original set and by l ($l \leq k$) objects in the sample. The variables N_{kl} are strongly constrained, since they must satisfy all the conditions:

$$\sum_{l=0}^{n} N_{kl} = N_k, \qquad\qquad \sum_{k=1}^{N} N_{kl} = n_l.$$

The probability $P_{\{N_{kl}\}}$ of a specific configuration $\{N_{kl}\}$ follows from the general probability formula [118]:

$$P_{\{N_{kl}\}} = \binom{N}{n}^{-1} \prod_{k=1}^{N} \left[N_k! \prod_{l=0}^{k} \frac{1}{N_{kl}!} \binom{k}{l}^{N_{kl}} \right],$$

subject to the constraint $\sum_{l=0}^{n} N_{kl} = N_k$.

The probability of finding a frequency distribution $\{n_l\}$ in a sample is then obtained by summing the probabilities $P_{\{N_{kl}\}}$ over all configurations satisfiying the constraint $\sum_{k=1}^{N} N_{kl} = n_l$.

In practice it is not possible to obtain simple closed formulas for $P_{\{n_l\}}$. However we shall be interested only in the expectation values $< n_l >$ of the frequency distribution, and these can be computed rather explicitly starting from the above expressions and from the relationship

$$< n_l >= \sum_{k=1}^{N} < N_{kl} >= \sum_{k=1}^{N} \sum_{\{N_{jm}\}} N_{kl} P_{\{N_{jm}\}}.$$

Straightforward manipulations lead to the results

$$< N_{kl} >= N_k \frac{\binom{k}{l}\binom{N-k}{n-l}}{\binom{N}{n}}, \qquad\qquad < n_l >= \frac{\sum_{k=1}^{N} N_k \binom{k}{l}\binom{N-k}{n-l}}{\binom{N}{n}}.$$

In order to fully appreciate the relevance of considerations based on the expectation values we must evaluate the weight of the fluctuations:

$$< n_l^2 > - < n_l >^2 = \sum_{k,k'} N_k N_{k'} \binom{k}{l}\binom{k'}{l} \left[\frac{\binom{N-k-k'}{n-2l}}{\binom{N}{n}} - \frac{\binom{N-k}{n-l}}{\binom{N}{n}} \frac{\binom{N-k'}{n-l}}{\binom{N}{n}} \right] + \sum_{k} N_k \binom{k}{l} \left[\frac{\binom{N-k}{n-l}}{\binom{N}{n}} - \binom{k}{l} \frac{\binom{N-2k}{n-2l}}{\binom{N}{n}} \right].$$

A very relevant limit of the above result may be obtained when considering the (rather typical) case $k, l << N, n$. In this limit

$$< n_l >= \sum_{k=1}^{N} N_k \binom{k}{l} (\frac{n}{N})^l (1 - \frac{n}{N})^{k-l} \equiv \sum_{k=1}^{N} N_k P_{kl},$$

where the definition of P_{kl} follows from the above equation.

We can also estimate the behavior of fluctuations when $k, l << N, n$:

$$< n_l^2 > - < n_l >^2 = \sum_k N_k P_{kl}(1 - P_{kl}).$$

The above expression is always smaller than $< n_l >$ and as a consequence fluctuations become irrelevant for sufficiently large values of $< n_l >$.

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: defining

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

replacing the expression found for $< n_l >$ in $g(z)$ and exchanging the order of summations we obtain

$$g(z) = G(z).$$

It is very important to be able to define a set of expectation values that are independent of the size of the sample, and therefore may reflect very directly the properties of the original frequency distribution.

Some algebraic manipulation allows to prove that

$$\binom{n}{p}^{-1} \sum_{l=p}^{n} < n_l > \binom{l}{p} = \binom{N}{p}^{-1} \sum_{k=p}^{N} N_k \binom{k}{p}.$$

Hence the following equations hold for all $p \leq n$:

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

The expectation values of the "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 $p \leq n$.

