CHAPTER 1

MULTIPLICATIVE PROCESSES IN SOCIAL SYSTEMS

Damián H. Zanette
Consejo Nacional de Investigaciones Científicas y Técnicas
Centro Atómico Bariloche and Instituto Balseiro
8400 Bariloche, Río Negro, Argentina
E-mail: zanette@cab.cnea.gov.ar

Susanna C. Manrubia
Centro de Astrobiología, INTA-CSIC
Ctra. de Ajalvir km 4, 28850, Torrejón de Ardoz, Madrid, España
E-mail: cuevasms@inta.es

Many quantitative properties of social systems display frequency distributions with long power-law tails. This ubiquitous feature, known as Zipf's law, can be understood as a consequence of the stochastic multiplicative mechanisms that underlie the evolution of those systems. In this contribution, several instances of Zipf's law in social processes are discussed. We review a class of models which have been put forward to explain the occurrence of power-law distributions in a wide variety of systems, ranging from word usage in languages to surname frequencies in human populations.

1. Introduction

Biological populations, including those formed by human beings, are collectively subject to a multitude of actions that shape their evolution and determine their fate within the ecosystem to which they belong. These actions may be of very disparate origins, but always involve a complex interplay between factors endogenous to the population, and external mechanisms, related to the interaction with other populations and with physical environmental factors. The fluctuating nature of such actions, as well as the diversity of their origin, call for a description based on stochastic processes.

Within this kind of formulation, it is explicitly assumed that the parameters that govern the evolution of the population can change with time in irregular ways. For instance, the change in the number $n(t)$ of individuals within the population during a certain time interval Δt can be modelled by means of the discrete stochastic equation

$$n(t + \Delta t) - n(t) = a(t)n(t) + f(t) \tag{1}$$

where $a(t)$ and $f(t)$ are random variables with suitably chosen distributions. The equation may be solved for a specific realization of these random variables but, usually, one is rather interested at finding the statistical properties of $n(t)$ –for example, the expectation value of n at a time t in the future– as a function of the statistical properties of $a(t)$ and $f(t)$. Equations of the type of (1) have been studied in detail by several authors in various contexts, as recently reviewed by Sornette.[1,2]

The two terms in the right-hand side of Eq. (1) have well-differentiated interpretations. The first term, $a(t)n$, represents the contributions to the evolution of n which are proportional to the population itself. Due to this proportionality, such contributions are called *multiplicative*. In a closed population, multiplicative processes are restricted to birth and death, and $a(t)$ stands for the difference between the birth and death rates per individual in the interval Δt. In open populations, the number of individuals is also affected by migration processes. In general, the contribution of emigration is multiplicative-like, because each individual has a certain probability of leaving the population per time unit. On the other hand, immigration has both multiplicative and *additive* effects. Immigration flows can, in fact, be favoured by a large preexisting population –as in big cities– but a portion of arrivals may also occur as a consequence of individual decisions that do not take into account how large the population is. Such additive contribution is accounted for by the second term in Eq. (1). This term can also stand for negative effects on the population growth, such as catastrophic events where a substantial part of the population dies irrespectively of the value of n.[3] More generally, the additive term $f(t)$ describes "reinjection" events, which insure that n remains finite even when multiplicative processes by themselves may imply unbounded growth or eventual extinction of the population.[1]

It can be readily shown that in the absence of reinjection, $f(t) \equiv 0$, and under very general conditions on the statistical properties of the random variable $a(t)$, Eq. (1) implies that the probability distribution $P(n, t)$ for the population n at time t is a log-normal function. If, on the other hand,

$f(t) \neq 0$, the distribution can have a complicated analytical form. It is nevertheless known that, for large n and long times, $P(n,t)$ depends on the population as

$$P(n,t) \sim n^{-1-\gamma}. \tag{2}$$

The exponent γ is determined by the equation $\langle (a+1)^{\gamma} \rangle = 1$, where $\langle \cdot \rangle$ indicates average over the distribution of the random variable a.[1]

Detecting the power-law distribution of Eq. (2) in real systems would require to have access to many realizations of the evolution of the same population –which, in practice, is rarely possible– or, alternatively, to follow the parallel evolution of several populations of the same type. In this second case, it would be necessary that all the populations under study are subject to similar conditions, such as to insure that the parameters that govern the evolution are uniform over the ensemble. These requirements are often met in populations formed by human beings. Due to social, historical, geographical, cultural, and/or economic reasons, human populations happen to be divided into groups of different types. Within each group, all individuals share a distinctive trait, and the "affiliation rules" are such that children belong to the same group as their parents. The creation of new groups is usually rare, and migration between groups is relatively limited.

Consider, for instance, the case of surnames. In the overwhelming majority of cases, they are transmitted unchanged from the father to his children. Surname mutation is infrequent, as it is mostly associated with migration to culturally distant populations. The voluntary change of an individual's surname is even rarer. As a result, human populations are divided into groups where all individuals bear the same surname, and the population in each group evolves almost autonomously. According to the above discussion, it is expected that the distribution of the number of individuals in such groups –given, for instance, by the probability of finding a surname borne by n individuals– displays a power-law tail. In fact, it does, and the same is true in groups such as the speakers of different languages, or the inhabitants of different cities.

Over the past century, the occurrence of power laws in the population distribution of human groups of various kinds has been reported by several authors, notably, by the philologist G. K. Zipf.[4] As a matter of fact, the power-law dependence of the frequency of groups as a function of their population came to be known as Zipf's law. Remarkably, however, the only case discussed in detail by Zipf does not involve the evolution of human populations, but the apparently unrelated question of word usage in written

and spoken language.[5] With the illustration of statistical data obtained by himself and others, Zipf pointed out that, in a text, the number $P(n)$ of words that are used exactly n times decreases with n as

$$P(n) \sim n^{-\zeta}. \tag{3}$$

Equivalently, the probability of finding a word with exactly n appearances follows Eq. (2), with $\gamma = \zeta - 1$. Zipf discovered that, for many texts in different languages, one has $\zeta \approx 2$. In an alternative formulation –which became famous as Zipf's rank analysis– all the different words in a text are ranked according to their number of appearances, with rank $r = 1$ for the most frequent word, $r = 2$ for the second most frequent, and so on. It can be shown that Eq. (3) implies, for the number of appearances n as a function of the rank r, a power-law dependence

$$n(r) \sim r^{-z}, \tag{4}$$

where $z = (\zeta - 1)^{-1} \approx 1$ is usually known as the Zipf exponent. The same type of power-law dependence between frequency and rank is found in surnames ranked by the number of individuals who bear them, languages by the number of speakers, and cities by their population.

The aim of this contribution is to review a class of models that predict the occurrence of Zipf's law in human groups of various kinds. All of them are extensions of Simon's model,[6] which is in turn based on a multiplicative mechanism for the population growth. In the next section, we present Simon's model in the frame where it was originally introduced –word frequency in language. The role of multiplicative mechanisms in language is clarified, in connection with the process of context creation. We discuss some refinements of the model, as well as its application to musical language. Next, we describe how Simon's model applies to the distribution of city sizes and of speakers of different languages, pointing out some open problems. Section 4 is the core of the contribution, and presents an extension of the model including mortality. This extension makes it possible to give a detailed quantitative explanation of the distribution of surnames observed in present-day populations, which may also apply to the distribution of certain genetic traits. Finally, we give a concluding summary.

2. Models for Zipf's law in language

A thorough formulation of a model for Zipf's law was provided in the 1950s by H. A. Simon,[6] elaborating on an idea previously advanced by Willis

and Yule.[7] Simon presented his model by referring to the case of language, which Zipf himself had discussed in detail in one of his books.[5] Some specific features for Zipf's law for language are the following. First, while the exponent z of the power-law decay of the number of occurrences as a function of the rank r, Eq. (4), is generally close to unity, systematic deviations are observed for texts in languages such as Latin and Russian, for which z can be considerably smaller than one. Those languages share the property of being highly inflected, due to the strong variation of both nouns in declensions and verbs in conjugations. For other languages, in contrast, z is larger than one. Second, at high ranks, the number of occurrences as a function of r abandons its power-law dependence, and displays a faster decay. These features are illustrated in Fig. 1.

