

# Drift as a mechanism for cultural change: an example from baby names

Matthew W. Hahn<sup>1</sup> and R. Alexander Bentley<sup>2\*</sup>

<sup>1</sup>Department of Biology, Duke University, Box 90338, Durham, NC 27708, USA

<sup>2</sup>AHRB Centre for the Evolutionary Analysis of Cultural Behaviour, University College London, 31–34 Gordon Square, London WC1H 0PY, UK

\*Author for correspondence (r.bentley@ucl.ac.uk).

Recd 01.04.03; Accptd 08.05.03; Online 13.06.03

**In the social sciences, there is currently no consensus on the mechanism by which cultural elements come and go in human society. For elements that are value-neutral, an appropriate null model may be one of random copying between individuals in the population. We show that the frequency distributions of baby names used in the United States in each decade of the twentieth century, for both males and females, obey a power law that is maintained over 100 years even though the population is growing, names are being introduced and lost every decade and large changes in the frequencies of specific names are common. We show that these distributions are satisfactorily explained by a simple process in which individuals randomly copy names from each other, a process that is analogous to the infinite-allele model of population genetics with random genetic drift. By its simplicity, this model provides a powerful null hypothesis for cultural change. It further explains why a few elements inevitably become highly popular, even if they have no intrinsic superiority over alternatives. Random copying could potentially explain power law distributions in other cultural realms, including the links on the World Wide Web.**

**Keywords:** cultural transmission; cultural evolution; random genetic drift; power laws

## 1. INTRODUCTION

Evolution is the process by which the frequencies of genetic variants in a population change over time. This process can be deterministic, whereby natural selection favours specific genetic variants, or it can be random, where change occurs through genetic drift. Random drift, also known as ‘unbiased transmission’ when applied to cultural change (Boyd & Richerson 1985), has been proposed as an appropriate null model for the way in which functionally neutral cultural elements come and go in society (Cavalli-Sforza & Feldman 1981; Neiman 1995). While many studies have described the transition from one dominant variant (norm) to another (e.g. Rogers 1995; Henrich 2001), few studies have compared change over time among multiple cultural variants coexisting within a population with a neutral model of random drift (but see Cavalli-Sforza & Feldman 1981; Neiman 1995; Bentley & Shennan 2003).

An example of a cultural trait potentially influenced by drift is the name given to newborn babies. Most parents choose a pre-existing name for their child, while others invent a new name. Over time, some names become more prevalent, while others decline in frequency and may even be lost from a population. Assuming that no name is intrinsically more valuable than another, and envisioning the copying of names as ‘replication’ and the invention of new names as ‘mutation’, this process is analogous to the population genetic mechanism of random genetic drift (Wright 1931; Crow & Kimura 1970).

We show that, for each decade of the twentieth century, the frequency distributions of male and female baby names used in the United States obey power laws with slopes that are consistent through the century. Fat-tailed distributions, including power laws, characterize a wide range of phenomena, including the much-discussed World Wide Web network (e.g. Albert & Barabasi 2002). Many different mechanisms can generate a power law distribution, including random multiplicative processes (e.g. Simon 1955; Laherrère & Sornette 1998) and critical phenomena (Jensen 1998). We follow others (Cavalli-Sforza & Feldman 1981; Neiman 1995) in using the model of drift in a finite population, an example of a random multiplicative process, as a null model against which to compare empirical data on cultural variants.

## 2. MATERIAL AND METHODS

### (a) Data collection

Frequency data for the 1000 most commonly used baby names in the United States in each decade of the twentieth century were collected from the US Social Security Administration (<http://www.ssa.gov/OACT/babynames>) on 20 July 2002. These data come from a sample of 5% of all social security cards issued to individuals who were born during each decade in the United States. New names were discovered by comparing each decade with the previous decade; the first decade of the century was considered to have no new names.

### (b) Simulation

In what amounts to a single-locus, multiple-allele model simulated in MATLAB (The MathWorks, Inc.), we arbitrarily assigned numerical names to a population of  $N$  individuals, which were then subject to repeated mutation and sampling; the appearance of every name in the population was followed cumulatively throughout the run. Each individual in the starting population had a unique name. In every time-step,  $N$  new individuals ‘chose’ a name randomly with replacement from the previous population. With some probability,  $\mu$ , an individual received a novel name (‘mutation’). All runs lasted for 1000 time-steps, keeping the value of  $\theta$  ( $= 4N\mu$ ) constant. There was no effect on variant frequency distributions seen for runs above 500 steps and therefore it is assumed that the simulations reached a steady state.