In the limit $k, l << N, n$ the definition of $m_p^{(n)}$ simplifies to

$$m_p^{(n)} \to \frac{1}{n^p} \sum_{l=p}^{n} n_l \binom{l}{p},$$

and the fact that $< m_p^{(n)} >$ are independent of the size of the sample then follows as a trivial consequence of the relationship $g(z) = G(z)$, holding in the same limit and allowing the interpretation of $G(z)$ as the generating function of the invariant moments. The explicit expression of the first few invariant moments is:

$$M_0 = S, \qquad M_1 = 1, \qquad M_2 \to \frac{1}{2\,N^2} \sum_{k=1}^{N} k(k-1) N_k = \frac{1}{2\,n^2} \sum_{l=0}^{n} l(l-1) < n_l > \equiv \frac{1}{2\,\alpha_F}.$$

An important test of randomness in sampling is offered by the measure of the correlation between two different samples. Let's consider two random samples, characterized by their sizes n and m. The expected correlation between the two samples can be expressed in terms of the above defined parameter α_F:

$$< C >= \frac{\frac{1}{\alpha_F} + \frac{1}{N}}{\sqrt{\frac{1}{\alpha_F} + \frac{1}{n}} \sqrt{\frac{1}{\alpha_F} + \frac{1}{m}}}.$$

Appendix B. Parametrization of surname distributions

For the purpose of fitting surname distributions originated by sampling of large populations, it is especially interesting to consider the class of negative binomial distributions, such that

$$N_k = \frac{N}{x}\frac{(1-x)^{1-c}}{\Gamma(1-c)}\frac{\Gamma(k-c)}{k!}x^k,$$

where $0 < x < 1$ and the parameter c is assumed to vary in the range $0 \leq c < 1$.

The asymptotic behaviour of the distribution for large k is easily obtained by observing that in this limit

$$N_k \to \frac{N}{x}\frac{(1-x)^{1-c}}{\Gamma(1-c)}\frac{x^k}{k^{1+c}}.$$

One can now compute the generating function for the expectation values of the samples according to the general rule previously discussed and, having defined

$$y = \frac{\frac{n}{N}x}{(1-x+\frac{n}{N}x)},$$

one finds the very convenient property that that the distribution of the samples has the same form as the original distribution, once the replacements $N \to n$ and $x \to y$ have been performed.

It is possible to define the invariant combination of parameters

$$\beta_F \equiv (1-c)\alpha_F = N\frac{1-x}{x} = n\frac{1-y}{y},$$

independent of the dimension of the sample, and it is useful to represent x and y in a form showing their dependence on the dimension of the sample and on the invariant parameter β_F:

$$x = \frac{N}{\beta_F + N}, \qquad\qquad y = \frac{n}{\beta_F + n}.$$

As long as β_F has a finite value, the large N limit corresponds to the limit $x \to 1$. The most attractive feature of these distributions is then the property that in the large N limit, for sufficiently large values of k, the values of N_k tend to satisfy the recursion equation characterizing the Yule function.

It is now possible to evaluate the invariant moments from the expression

$$G_c(z) - G_c(0) = \frac{\beta_F}{c}\Big[1 - \Big(1 + \frac{z}{\beta_F}\Big)^c\Big],$$

showing no explicit parametric dependence on N, as expected; we therefore obtain (for $p \neq 0$)

$$M_p = \frac{\beta_F^{1-p}}{\Gamma(1-c)}\frac{\Gamma(p-c)}{p!} \to \frac{1}{\Gamma(1-c)}\frac{\beta_F^{1-p}}{p^{1+c}}.$$

The limit of the above results when $c \to 0$ is smooth, and it corresponds to Fisher distribution, such that

$$N_k = \beta_F\frac{x^k}{k} \qquad n_l = \beta_F\frac{y^l}{l}, \qquad M_p = \frac{\beta_F^{1-p}}{p}, \qquad \beta_F = \alpha_F.$$

References

[1] G.H. Darwin, *Marriages between first cousins in England and their effects*, Journal of the Statistical Society 38 (1875) 153-184

[2] G.B.L. Arner, *Consanguineous marriages in the American population*, Columbia Univerity Studies in History, Economics and Public Law (XXXI) 3, New York 1908