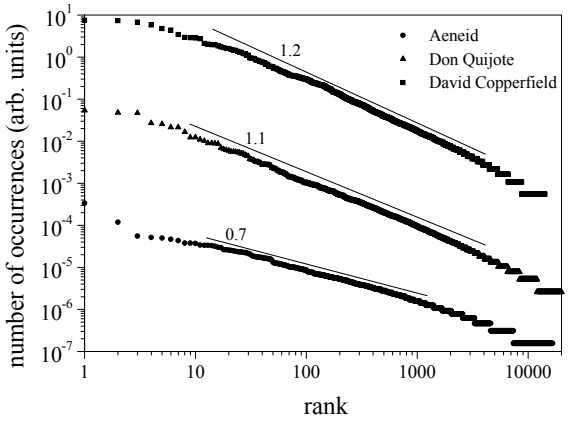

Fig. 1. Zipf's rank plots for Virgil's Aeneid, Don Quijote, by Miguel de Cervantes Saavedra, and David Copperfield, by Charles Dickens. For clarity, the plots have been mutually shifted in the vertical direction, so that the units for the number of occurrences are arbitrary. Straight lines have the slope of least square fittings in the zone where the power-law decay is well defined; labels indicate the slope value.

Simon's model mimics the generation of a text as a stochastic process. At each step, a word is added to the text, according to the following rules. (i) With probability α, a new word –not yet present in the text– is added. (ii) With the complementary probability $1 - \alpha$, an already used word is added. In this case, the word to be added is chosen with a probability proportional to its previous occurrences. Rule (i) implies that the lexicon grows, on the average, at a constant rate as the text progresses. Rule (ii) introduces a multiplicative mechanism that favours the occurrence of those

words which are already frequently used in the text. In this formulation, the only parameter of Simon's model is α, the probability of appearance of a new word.

The two rules defining Simon's model can be translated into mathematical terms, in the form of an evolution equation for $P(n, s)$, the number of words that have occurred exactly n times up to step s. For $n = 1$, we have

$$P(1, s+1) = P(1, s) + \alpha - \frac{1-\alpha}{N(s)} P(1, s), \tag{5}$$

while, for $n > 1$,

$$P(n, s+1) = P(n, s) + \frac{1-\alpha}{N(s)}[(n-1)P(n-1, s) - nP(n, s)]. \tag{6}$$

Here, $N(s)$ is the total text length at step s. If the text generation is assumed to have begun with one word at $s = 0$, we have $N(s) = s + 1$. The above deterministic equations govern the mean evolution of $P(n, s)$. Their solution must be understood as the mean number of words with exactly n occurrences, averaged over many realizations of the stochastic rules (i) and (ii).

Simon himself proved that Eqs. (6) and (5) admit a solution which decays with n as [6]

$$P(n, s) \sim N(s) n^{-1-1/(1-\alpha)}. \tag{7}$$

In the rank plot, this implies a power-law decay with exponent $z = 1 - \alpha$. He showed moreover that this special solution describes the asymptotic distribution $P(n)$ for any initial condition. Thus, a sufficiently long text generated following the rules of Simon's model verifies Zipf's law with the above exponent. Note that the exponent tends to the typical value $z = 1$ for a vanishingly small probability of appearance of new words. For finite α, we have $z < 1$.

Simon's model can be interpreted as an attempt to represent the creation of context as a text is generated. Context is the global property of a structured message that sustains its coherence or, in other words, its intelligibility.[8] A long chain of words, even if they constitute a grammatically correct text, would result incomprehensible if it does not succeed at defining a contextual framework. It is in this framework, created by the message itself, that its perceptual elements become integrated into a meaningful coherent structure. As words are successively added to the text, a context is created which favours the later appearance of certain words –in

particular, those that have already appeared– and inhibits the use of others. The model aims at capturing the essentials of the mechanism which, by repeated use of certain words, is at work in the construction of a structured, comprehensible text. The repetition of perceptual elements is one of the basic ingredients in the conception of intelligible structures and in the ensuing cognitive response to their reception, including the creation and retrieval of memories.[9] Such notion lies at the basis of the cognitive processes associated with written and spoken communication.

Thus, Simon's model interprets Zipf's law as a statistical property of word usage during the creation of context, as a text is progressively generated. Context emerges from the mutually interacting meanings of words, and represents a collective expression of the semantic contents of the message, arising from the multiple structured relations between language elements. Semantics is in fact essential to the function of language as a communication system.

Incidentally, let us mention that B. Mandelbrot pointed out a different –and, in a sense, simpler– mechanism able to give rise to a Zipf-like law for written texts.[10] He proposed to generate a "text" as an array of characters chosen at random from a given alphabet, where the blank space has also a certain fixed probability. "Words" are defined as the sub-arrays between any two consecutive blank spaces. For sufficiently long "texts" of this type, rank plots constructed by counting the number of occurrences of each "word" show a power-law decay with an exponent close to $z = 1$, as in real texts. If Mandelbrot's explanation were right, Zipf's law would lack any linguistic significance. At the level of rank statistics, in fact, a text would not be distinguishable from a random array of characters. Zipf's law should be thought of as a trivial manifestation of this "quasi-randomness" of real texts. This observation gave origin to a lively discussion between Mandelbrot and Simon themselves.[11,12]

Though, sometimes, Mandelbrot's model is still invoked as an explanation for Zipf's law in language, a few important drawbacks strongly suggest that such explanation is not correct. First, the exponent z predicted by Mandelbrot's model depends of the length of the involved alphabet.[13] This dependence of z on the alphabet length is not observed in real texts. Second, Mandelbrot's model implies a specific prediction for the distribution of word lengths. If p_0 is the probability of having a blank space, the probability distribution for the word length l is the exponential $p(l) = p_0(1 - p_0)^{l-1}$. This result, however, bears no relation to real word-length distributions. In the first place, they usually show a maximum at small lengths. In the

case of English, mainly due to the high frequency of the words THE and AND, this maximum occurs at $l = 3$. Moreover, real distributions do not decay exponentially. Language usage heavily penalizes very long words – in English, beyond about $l = 12$. Consequently, the decay of word-length distributions is usually faster than exponential. Finally, we mention that if Mandelbrot's model were correct, the number of different words of a given length l should grow exponentially with l, which is also in disagreement with data from real languages.

As discussed above, Simon's model is able to explain Zipf's exponents lower than one, $z < 1$. However, rank plots for certain languages (such as English and Spanish; see Fig. 1) typically exhibit exponents above unity. To explain this discrepancy, Simon's model can be refined on the basis of linguistically sensible assumptions.[14,15] In fact, probably the most unrealistic hypothesis in the model is the fact that the probability of appearance of new words, α, does not vary as the text progresses. In real texts, this is manifestly false. While during the first stages of the process new words are frequently needed to settle the context, in later stages the lexicon becomes better established and, consequently, its growth rate is lower. A phenomenological representation of this feature consists in assuming that the probability of appearance of new words decays as $\alpha(s) = \alpha_0 s^{\nu-1}$, with $0 < \nu < 1$, as the text is generated. This form for $\alpha(s)$ implies that the lexicon size, i.e. the number of different words, increases as $V(s) \sim s^{\nu}$, while the text length grows as $T(s) \sim s$.