[3] M. Kamizaki, *Frequency of isonymous marriages*, Seibutsu Tokei-gaku Zassi, 2 (1954) 292-298

[4] R.F. Shaw, *An index of consanguinity based on the use of the surname in Spanish speaking countries*, Journal of Heredity 51 (1960) 221-230

[5] N. Yasuda and N.E. Morton, *Studies in human population structure*, in "Proceedings of the 3rd International Congress of Human Genetics', eds J.F. Crow and J.V. Neel, pp. 249-265, Johns Hopkins University Press, Baltimore 1967

[6] S. Wright, *Coefficients of inbreeding and relationship*, American Naturalist 56 (1922) 330-338

[7] J.F. Crow and A.P. Mange, *Measurement of inbreeding from the frequency of marriages between persons of the same surname*, Eugenics Quarterly 12 (1965) 199-203;

[8] J.F. Crow, *The estimation of inbreeding from isonymy*, Human Biology 52 (1980) 1-14

[9] P. Holgate, *Drifts in the Random Component of Isonymy*, Biometrics 27 (1971) 448-451

[10] L.L. Cavalli-Sforza and W.F. Bodmer, *The Genetics of Human Populations*, pp. 473-480, Freeman, San Francisco 1971

[11] G.W. Lasker, *A coefficient of relationship by isonymy: A Method for Estimating the genetic Relationship between Populations*, Human Biology 49 (1977) 489-493

[12] I. Hussels, *Genetic structure of Saas, a Swiss isolate*, Human Biology 41 (1969) 469-479

[13] N.E. Morton and I. Hussels, *Demography of inbreeding in Switzerland*, Human Biology 42 (1970) 65-78

[14] N.E. Morton, S. Yee, D.E. Harris and R. Lew, *Bioassay of kinship*, Theoretical Population Biology 2 (1971) 507-524

[15] N.E. Morton, D. Klein, I. Hussels, P. Donival, A. Todorov, R. Lew and S. Yee, *Genetic structure of Switzerland*, American Journal of Human Genetics 25 (1973) 347-361

[16] J. Friedl and W.S. Ellis, *Inbreeding, isonymy and isolation in a Swiss community*, Human Biology 46 (1974) 699-712

[17] W.S. Ellis and J. Friedl, *Inbreeding as measured by isonymy and by pedigrees in Kippel, Switzerland*, Social Biology 23 (1976) 158-167

[18] W.S. Ellis and W.T. Starmer, *Inbreeding as measured by isonymy, pedigrees and population size in Törbel, Switzerland*, American Journal of Human Genetics 30 (1978) 366-376

[19] J. Bourgoin and J. Vu Tien Khang, *Quelques aspects de l'histoire génétique de quatre villages pyrenées depuis 1740*, Population 3 (1978) 633-659

[20] K. Gottlieb (ed.), *Surnames as markers of inbreeding and migration*, Human Biology 55 (1983) 209-408

[21] J.F. Crow, *Discussion*, Human Biology 55 (1983) 383-398

[22] G.W. Lasker and B.A. Kaplan, *Surnames and genetic structure: repetition of the same pairs of names in married couples, a measure of subdivision of population*, Human Biology 57 (1985) 431-440

[23] E.J. Devor, *Matrix Methods for the Analysis of Isonymous and Nonisonymous Surname Pairs*, Human Biology 55 (1983) 277-288

[24] G.W. Lasker, *Surnames in the Study of Human Biology*, American Anthropologist 82 (1980) 525-538

[25] G.W. Lasker, *Surnames and Genetic Structure*, Cambridge University Press, Cambridge 1985

[26] F. Galton and H.W. Watson, *On the Probability of the Extinction of Families*, Journal of the Anthropological Institute of Great Britain and Ireland 4 (1874) 138-144

[27] T.E. Harris, *The theory of branching processes*, Springer, Berlin 1963

[28] K.B. Athreya, P.E. Ney, *Branching processes*, Springer, Berlin 1972

[29] A.J. Lotka, *The extinction of families I-II*, Journal of the Washington Academy of Sciences 21 (1931) 377-380, 453-459

[30] S. Karlin and J. McGregor, *The number of mutant forms maintained in a population*, Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability 4 (1967) 415-438

[31] M. Kimura and J.F. Crow, *The number of alleles that can be maintained in a finite population*, Genetics 49 (1964) 725-738

[32] R.A. Fisher, A.S. Corbet and C.B. Williams, *The Relation Between the Number of Species and the Number of Individuals in a random sample of an Animal Population*, J. Animal Ecology 12 (1943) 42-58

[33] N. Yasuda, L.L. Cavalli-Sforza, M. Skolnick and A. Moroni, *The Evolution of Surnames: An Analysis of Their Distribution and Extinction*, Theoretical Population Biology 5 (1974) 123-142

[34] L.L. Cavalli-Sforza, A. Moroni and G. Zei, *Consanguinity, inbreeding, and genetic drift in Italy*, Princeton University Press, Princeton 2004

[35] W.R. Fox and G.W. Lasker, *The Distribution of Surname Frequencies*, International Statistical Review 51 (1983) 81-87

[36] G. Zei, C.R. Guglielmino, E. Siri, A. Moroni and L.L. Cavalli-Sforza, *Surnames as Neutral Alleles: Observations in Sardinia*, Human Biology 55 (1983) 357-365

[37] G. Zei, R. Guglielmino Matessi, E. Siri, A. Moroni and L. Cavalli-Sforza, *Surnames in Sardinia I. Fit of frequency distributions for neutral alleles and genetic population structure*, Annals of Human Genetics 47 (1983) 329-352

[38] G. Zei, A. Piazza, A. Moroni and L.L. Cavalli-Sforza, *Surnames in Sardinia III. The spatial distribution of surnames for testing neutrality of genes*, Annals of Human Genetics 50 (1986) 169-180

[39] J. Panaretos, *On the Evolution of Surnames*, International Statistical Review 57 (1989) 161-167

[40] J. Panaretos, *A Probability Model Involving the Use ofthe Zero-Truncated Yule Distribution for Analysing Surname Data*, IMA Journal of Mathematics Applied in Medicine and Biology 6 (1989) 133-136

[41] P.C. Consul, *Evolution of Surnames*, International Statistical Review 59 (1991) 271-278

[42] M.N. Islam, *A Stochastic Model for Surname Evolution*, Biometrical Journal 37 (1995) L119-126

[43] H.A. Simon, *On a class of skew distribution functions*, Biometrika 42 (1955) 425-440

[44] G.K. Zipf, *Human Behavior and the Principle of Least Effort*, Addison-Wesley, Cambridge 1949

[45] K.-H. Chen and L.L. Cavalli-Sforza, *Surnames in Taiwan: Interpretations based on geography and history*, Human Biology 55 (1983) 367-374

[46] E. Wijsman, G. Zei, A. Moroni and L.L. Cavalli-Sforza, *Surnames in Sardinia II. Computation of migration matrices from surname distributions in different periods*, Annals of Human Genetics 48 (1984) 65-78

[47] A. Piazza, S. Rendine, G. Zei, A. Moroni and L.L. Cavalli-Sforza, *Migration rates of human populations from surname distributions*, Nature 329 (1987) 714-716

[48] A. Piazza, N. Cappiello, E. Olivetti and S. Rendine, *A genetic history of Italy*, Annals of Human Genetics 52 (1988) 203-213

[49] P. Darlu and J. Ruffié, *L'immigration dans les dpartements franais tudie par la mthode des patronymes*, Population 47 n.3 (1992) 719-734

[50] P. Darlu and J. Ruffié, *Relationships between consanguinity and migration rate from surname distributions and isonymy in France*, Annals of Human Biology 19 (1992) 133-137

[51] P. Darlu, A. Degioanni, J. Ruffié, *Quelques statistiques sur la distribution des patronymes en France*, Population 52 n.3 (1997) 607-634

[52] A. Degioanni, A. Lisa, G. Zei, P. Darlu, *Patronymes italiens et migration italienne en France*, Population 51 n.6 (1996) 1153-1180