While, in general, it is not possible to solve Eqs. (5) and (6) for s-dependent α, an approximate solution can be found, following the same argument as Simon, if $\alpha(s) = \alpha_0 s^{\nu-1} \ll 1$. Certainly, this inequality holds at least when the initial stages in the text generation have elapsed. Under these conditions, it has been shown that the number of words with exactly n appearances decreases with n as $P(n) \sim n^{-1-\nu}$. This implies

$$z = \frac{1}{\nu} \tag{8}$$

for the power-law exponent in the Zipf's rank plot. Thus, within this extension of Simon's model, exponents larger than one can also be reproduced. Moreover, the result is in agreement with the empirical observation that highly inflected languages (such as Latin) have Zipf exponents smaller than those of less inflected languages (such as English). In fact, as for the number of different words, poorly inflected languages have a more limited lexicon. The vocabulary of texts written in such languages is therefore expected to increase slowly as the text progresses, which corresponds to relatively small

values of ν and, accordingly, large z.

A further extension of Simon's model makes it possible to explain the faster decay of the number of occurrences for high ranks. This extension is also based on linguistic considerations regarding the creation of context as a text is generated. It can be argued that a single appearance of a given word is not enough to establish its role in defining the context. Rather, there should be a threshold in the number of occurrences of a word, before it enters the regime where the multiplicative process of Simon's rule (ii) acts. This effect can be implemented by modifying the probability that a newly introduced word is used again. Namely, the probability that a word with n previous occurrences appears at the current step is taken to be proportional to $\max\{n, \eta\}$, where η is the threshold. In this way, a given word has to appear η times before the multiplicative process begins to act. Until then, the probability of occurrence is constant. The threshold η may be different for each word. Numerical simulations of the extended Simon's model with an exponential distribution for the value of η assigned to each word are able to satisfactorily reproduce the observed decay for high ranks. Within this extension, the fast-decaying tail of Zipf's plot is interpreted as containing those words whose number of occurrences has remained below the corresponding threshold.

In view of the interpretation of Simon's model as capturing the essential mechanisms of the creation of linguistic context, it is natural to pose the question as whether the same model can be applied to other communication systems with a meaningful notion of context. An appealing candidate is music, which –supposedly– shares with language at least some neural mechanisms related to acquisition and perception processes.[16] The crucial difference in nature between the information conveyed by music and language, however, makes it difficult to extend linguistic concepts to the realm of musical expression. Often, such extension remained at a metaphorical level though, recently, scientifically sound definitions for musical syntax, grammar, and semantics have been put forward. On the other hand, the notion of context admits a straightforward extension to music. Musical context is determined by a hierarchy of intermingled patterns occurring at different time scales. The tonal and rhythmic structure of melody motifs constitutes the most evident contribution to musical context. The repetitions, variations, and transpositions of those motifs shape the thematic base of a composition. At larger scales, the recurrence of long sections and certain standard harmonic progressions determine the musical form. Crossed references between different movements or numbers of a given work estab-

lish patterns over even longer times. Meanwhile, at the opposite end of time scales, the duration and pitch relation of a few notes are enough to determine tempo, rhythmic background, and tonality.

Applying Zipf's analysis to music requires first to solve the task of giving a convincing definition to the musical equivalent of "word." The multiplicity of levels at which musical context can be defined suggests several possible identifications for "words" in music, ranging from single notes to rhythmic patterns, to melodic phrases. Many of them have in fact been used to construct Zipf's rank plots for musical compositions. Unfortunately, such studies did not go beyond a phenomenological description, and established no connection with possible models for Zipf's law.[17,18]

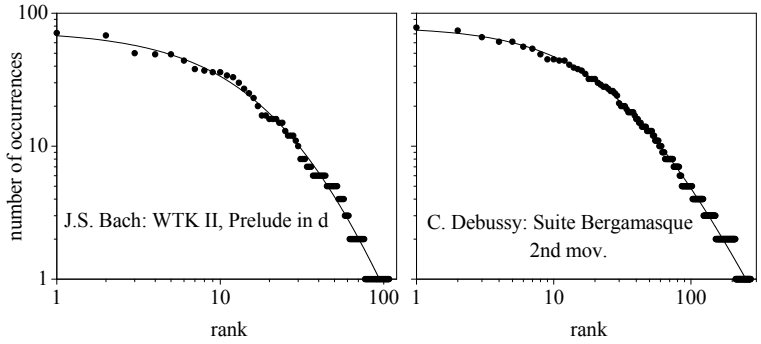

Fig. 2. Zipf's rank plots for the Prelude N. 6 in d from the second book of Das Wohltemperierte Klavier, by J. S. Bach, and the second movement, Menuet, from the Suite Bergamasque by C. Debussy. Curves correspond to least-square fittings with Eq. (9). The resulting exponent is $\nu = 0.28$ for Bach and $\nu = 0.48$ for Debussy.

More recently, however, the significance of Simon's model in music has been assessed on the basis of Zipf's analysis for a set of classical compositions.[19] Due to operational convenience, "words" were identified with single notes, defined by their individual pitch and duration. The contribution of notes to the creation of musical context, determining tonality and rhythm through they relative pitches and lengths, is particularly transparent. Figure 2 shows Zipf's plots for two compositions for keyboard: the Prelude N. 6 in d from the second book of Das Wohltemperierte Klavier, by J. S. Bach, and the second movement, Menuet, from the Suite Bergamasque by C. Debussy. Note that these plots lack the power-law high-rank regime of Zipf's plots for language (Fig. 1). This feature, which can be ascribed

to the relative small "lexicon" size (number of different notes) and "text" length (total number of notes) of musical compositions as compared with language corpora, does not preclude, however, the application of Simon's model. In fact, imposing to Simon's model the additional condition that any given "word" can appear at most a predefined number of times, the functional form of the number of occurrences n in terms of the rank r is

$$n(r) = (a + br)^{-1/\nu}. \tag{9}$$

Here, a and b are constants, and ν is the exponent that defined the "lexicon" growth, $V \sim s^{\nu}$, as discussed above. Least-square fittings of Zipf's plots with Eq. (9) are in excellent agreement with empirical data, supporting the applicability of Simon's model, as a representation of context creation, to musical compositions. The difference in the values of the exponent ν for Bach ($\nu = 0.28$) and Debussy ($\nu = 0.48$) is not unexpected. The exponent becomes even larger for atonal compositions, where the use of elements that determine the tonality context is avoided on purpose. As discussed in the case of language, small exponents correspond to a compact lexicon, determining a rather robust, stable context. Large exponents, on the other hand, determine an abundant lexicon, related to a ductile, more tenuously defined context. The merest comparison of the above compositions clearly reveals this difference to the listener.

3. City sizes and the distribution of languages

Before moving to the core of this contribution, we briefly review in this section two instances of occurrence of Zipf's law in direct relation to human populations. As discussed in the introduction, the nature of the reproduction mechanism of living organisms implies that the overall evolution of any biological population is inherently driven by stochastic multiplicative processes. In the two instances considered here, these processes are reflected in the size distribution of human groups, as their population grows.

Our first instance regards the distribution of city sizes. It is an evident fact that the geographical, political, and socioeconomic factors that determine the sizes of cities, as measured by their populations, are broadly heterogeneous. Accordingly, changes in city populations are quite disparate, even for closely related cities. Think of the fate of a few Western urban settlements during the last five hundred to one thousand years. Venice, for instance, which in the Middle Ages was one of the largest cities in Europe, bears now some $60,000$ inhabitants –approximately, half of its population

three centuries ago. In the same period, Rome multiplied its population by a factor of 100, reaching its present few millions. By the beginning of the thirteenth century, Paris and Florence had approximately equal sizes; now, the former is some 20 times bigger than the latter. As for the cities of the New World, initially modest and precarious settlements such as México, São Paulo, Buenos Aires, and New York have become, in five hundred years, some of the largest metropolitan areas in the globe.

Yet, a rank plot of populations for all the cities in the world shows a well-defined power-law regime over several orders of magnitude, revealing an unexpected regularity in the result of the very non-uniform process of urban growth. And, perhaps more surprisingly, Zipf's law occurs also when the sample is limited to the cities of a given country or region. This is one of the best known occurrences of Zipf's law; it was already quoted by Zipf and Simon themselves. Figure 3 displays rank plots for the largest urban settlements in India, Argentina and France, including some 200 cities each. Data have been obtained from www.citypopulation.de, and correspond to 2001 for India and Argentina, and to 2004 for France.

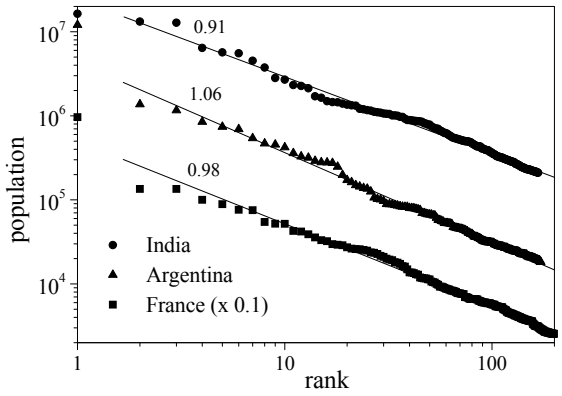

Fig. 3. Zipf's rank plots for the population of the largest cities in India (2001), Argentina (2001), an France (2004). Some 200 cities are considered in each case. Data for France have been multiplied by 0.1, for clarity in the display. Straight lines stand for least square fittings. The corresponding Zipf's exponents are shown as labels. Source: www.citypopulation.de.

Such ubiquitous regularity calls for an explanation based on universal mechanisms and, of course, it is natural to think of the multiplicative processes that govern the evolution of populations. Larger cities grow faster,

first, due to the reproduction of its inhabitants. But also the effect of immigration, which cannot be neglected in the change of city sizes, is expected to be multiplicative in nature. The accumulation of wealth and resources in a given city should be proportional to its size, at least within geopolitically uniform regions. Consequently, its appeal to immigration should increase as its population grows. The basic mechanism of rule (ii) in Simon's model is thus at work. Each time a new inhabitant is added to the system, their destination city is chosen with a probability proportional to its current population. Rule (i) requires, in addition, to have a finite probability of foundation of a new city when the new inhabitant appears. In practice, such probability must be extremely small.

For city sizes, the variation of the Zipf exponent z between countries is more restricted than in the case of word frequencies between different languages. In the former case, Zipf exponents are rarely below 0.9 or above 1.1. A regularity has however been reported in the variation of z: the Zipf exponent is systematically smaller for old countries (as, for instance, in Europe and Asia) than for young countries (as in the Americas). Figure 3 illustrates this fact. Exponents larger than one –such as that of Argentina, $z = 1.06$– can be readily explained using the extension of Simon's model discussed in the previous section, which admits that the probability of creation of new cities decreases as time elapses. On the other hand, while the original form of Simon's model could explain an exponent lower than one –such as that of India, $z = 0.91$– it would require a very large value of the probability α. In the case of India, we would have $\alpha = 1 - z = 0.09$, which would imply that, roughly, a new city is created every ten new inhabitants in the country! Clearly, another mechanism is needed to explain such small exponents as that of India. Geographers suggest that an important ingredient may be given by the fact that the growth rate of an existing city is not necessarily proportional to its current size, as assumed in rule (ii) of Simon's model. In particular, a dependence on the size that penalizes large populations would produce an overall flattening of the rank plot, with the ensuing decrease of z. To our knowledge, the extension of Simon's model with size-dependent growth rates has not been studied yet.

The application of Zipf's analysis and Simon's model to urban settlements is implicitly assuming that individual cities are well-defined entities. Actually, urbanists may not agree on this point. The modern city is such a complex of intermingled systems that it defies a definition in terms of traditional classification schemes, and requires a wider concept of class.[20] Figure 4 illustrates the fact that, while urban settlements can be distinctly

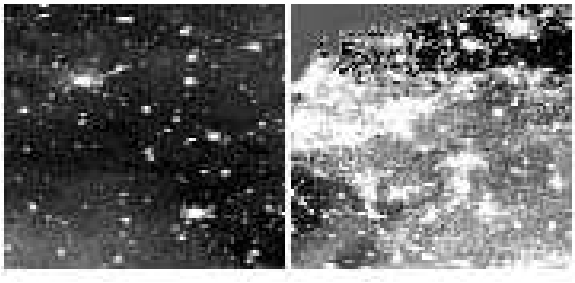

Fig. 4. Two satellite images of the Earth by night. Left: Central Ukraine. Right: Northwestern Germany. Each image covers an area of, roughly, 500×500 km^2. Source: visibleearth.nasa.gov.

identified in some regions, in other places the situation is much less clear cut. Currently, it is accepted that –at the level of big cities– the entities to be considered in Zipf's analysis are the clusters resulting from the growth and aggregation of initially separated settlements. Administrative divisions, usually inherited from those initial conditions, do not play a substantial role in defining such metropolitan areas. Figure 3 was drawn taking this criterion into account. This discussion rises the question on the origin of Zipf's law for urban agglomerations. It would be interesting to consider an extension of Simon's model incorporating the formation of aggregates, and determine which features in the aggregation mechanism insure that Zipf's law holds for the resulting system of cities and urban clusters.

The second instance of Zipf's law considered in this section regards the number of speakers of different human languages. At the present day, some $5,000$ to $6,000$ different languages are spoken all over the world. Their distribution and diversity, which have been determined by both historical and geographical factors, are extremely heterogeneous. For instance, about $1,000$ different languages –all of them belonging to the Indo-Pacific family– are spoken in New Guinea and neighbouring islands while, in turn, practically all the American countries to the south of the United States (Brazil being the most noticeable exception) have Spanish as their main mother language. The number of Native American languages, on the other hand, had certainly reached several hundreds before the European invasion in the sixteenth century.[21] In correspondence with this heterogeneity, the number of speakers per language varies between several hundred millions for Chinese and some languages of the Indo-European family, to a mere handful of speakers for those hundreds of languages that are presently on the edge

of extinction.

Notice that the same warning put forward above on the entity of cities applies to languages. Usually, a language is accompanied by a host of regional variations, dialects, and jargons, that make it difficult to give a neat definition of geographical boundaries and historical domains. Nevertheless, linguists seem to have reached a reasonably general agreement on the entity of a large number of languages, and the size of the respective populations has been determined. Figure 5 shows a rank plot of the first $1,000$ languages ordered according to the corresponding number of speakers. The plot begins with a zone where the Zipf exponent is close to unity. Soon, however, the exponent changes to a much higher value, $z \approx 1.8$. This is, in fact, the highest Zipf exponent among the several instances discussed in this article.

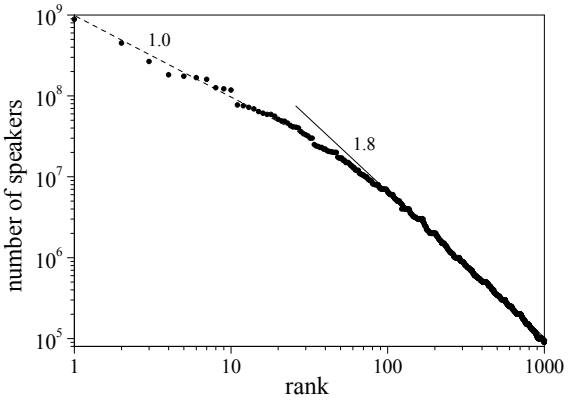

Fig. 5. Zipf's rank plot for the number of speakers per language, for languages with more than $\sim 10^5$ speakers. Source: www.etnologue.org.

The occurrence of Zipf's law for the number of speakers per language can be readily understood in terms of the multiplicative mechanisms that underly the growth of the respective populations. In this process, it is essential that –in the overwhelming majority of cases– an individual inherits the language of their parents, so that they belong to the same speaker population. The situation is similar to that of family names, that we discuss in detail in next section. The probability of creation of new languages should be very small. In the frame of Simon's model, a Zipf's exponent $z \approx 1.8$ can be explained by means of the extension discussed in the previous section, with a decreasing frequency of language creation. According to Eq. (8), the corresponding exponent would be $\nu \approx 0.56$.