[53] G. Zei, G. Barbujani, A. Lisa, O. Fiorani, P. Menozzi, E. Siri and L.L. Cavalli-Sforza, *Barriers to gene flow estimated by surname distribution in Italy*, Annals of Human Genetics 57 (1993) 123-140

[54] C.R. Guglielmino, G. Zei and L.L. Cavalli-Sforza, *Genetic and Cultural transmission in Sicily as Revealed by Names and Surnames*, Human Biology 63 (1991) 607-627

[55] A. De Silvestri and C.R. Guglielmino, *Sicilian Provinces: Population Subdivisions Revealed by Surname Frequencies*, Human Biology 76 (2004) 901-920

[56] A. Rodriguez-Larralde, A. Pavesi, C. Scapoli, F. Conterio, G. Siri and I. Barrai, *Isonymy and the genetic structure of Sicily*, Journal of Biosocial Science 26 (1994) 9-24

[57] A. Piazza et al., *Towards a genetic history of Sicily*, Journal of Cultural Heritage 1 (2000) Sup.39-42

[58] A. Pavesi, P. Pizzetti, E. Siri, E. Lucchetti and F. Conterio, *Coexistence of Two Distinct Patterns in the Surname Structure of Sicily*, American Journal of Physical Anthropology 120 (2003) 195-199

[59] E. Lucchetti, D.M. Battisti, G. Ghisolfi, L. Soliani, *Delimitation and Aggregation between Popultions analyzed by Surname Structure*, International Journal of Anthropology 5 (1990) 49-62

[60] P. Pizzetti, E. Lucchetti, L. Soliani, *L'uso dei cognomi nella ricerca biodemografica*, Popolazione e Storia 1 (2001) 107-136

[61] K. Skorecki, S. Selig, S. Blazer, R. Bradman, N. Bradman, P.J. Waburton, M. Ismajlowicz and M.F. Hammer, *Y chromosomes of Jewish priests*, Nature 385 (1997) 32

[62] M.G. Thomas, K. Skorecki, H. Ben-Amiv, T. Parfitt, N. Bradman, D.B. Goldstein, *Origins of Old Testament priests*, Nature 394 (1998) 138-140

[63] E.W. Hill, M.A. Jobling and D.G. Bradley, *Y chromosome variation and Irish origin*, Nature 404 (2000) 351-352

[64] M.A. Jobling, *In the name of the father: surnames and genetics*, Trends in Genetics 17 (2001) 353-357

[65] B. Sykes and C. Irven, *Surnames and the Y Chromosome*, American Journal of Human Genetics 66 (2000) 1417-1419

[66] G. Zei, A. Lisa, O. Fiorani, C. Magri, L. Quintana-Murci, O.Semino and S. Santachiara-Benericetti, *From surnames to the history of Y chromosomes: the Sardinian populationas a paradigm*, European Journal of Human Genetics 11 (2003) 802-807

[67] F. Manni, B. Toupance, A. Sabbagh and E. Heyer, *New Method for Surname Studies of Ancient Patrilineal Population Structures, and possible Application to Improvement of Y-Chromosome Sampling*, American Journal of Physical Anthropology 126 (2005) 214-228

[68] T.E. King, S.J. Ballereau, K. Schrer, M.A. Jobling, *Genetic signatures of coancestry within surnames*, Current Biology 16 (2006) 384-388

[69] B. McEvoy and D.G. Bradley, *Y-chromosomes and the extent of patrilineal ancestry in Irish surnames*, European Journal of Human Genetics 15 (2007) 288-293

[70] G.R. Bowden et al., *Excavating past population structures by surname-based sampling: the genetic legacy of the Vikings in Northwest England*, Molecular Biology and Evolution 25 (2007) 301-309

[71] T.E. King and M.A. Jobling, *Founders, drift and infidelity: the relationship between Y chromosome diversity and patrilineal surnames*, Molecular Biology and Evolution 26 (2009) 1093-1102

[72] T.E. King and M.A. Jobling, *What's in a name? Y chromosomes, surnames and the genetic genealogy revolution*, Trends in Genetics 25 (2009) 351-360