The presence of a power-law regime with a different Zipf exponent for the languages with the largest numbers of speakers –some 20 languages spoken by, roughly, more than $50,000,000$ people– is intriguing.[22] However, the populations associated with most of these languages have evolved in the last few centuries through mechanisms that may not be well described by the local multiplicative processes of Simon's model and its variations.[23] The relatively rapid expansion of these languages over vast geographical domains, through invasion –pacific or violent–, conquest, and massive migration, may imply that the spatial variable cannot be ignored in a description of their evolution. The already mentioned case of Spanish is a clear example: some 90 % of the present-day Spanish speaking population was not born in Spain, and much of it is ethnically non-European. A case not related to classical colonialism is that of Turkish: it is spoken by more than 60 million people, one third of them outside Turkey. The quantitative modelling of the distribution of these geographically very extended languages is an open problem.

4. Family names

It belongs to common experience that the pedigree of an individual can be traced back for many generations, often following the line that links fathers to sons. The frequency of surnames is one of the clearest cases of multiplicative growth of a cultural feature, and has been studied using different approximations for at least one century. The similarity of this problem with some questions put forward in the field of population genetics has favored that, nowadays, we enjoy a deep understanding of the main mechanisms at play. In this section, we briefly review the historical development of problems related to surname inheritance and the models proposed to explain its dynamics, and analyze the sociological and historical context of a number of present-day populations.

The end of the nineteenth century witnessed the first attempt to formulate and solve a sociological problem mathematically. The problem arose when it was noted that certain families "of men of genius" tended to perish, as the disappearance of certain surnames seemed to indicate. The problem was qualitatively addressed by Sir Francis Galton, who at the time gave an explanation based on his belief that a rise in intellectual capacity somehow implied a diminution in fertility. A contrasting point of view was that of Alphonse de Candolle, who pointed out that the unavoidable fate of a surname is to disappear simply due to the stochastic nature of the inheritance process. The mathematical formulation of the problem, and a first solution,

came from the study of Rev. H. W. Watson, who correctly concluded that any surname is bound to disappear in constant or shrinking populations, without the need to invoke differential fertility of the individuals.[24]

It took several decades to relate the problem of family name inheritance to the genealogy of non-recombining alleles (or of genetic heterogeneity) in a population.[25] Some parts of the human genome, among them the Y chromosome and the mitochondrial DNA, are inherited from one of the parents only, and do not experience recombination in the process. Hence, they are transferred unaltered, except for rare mutations, from generation to generation. The dynamics of this process correspond to a monoparental way of transmission affected by population fluctuations, and is completely analogous to surname inheritance. The correlation between the two processes is strong enough that, occasionally, the surname of certain patrilineal families clearly correlates with the inherited characteristics of the Y chromosome.[26]

Regarding the disappearance of surnames, the interest was initially directed to estimate the probability that a surname perished as a result of the randomness inherent to the transmission process. To solve that problem, a formulation fully analogous to the fixation of a mutant allele in a population was proposed.[27] The first statistical approaches to the description of surname abundance [28] came much later, and took advantage of neutral models initially devised to quantify the number of different alleles that could be maintained in a population.[29]

In the framework of those stochastic models, the trait under consideration evolves neutrally, that is, it does not confer any selective advantage to the individual carrying it. While this statement is difficult to prove in a genetic context, it is much more easily verified in the case of family names. This approach yields a number of exact results, including the probability for a trait to survive at any time in the future and the average number of different traits that can stably coexist in a large population. In particular, for a population to be heterogeneous with respect to a certain trait, a sufficiently high rate of appearance of new variants is required.[30] Consider a population of constant size evolving by non-overlapping generations, and initially homogeneous with respect to a certain character. Suppose that a mutant appears. Neutral theory states that the typical number of generations g for the mutant to be fixed under the action of random drift is of the order of the size N of the population, $g \sim N$. If the rate of appearance of mutants is r per generation, then rgN mutants appear in g generations. Hence, only when $r \ll N^{-2}$ is the population homogeneous with respect to that character. For larger values of the mutation rate a number of different

haplotypes (or of different surnames) coexist at the statistically stationary state. In the case of exponentially growing populations, the composition of the population crosses over from homogeneity to heterogeneity when the number of individuals becomes large enough, and if growth continues the number of coexisting variants keeps increasing. In the case that will be tackled in this section –the abundance of families of a certain size– the mutation rate is high enough that all the societies studied maintain high degrees of heterogeneity.

The inheritance of surnames or of non-recombining alleles is characterized by three main mechanisms involved in the transmission process from one of the parents to the offspring: (i) the probability that a newborn inherits a certain surname or gene is proportional to the number of individuals in the population bearing it; (ii) the surname (or form of the gene) remains unchanged in most cases, though with a small probability α the surname changes or the gene mutates, and a different group, initially constituted by a single individual, appears; (iii) individuals carrying that surname (or allele) can die at any time with a given probability. Associating an evolution step with the appearance of a newborn in the population, rules (i) and (ii) correspond, respectively, to rules (ii) and (i) in the formulation of Simon's model for Zipf's law in language, as presented in Sect. 2. In addition to mutations, rule (ii) also takes into account migration of individuals to the population. The third rule introduces a new mechanism –mortality– essential to the problem that we are now dealing with: surnames or alleles can disappear whenever they are carried by a single individual, if that individual dies. We call μ the probability that a single individual dies per evolution step. The model described by rules (i), (ii), and (iii) corresponds to an exponentially growing population for any $\mu < 1$. In that scenario, it can be shown that, similarly to the asymptotic behaviour described for Simon's model, the system eventually attains a statistically stationary state where the distribution of family sizes reaches a fixed profile. This distribution will be broad whenever α is large enough.

The analysis of real data for family abundance in different societies reveals remarkable quantitative differences. For example, there are broadly different degrees of heterogeneity regarding surname distribution. The data shown in Fig. 6 imply that there are about 50 different surnames in the USA for each surname in China. Though the transmission process is the same in both cases, each of them should be described by very different values of the relevant parameters. Indeed, actual values of α depend on the accuracy of transmission of surnames and on immigration flows. Changes

of country, of writing system, spelling errors and, in some cases, voluntary changes, together with the appearance of new surnames due to the arrival of foreign families, might translate into very different values for α in different societies. The parameter μ determines the growth rate of the population, and can be highly variable in time. Finally, the distance to the asymptotic form of the distribution depends on the initial condition (number and size of the founding families), and on the genealogical depth of a population, that is, on the time since surnames started to be systematically used as cultural and sociological markers. Thus, real data indicate that countries with different surname distributions differ at least in one of the following conditions: either their values for the parameter α or for the growth rate μ are different, or they are still at the transient phase and have not reached stationarity. This nonetheless, the deep relationship between non-recombining alleles and surname inheritance has made of surname distributions a powerful tool to quantify the genetic heterogeneity of a population, the amount of inbreeding, and the historical degree of mixing in some human communities.[31]

In China, the tradition of using surnames dates back at least to about 2200 B.C. Nowadays, the Chinese society has little diversity regarding surnames, partly due to its genealogical depth, which spans 160 to 200 generations.[32] However, there is probably a second reason explaining why almost 90% of Chinese people share only 100 different surnames: the writing system. Most surnames in China correspond to a well-defined concept, which is represented using a symbol common to most languages and dialects spoken in the country. Mutation thus becomes extremely rare, and the value of α is consequently low, favouring in this way the fixation of a given surname in a large fraction of the population. For example, the surname "King" or "Royal" (often transcribed as Huang), which ranks fifth in abundance, is pronounced Wang2 [a] in Mandarin, Heng in Teochew, and Wong in Cantonese. When people of Chinese origin bearing that surname move to countries using phonetic writing systems, many different transcriptions might arise, such that at present surnames as Huang, Henk, Hank, Wenk or Wank also exist in the USA, though they probably stem from a single original ideogram. Interestingly, the study of large isonymous groups in China [33] demonstrated that the Y chromosome displays multiple haplotypes within that population. This was interpreted as polyphyletism in the surname, meaning that the population under study originated from differ-

[a]The number refers to the tonal form of the word.

ent unrelated founders bearing the same surname. However, an alternative explanation could be that the mutation rate of the Y chromosome is larger than that of surnames, such that changes in the two markers are different enough and the correlation between them decays with time.

Surnames in Europe began to be used in the Middle Ages, meaning that this society has a genealogical depth of 20 to 30 generations. Taking into account the writing system, the values of α are predictably much higher than in the Chinese case. Indeed, there are many surnames that differ just in one or two characters (changes in one letter, including insertions and deletions), and some of them constitute closely related groups. For example, the surnames Kemmingway and Hemaway can be "linked" through a chain of surnames, all of them in use nowadays, that differ in just one character: Hemmingway, Hemingway, Heminway, Hemenway, and Hemanway.[34] While some centuries ago the European population experienced a fast growth (implying a low value of μ), at present it has reached a close-to-stationary value, such that $\mu \simeq 1$. Changes in growth rates, and in particular the limit case $\mu = 1$, can cause qualitative changes in the expected distribution of surname abundance, as shown below. An interesting case in Europe is that of Sweden. Prior to 1862 it was not permitted that common people retained family names, such that the surname changed at each generation, and the old family name disappeared.[35] Moreover, the way of construction of most surnames added the suffix "son" ("daughter") to the given name of a boy's (girl's) father (mother). Due to this procedure, Swedish surnames are highly polyphyletic. Hence, the use of family names as genetic markers is those populations is not feasible.

Japan has a genealogical depth comparable to that of Sweden, since surnames have been systematically used only during the last 120 years.[36] Though the mutation rate in the Japanese system is probably quantitatively similar to the Chinese case –at least as far as the writing system is concerned– its youth still maintains a relatively high diversity at present. Another interesting case is that of American countries which grew fast in population and whose founders were a mixture of European immigrants. Such is the case of Argentina [37] and the USA, where the actual distribution of surnames had as initial condition a relatively large population with high heterogeneity and a few individuals per surname.

Figure 6 shows rank plots for surname abundance in two of the cases discussed. The influence of the genealogical depth, and the low value of α in the Chinese case are particularly visible. Summarizing, we can conclude that different historical contexts, the time at which surnames appeared, and the

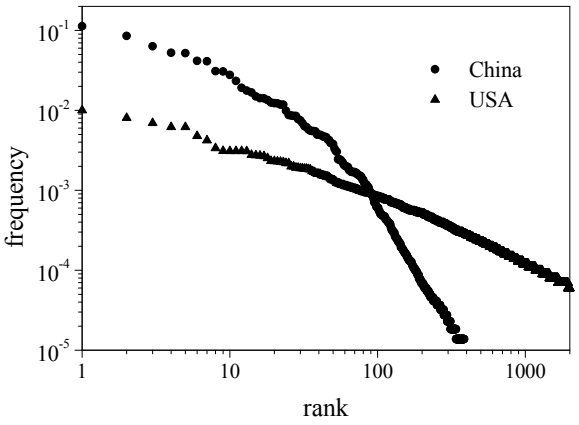

Fig. 6. Zipf's rank plots for surname abundance in some representative societies. Data are from http://technology.chtsai.org/namefreq/ (China), http://www.census.gov/ (USA). While the three most common Chinese surnames (Li, Wang, and Zhang) are borne by almost 10% of the population each, the most common surname in the USA (Smith) is borne by only 1% of the population.

accuracy to which they are transmitted from generation to generation are three factors reflected on the shape of the surname abundance distribution at present.

4.1. *The effects of mortality*

The introduction of the parameter μ in Simon's model is necessary in order to consider the death of individuals in the population, which is the only mechanism leading to the eventual disappearance of surnames. In addition, mortality has immediate consequences in other quantities describing population dynamics. First, the average growth of the population is exponential in time for $\mu < 1$,

$$N(t) = N_0 \exp[\nu(1-\mu)t], \tag{10}$$

with ν standing for the birth rate per individual and unit time, and the product $\mu\nu$ yielding the corresponding death rate.[b] The quantity N_0 is the size of the initial population. In principle, the N_0 initial individuals can be

[b]The relation between the step variable s, which gives the total number of individuals added to the population, and the real time t comes from noticing that the birth frequency is proportional to the total population, such that the elementary increment in time δt is inversely proportional to $N(t)$, $\delta t(s) = (\nu N(s))^{-1}$. The frequency ν fixes time units.

distributed among a number of families of different sizes. The initial condition becomes fully specified once the number of surnames initially borne by exactly n individuals, $P(n,0)$, is known. Polyphyletism corresponds to a situation where $P(n,0) \geq 1$ for at least one value of $n > 1$. The opposite case, where $P(n,0) = 0$ for all $n > 1$, is to be associated with monophyletism. Note that $N_0 = \sum_n nP(n,0)$.

The second consequence of mortality is that individuals have a life expectancy $1/\nu\mu$. During their lifetime, the probability to have m children who inherit their parent's surname turns out to be an exponential distribution of the form

$$p(m) = (1-\alpha)\mu[1+(1-\alpha)\mu]^{-m-1}. \tag{11}$$

Though it is usually assumed that the distribution of offspring is Poisson-like, data collected over short periods of time yield distribution of offspring close to exponential,[38] thus supporting the use of this model at least in appropriate social contexts.

The third consequence of mortality is that the total number of different surnames in a population might decrease. This situation holds, for instance, when the diversity is high and μ changes from small values to values close to one. This represents a situation where the exponential growth stops and the size of a population keeps approximately constant. This is frequent in developed societies, as in Europe nowadays, where the fast growth experienced in the last two centuries has come to a halt.

For $\mu = 0$ the dynamical equations describing the process are (5) and (6), which are completed with an initial condition specifying in this case number and size of the founding families. When mortality is turned on, the update of the population has to be modified in order to include death events. To this end, it is useful to split the dynamics into two sub-steps, as follows. Equations (5) and (6) are used to yield intermediate values $P'(1,s+1)$ and $P'(n,s+1)$, and the total population becomes $N'(s+1) = N(s)+1$ at the first sub-step. The effect of mortality can be accounted for immediately after growth and mutation are applied, such that the final value for the total population once the update is completed reads

$$N(s+1) = N'(s+1) - w(s), \tag{12}$$

with $w(s)$ representing a stochastic dichotomic process that takes the value 1 with probability μ and 0 with probability $1-\mu$. The corresponding evo-

lution equation for the abundance of families of size n is

$$P(n,s+1) = P'(n,s+1) + \overline{\left[\frac{w(s)}{N'(s+1)}\right]}\left[(n+1)P'(n+1,s+1) - nP'(n,s+1)\right], \tag{13}$$

where the bar indicates average over different realizations of the stochastic process. This dynamical equation cannot be solved exactly, though some reasonable assumptions make it possible to obtain approximate solutions. Assuming that the solution varies slowly with n and s, a continuous approximation becomes feasible, where the family size n and the step index s are replaced by continuous variables y and z, respectively.[37]

A relevant problem when analyzing real data for surname abundance is the typical time required to develop the asymptotic form of the solution in a reasonable range of family sizes, and starting with arbitrary initial conditions.[39] Considering that the use of surnames is relatively recent in history, it is important to estimate whether present day societies would be close enough to the asymptotic regime, and thus whether the model can be applied to real situations. A quantitative answer to this question can be obtained by solving the model for surname dynamics using a first-order expansion in the continuous variables y and z. In this approximation, the solution consists of two parts. For $y < y_D(z)$,

$$P(y,z) = \alpha \frac{N_0 + (1-\mu)z}{1-\alpha-\mu} y^{-\zeta} \tag{14}$$

with

$$\zeta = 1 + \frac{1-\mu}{1-\alpha-\mu}. \tag{15}$$

For $y > y_D(z)$,

$$P(y,z) = y_D^{-1} P(y/y_D(z), 0). \tag{16}$$

The family size $y_D(z)$ that separates the two parts of the solution grows as time elapses,

$$y_D(z) = \left(1 + \frac{1-\mu}{N_0} z\right)^{1/(\zeta-1)}, \tag{17}$$

and is directly related to the genealogical depth of the population. As a function of real time, $y_D(t) = \exp[\nu(1-\alpha-\mu)t]$. This means that the transient time t_0 needed to observe the asymptotic regime (dominated by a power-law with exponent ζ) in the family size distribution is logarithmic

in the family size, $t_0 \propto \ln y_0$. This explains why many real distributions of surname abundance are well described by the asymptotic solution in a broad range of values, even if the genealogical depth of most systems seems relatively small.