[73] G. Redmonds, T. King and D. Hey, *Surnames, DNA, and Family History*, Oxford University Press, Oxford 2011

[74] S.E. Colantonio, G.W. Lasker, B.A. Kaplan and V. Fuster, *Use of Surname Models in Human Population Biology: A Review of Recent Developments*, Human Biology 75 (2003) 785-807

[75] P. Darlu, *Patronymes et dmographie historique*, Annales de Démographie Historique 2 (2004) 53-65

[76] C. Scapoli, E. Mamolini, A. Carrieri, A. Rodriguez-Larralde, I. Barrai, *Surnames in Western Europe: A comparison of the subcontinental populations through isonymy*, Theoretical Population Biology 71 (2007) 37-48

[77] C. Scapoli, A Rodriguez-Larralde, M. Beretta, C. Nesti, A. Lucchetti, I. Barrai, *Correlations between Isonymy Parameters*, International Journal of Anthropology 12 (1997) 17-37

[78] J. Cheshire, P. Mateos, P.A. Longley, *Delineating Europe's Cultural Regions: Population Structure and Surname Clustering*, Human Biology 83 (2011) 573-598

[79] A. Boattini, A. Lisa, O. Fiorani, G. Zei, D. Pettener and F. Manni, *General method to unravel ancient population structures through surnames, final validation on Italian data*, Human Biology 84 (2012) 235-270

[80] M. Kimura, *The Neutral Theory of Molecular Evolution*, Cambridge University Press, Cambridge 1983

[81] B. Derrida and L. Peliti, *Evolution in a flat fitness landscape*, Bullettin of Mathematical Biology 53 (1991) 355-382

[82] B. Derrida and D. Bessis, *Statistical properties of valleys in the annealed random map model*, Journal of Physics A 21 (1988) L509-L515

[83] M. Serva and L. Peliti, *A statistical model of an evolving population with sexual reproduction*, Journal of Physics A 24 (1991) L705-L709

[84] B. Derrida, S.C. Manrubia and D.H. Zanette, *Statistical Properties of Genealogical Trees*, Physical Review Letters 82 (1999) 1987-1990

[85] B. Derrida, S.C. Manrubia and D.H. Zanette, *Distribution of repetitions of ancestors in genealogical trees*, Physica A 281 (2000) 1-16

[86] B. Derrida, S.C. Manrubia and D.H. Zanette, *On the genealogy of a population of biparental individuals*, Journal of Theoretical Biology 203 (2000) 303-315

[87] S. Ohno, *The Malthusian Parameter of Ascents: What Prevents the Exponential Increase of Ones Ancestors?*, Proceedings of the National Academy of Sciences of the U.S.A. 93 (1996) 15276-15278

[88] J.T. Chang, *Recent common ancestors of all present-day individuals*, Advances in Applied Probability 31 (1999) 1002-1026

[89] D.L.T. Rohde, S. Olson and J.T. Chang, *Modelling the recent common ancestry of all living humans*, Nature 431 (2004) 562-566

[90] S. Miyazima, Y. Lee, T. Nagamine and H. Miyajima, *Power Law Distribution of Family Names in Japanese Societies*, Physica A 278 (2000) 282-288

[91] D. H. Zanette and S. C. Manrubia, *Vertical transmission of culture and distribution of family names*, Physica A 295 (2001) 1-8

[92] S. C. Manrubia and D. H. Zanette, *At the Boundary between Biological and Cultural Evolution: the Origin of Surname Distributions*, Journal of Theoretical Biology 216 (2002) 461-477

[93] S.C. Manrubia, B. Derrida and D.H. Zanette, *Genealogy in the Era of Genomics*, American Scientist 91 (2003) 158-165

[94] G.U. Yule, *A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis*, Philosophical Transactions of the Royal Society of London B 213 (1925) 21-87

[95] D.L. Bartley, T. Ogden, R. Song, *Frequency distributions from birth, death and creation processes*, BioSystems 66 (2002) 179-191

[96] Y.E. Maruvka, N.M. Shnerb, D.A. Kessler, *Universal features of surname distribution in a subsample of a growing population*, Journal of Theoretical Biology 262 (2010) 245-256