A more accurate solution to the problem with mortality is obtained by using a second-order expansion of Eq. (13). It reads

$$P(y,z) = \frac{\alpha N(z)}{1-\alpha-\mu}\left(2\frac{1-\alpha-\mu}{1-\alpha+\mu}\right)^{\zeta-1} y^{-1} U\left(\zeta-1,0,2\frac{1-\alpha-\mu}{1-\alpha+\mu}y\right), \tag{18}$$

where $U(a,b,x)$ is the logarithmic Kummer's function.[40] For large family sizes, $y \to \infty$, this solution again predicts a power-law behavior of the form $n(y,z) \propto y^{-\zeta}$. The exponent ζ, defined in Eq. (15), presents two relevant limits. First, for $\mu = 0$ the known solution for Simon's model, Eq. (7), is recovered. Second, the limit $\alpha \to 0$ always converges to $\zeta = 2$, irrespectively of the value of μ. For small family sizes, Eq. (18) yields a probability lower than in the case $\mu = 0$. This downward bending of the distribution of surname abundance at small sizes is in agreement with field data. Figure 7 represents several sets of data and the corresponding fits obtained from Eq. (18).

A similar continuous approximation to calculate frequency distributions in processes with birth, death, and mutation, yields a solution for this problem equivalent to Eq. (18).[41] When that solution was used to fit the distribution of surnames in several European countries and in the USA, a good agreement between data and theoretical prediction was obtained. This reinforces the idea that the genealogical depth of those relatively young systems suffices to be close enough to the asymptotic, power-law regime.

The case $\mu = 1$ deserves some separate comments, since in this limit the qualitative properties of the system change. This situation corresponds to populations that are stationary in size $N(s) = N_0$, where the number of births equals the number of deaths. This model was used in the context of genetic inheritance to study the probability of fixation of alleles:[42] Moran's model is analogous to Simon's model in populations of constant size. Eventually, the diversity supported by a population of constant size will reach a constant value, though the transient until this regime sets in depends, as it does for $\mu < 1$, on the initial condition. Further, it turns out that, for constant populations, the functional form of the surname abundance distribution changes with the actual values of the parameters: the solution to the dynamical equations depends on how the product αN_0 compares with

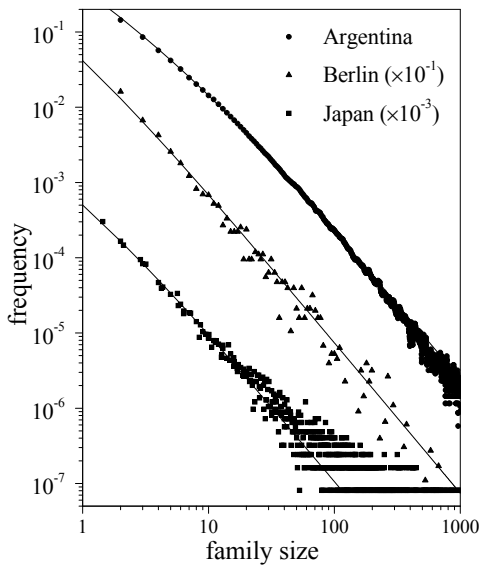

Fig. 7. Frequency of appearance of families with a given size. Data for Argentina correspond to almost 350,000 surnames in the whole 1996 Argentinian telephone book; for Berlin, 6400 surnames beginning by A in the 1996 telephone book have been used; data for Japan are adapted from Miyazima *et al.* (2000).

unity. If mutation is frequent enough such that $\alpha N_0 > 1$, the asymptotic distribution of family sizes is exponential,

$$P(n) \simeq \frac{\alpha N_0}{n}(1-\alpha)^{n-1} \quad \text{for} \quad \alpha N_0 > 1, \tag{19}$$

and the stationary number S of different surnames is

$$S \simeq \frac{\alpha N_0}{1-\alpha}|\ln \alpha|. \tag{20}$$

If, on the other hand, mutation is rare enough to yield $\alpha N_0 < 1$, the distribution behaves as a power-law,

$$P(n) \simeq n^{-1} \quad \text{for} \quad \alpha N_0 < 1. \tag{21}$$

In those cases where mutation is rare enough, in the limit $\alpha \to 0$, the population becomes homogeneous (there is a single family, $S = 1$) and the distribution consists of a single peak at $n = N_0$.

This could in principle be the fate of conservative societies where inheritance is very accurate and the appearance of new surnames is strongly suppressed. However, the limit situation where surname diversity disappears lacks any cultural meaning, since the value that an individual assigns

to his family name progressively fades out as the society becomes more homogeneous.

4.2. *The distribution of given names*

A person's full name identifies the individual and is frequently carried with pride. The low variability of surnames in certain societies can be balanced by a higher diversity in given names, such that the number of full names in use is large enough to be rarely repeated within a population. We conclude this section with a brief review of the distribution of given names.

One of the consequences of the very low surname diversity in the Chinese society may be that the family name is no longer a strong sign of individuality, but of a very large community of individuals among which close contacts do not always exist. This is probably one of the reasons that Chinese given names are extremely diverse and often complex in meaning: they add singularity to the individual and help distinguishing him within a large population isonymous with respect to the surname. The distribution of given names in different cultures seems to bear an inverse relationship with the distribution of family names. With the evidence at hand, one could argue that the full name arises from a compromise between "being different" and "belonging to a community."

Figure 8 represents Zipf's rank plots for given names abundance in China and USA. Those data correspond exactly to the same samples represented in Fig. 6, there ranked by surname abundance. In these two representative cases, it is interesting to note that the combinatorial variability of full names, defined as the product between the number of different surnames and the number of different given names, return similar quantities. In China, the number of surnames in use is of order 10^2, while the amount of different given names rises to 10^5; in the USA, 10^3 different surnames can combine with 10^4 different given names. Hence, in both societies the number of different full names is of order 10^7.

In societies where many common surnames occur, and where given names are also subject to tradition –such that their variability is lower than, for instance, in the Chinese case– it seems that other cultural mechanisms might act in order to increase the singularity of the full name for each individual. Such mechanisms could be the use of middle names, or the inclusion of the mother's surname after the father's one, as is done in Spain and several Latin American countries.

Finally, let us remark the qualitative similarity between the distribu-

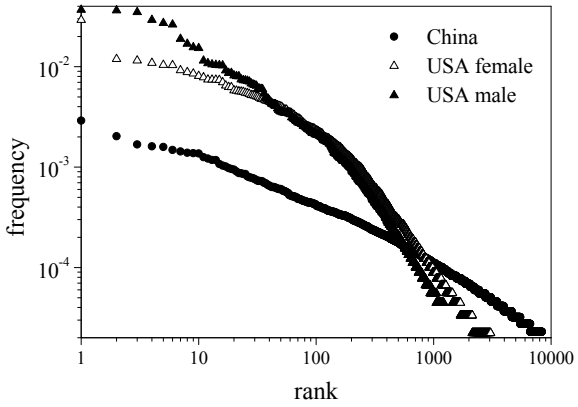

Fig. 8. Zipf's rank plots for given name abundance in two societies. Data are from http://technology.chtsai.org/namefreq/ (China) and http://www.census.gov/ (USA). Compare these distributions with those of surnames in the same populations (Fig. 6). Most common given names in USA (Mary and James rank 1, Patricia and John rank 2 for females and males, respectively) are carried by 2-3% of the population. In China, the most common given name is only shared by three people out of a thousand.

tions shown in Fig. 8 and those corresponding to surname abundance. Although the dynamics followed by the abundance in time of a given name does not precisely conform to the inheritance model followed by surnames, the distribution has characteristics that point to a broader applicability of multiplicative models in sociology. We believe that the main mechanisms shaping the distribution of given names might follow dynamics closely related to those of fashion, which, in a broad sense, underlies many of our daily habits and preferences.