[97] Y.E. Maruvka, N.M. Shnerb, D.A. Kessler, *The birth-death-mutation process: A new paradigm for fat tailed distributions*,PLoS ONE 6(11):e26480, 11 2011

[98] W.J. Reed and B.D. Hughes, *From gene families to incomes and internet file sizes: Why power laws are so common in nature*, Physical Review E 66 (2002) 067103

[99] W.J. Reed and B.D. Hughes, *On the distribution of family names*, Physica A 319 (2003) 579-590

[100] M.E.J. Newman, *Power laws, Pareto distributions and Zipf's law*, Contemporary Physics 46 (2005) 323-351

[101] P. Bak, C. Tang and K. Wiesenfeld, *Self-organized criticality: An explanation of the 1/f noise*, Physical Review Letters 59 (1987) 381-384

[102] P.Bak and K. Sneppen, *Punctuated equilibrium and criticality in a simple model of evolution*, Physical Review Letters 71 (1993) 4083-4086

[103] H. Flyvbjerg, P.Bak and K. Sneppen, *Mean field theory for a simple model of evolution*, Physical Review Letters 71 (1993) 4087-4090

[104] J. de Boer, B. Derrida, H. Flyvbjerg, A.D. Jackson and T. Wettig, *Simple Model of Self-Organized Biological Evolution*, Physical Review Letters 73 (1994) 906-909

[105] H.S. Yamada, K. Iguchi, *q-exponential fitting for distributions of family names*, Physica A 387 (2008) 1628-1636

[106] B.J. Kim and S. M. Park, *Distribution of Korean Family Names*, Physica A 347 (2005) 683-694

[107] H.A.T. Kiet, S.K. Baek, B.J. Kim, H. Jeung, *Korean Family Name Distribution in the Past*, Journal of the Korean Physical Society 51 (2007) 1812-1816

[108] S.K. Baek, H.A.T. Kiet and B.J. Kim, *Family name distributions: Master equation approach*, Physical Review E 76 (2007) 046113

[109] M. Doi, *Second quantization representation for classical many-particle system*, Journal of Physics A 9 (1976) 1465-1478

[110] N. Goldenfeld, *Kinetics of a model for nucleation-controlled polymer crystal growth*, Journal of Physics A 17 (1984) 2807-2821

[111] L. Peliti, P*ath integral approach to birth-death processes on a lattice*, Journal de Physique 46 (1985) 1469-1483

[112] P.D. Jarvis, J.D. Bashford and J.G. Sumner, *Path integral formulation and Feynman rules for phylogenetic branching models*, Journal of Physics A 38 (2005) 9621-9647

[113] A. De Luca, P. Rossi, *Renormalization group evaluation of exponents in family name distributions*, Physica A 388 (2009) 3609-3614

[114] P. Darlu, G. Bloothooft, A. Boattini, L. Brouwer, *The Family Name as Socio-Cultural Feature and Genetic Metaphor: From Concepts to Methods*, Human Biology84 (2012) 169-214

[115] P. Rossi, *La distribution des noms de famille comme outil pour l'analyse des dynamiques migratoires*, in "Un juego de engaños. Nombres, apellidos y movilidad en los siglos XV al XVII" (ed. G. Salinero), pp. 153-159, Madrid 2010

[116] P. Rossi, *La distribuzione dei cognomi come strumento per l'analisi sociale: l'esempio della docenza universitaria*, in "L'Italia dei cognomi" (eds. A. Addobbati, R. Bizzocchi, G. Salinero), pp. 203-207, Pisa University Press, Pisa 2012

[117] S. Nelli, P. Rossi, R. Bizzocchi, *Un progeto di analisi statistica dei dati genealogici relativi a Montecarlo di Lucca in età moderna*, in "L'Italia dei cognomi" (eds. A. Addobbati, R. Bizzocchi, G. Salinero), pp. 209-212, Pisa University Press, Pisa 2012

[118] P. Rossi, *On sampling and parametrization of discrete frequency distributions*, arXiv:1210.1410v1 [physics.data-an]