5. Conclusion

The dynamics of several of the cultural features discussed in this review are clearly dominated by a hereditary component. Languages and surnames are mostly passed unchanged from one generation to the next, such that their transmission is in the vertical direction. This fully justifies the use of stochastic multiplicative models to analyze their statistical properties. It could be argued that other systems, as cities, are not so clearly described by a multiplicative model, though it is reasonable to assume that city growth is dominated by reproduction of its inhabitants and the arrival of new individuals, this last process having a strong multiplicative component as well. The situation is less clear for the last example discussed: the distribution

of given names.

Cultural features are often determined by the sociological pressure exerted by groups of akins. The hobbies, religious beliefs, TV programs watched, or books read by an individual, are not independent of the majority preferences within his or her social group. It is arguable that, the larger the group sharing a given characteristic, the higher the probability that a new individual acquires that characteristic. This dynamics is intrinsically multiplicative, and though the form of transmission of the considered feature is horizontal in this framework –thus not inherited from one generation to the next– it suggests that coarse-grained multiplicative models where the relevant variable is the size of groups might be of general application in sociological problems. This calls for extensions of the models discussed in this contribution, for instance by adding horizontal flows between groups proportional to their sizes, superimposed to pure vertical transmission. Other modifications might include size-dependent growth rates, for instance in the form of higher-order terms in the dynamical equations. The splitting of very large groups or the merging of small ones, as often observed in real societies, would be worth considering as well.

The quantitative analysis of cultural evolution through phylogenetic methods is an increasingly used approach in the sociological community. Vertical transmission of cultural characters, including in particular languages, seems to be much more determinant in shaping the evolution and distribution of cultural groups than horizontal transmission. Nonetheless, this is a changing paradigm since, until the second half of the twentieth century, blending processes were considered as the main mechanism conforming cultural history.[43] If inheritance in its broader sense (that is, growth proportional to the group size) is indeed the dominant form of transmission of cultural traits, then models similar to Simon's offer a promising way of explaining the statistical abundance and evolution of a large number of cultural features.

Acknowledgments

Discussions with Marcelo Montemurro and Chema Ruiz are gratefully acknowledged. SCM benefits from a Ramón y Cajal contract of MEC (Spain).

References

1. D. Sornette, *Phys. Rev. E* **57**, 4811 (1998).
2. D. Sornette, *Critical Phenomena in Natural Sciences. Chaos, Fractals, Self-organization and Disorder: Concepts and Tools* (Springer, Berlin, 2000).

3. S. C. Manrubia and D. H. Zanette, *Phys. Rev. E* **59**, 4945 (1999).
4. G. K. Zipf, *Human Behaviour and the Principle of Least-Effort* (Addison-Wesley, Cambridge, 1949).
5. G. K. Zipf, *The Psycho-Biology of Language* (Houghton-Mifflin, Boston, 1935).
6. H. A. Simon, *Biometrika* **42**, 425 (1955).
7. J. Willis and G. Yule, *Nature* **109**, 177 (1922).
8. F. H. van Eemeren, *Crucial Concepts in Argumentation Theory* (University of Chicago Press, Chicago, 2001).
9. C. M. Brown and P. Hagoort, *The Neurocognition of Language* (Oxford University Press, Oxford, 2000).
10. B. B. Mandelbrot, in *Communication Theory*, Ed. W. Jackson (Butterworth, London, 1953), p. 486.
11. B. B. Mandelbrot, *Inform. Control* **2**, 90 (1959); **4**, 198 (1961); **4**, 300 (1961).
12. H. A. Simon, *Inform. Control* **3**, 80 (1960); **4**, 217 (1961); **4**, 305 (1961).
13. W. Li, *IEEE Trans. Inf. Theory* **38**, 1842 (1992).
14. M. A. Montemurro and D. H. Zanette, *Glottometrics* **4**, 86 (2002).
15. D. H. Zanette and M. A. Montemurro, *J. Quant. Linguistics* **12**, 29 (2005).
16. A. D. Patel, *Nature Neurosci.* **6**, 674 (2003).
17. M. G. Boroda and A. A. Polikarpov, *Musikometrika* **1**, 127 (1988).
18. B. Manaris, D. Vaughan, C. Wagner, J. Romero, and R. B. Davis, in *Lecture Notes in Computer Science: Applications of Evolutionary Computing*, Eds. F. Rothlauf *et al.* (Springer, Berlin, 2003), p. 522.
19. D. H. Zanette, *Musicae Scientiae* **10**, 3 (2006).
20. J. Portugali, *Self-Organization and the City* (Springer, Berlin, 2000).
21. M. Ruhlen, *A Guide to the World's Languages* (Stanford University Press, Stanford, 1991).
22. M.A.F. Gomes, G.L. Vasconcelos, I.J. Tsang, and I.R. Tsang, *Physica A* **271**, 489 (1999).
23. D.M. Abrams and S.H. Strogatz, *Nature* **424**, 900 (2003).
24. H. W. Watson and F. Galton, *J. Roy. Anthropol. Inst.* **4**, 138 (1874).
25. R. A. Fisher, *Proc. Roy. Soc. Edin.* **42**, 321 (1922).
26. B. Sykes and C. Irven, *Am. J. Hum. Genet.* **66**, 1417 (2000).
27. T. E. Harris, *The theory of branching processes.* (Springer-Verlag, Berlin, 1963).
28. N. Yasuda, L. L. Cavalli-Sforza, M. Skolnick, and M. Moroni, *Theor. Pop. Biol.* **5**, 123 (1974).
29. M. Kimura and J. F. Crow, *Genetics* **49**, 725 (1964).
30. M. Kimura and T. Ohta, *Theoretical Aspects of Population Genetics.* (Princeton University Press, 1971).
31. S. E. Colantonio, G. W. Lasker, B. A. Kaplan, and V. Fuster, *Hum. Biol.* **75**, 785 (2004).
32. Y. D. Yuan, C. Zhang, Q. Y. Ma, and H. M. Yang, *Yi Chuan Xue Bao* **27**, 471 (2000).
33. L. Jin, B. Su, J. Xiao, *et al.*, *Am. J. Hum. Gen.* **65**, 1136 (1999).
34. S. C. Manrubia, B. Derrida, and D. H. Zanette, *Am. Sci.* **91**, 158 (2003).

35. http://www.newulmtel.net/ jatakuck/lower/AckDoc1.html
36. S. Miyazima, Y. Lee, T. Nagamine, and H. Miyajima, *Physica A* **278**, 282 (2000).
37. S. C. Manrubia and D. H. Zanette, *J. theor. Biol.* **216**, 461 (2002).
38. D. M. Hull, *Theor. Popul. Biol.* **54**, 105 (1998).
39. D. H. Zanette and S. C. Manrubia, *Physica A* **295**, 1 (2001).
40. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, (Dover, New York, 1970).
41. D. L. Bartley, T. Ogden, and R. Song, *BioSystems* **66**, 179 (2002).
42. P. A. P. Moran, *The Statistical Processes of Evolutionary Theory.* (Clarendon Press, Oxford, 1962).
43. R. Mace and C. J. Holden, *Trends in Ecol. Evol.* **20**, 116 (2005).