## 3. RESULTS AND DISCUSSION

For each decade of the twentieth century, the frequency distribution for the 1000 most commonly used baby names in the United States (both male and female) fits a power law with an  $r^2$  value above 0.97 (figure 1; table 1). This distribution shows that there are a very few names that are highly popular (in frequencies approaching 10%), whereas there are many names present at very low frequencies (at or below 0.01%). Although a log-normal function with a small mean and large variance can exhibit a fat tail similar to a power law, a power law most accurately describes the distributions in figure 1, which, unlike the log-normal, are monotonically decreasing. In referring to these distributions as power laws, we stress that our main objective is to model a simple random drift process yielding a fat-tailed distribution that matches the data closely.

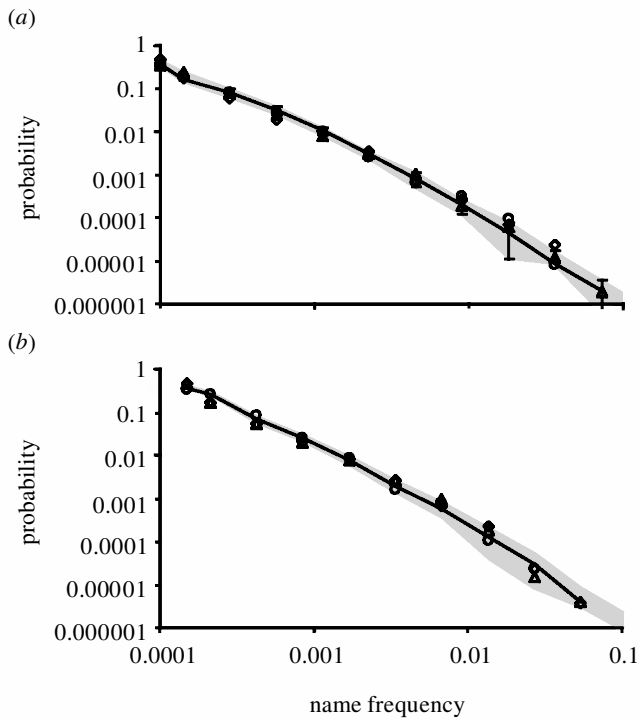


Figure 1. Power law distributions of baby names. Frequencies of the top 1000 (a) male and (b) female baby names in the United States, for three representative decades during the twentieth century. The  $x$ -axis represents the frequency of a name in the total sample of individuals, and the  $y$ -axis represents the probability that a certain name would fall within the bin at that frequency (proportional to the fraction of names in the top 1000). As common for such log-log plots (e.g. Jensen 1998; Zanette & Manrubia 2001; Albert & Barabasi 2002), the bin sizes increase in powers of 2 (0.0001–0.0002, 0.0002–0.0004, 0.0004–0.0008,...), data are plotted at the middle of each bin and probabilities are normalized for the increasing bin sizes. Also shown are the mean (solid line) and 95% confidence intervals (grey ribbon) resulting from 20 runs of the neutral-trait model with  $\theta = 4N\mu = 4$ . A regression between  $\log(\text{average model value})$  and  $\log(\text{average data value})$  yields  $r^2 = 0.993$  for male names and  $r^2 = 0.994$  for female names. Key: triangles, 1900–1909; diamonds in (a), 1950–1959; diamonds in (b), 1940–1949; circles, 1980–1989; solid line,  $\theta = 4$  model.

Figure 1 shows that these power law distributions extend over almost three orders of magnitude in terms of the frequency of names out of the total population (from 0.01% to almost 10%) and seven orders of magnitude in terms of the probability of a name having a certain frequency (proportional to the fraction of names in the top 1000). The power laws for both male and female names have persisted despite large population growth, driven by birth and immigration and changes in the ethnic make-up of the population, as well as gain, loss and changes in the frequency of specific names.

To explain the stable distribution of name frequencies despite the change in baby name usage over time, we suggest that baby names are value-neutral cultural traits chosen proportionally from the population of existing names, created by ‘mutation’ and lost through sampling. First names are good candidates for neutral cultural traits governed by drift because they are chosen independently at each birth, not transmitted from parents to children as

surnames are (Zanette & Manrubia 2001) and because they have no difference in functional value.

In population genetics, it has been shown analytically (Kimura & Crow 1964) that the number of neutral alleles (variants) with frequency  $x$  for a single moment at equilibrium is given by  $\theta x^{-1}(1-x)^{\theta-1}$ , where  $\theta = 4N\mu$ , as mentioned above. Unfortunately, the data available to us do not represent an instantaneous distribution, as they are the cumulative number of baby names in a decade. We therefore modelled the random copying of names by computer simulation, keeping track of the cumulative number of times that each name appeared during the simulation (see § 2 for details). Each time-step in the simulation represents a set of  $N$  newly born babies, each of whom is named by copying the name of a randomly chosen baby (with replacement) from the previous time-step. In addition, in each time-step a small fraction,  $\mu$ , of the  $N$  individuals receive a unique name. This simple model is equivalent to the infinite-allele model of population genetics for a single-locus, multiple neutral-allele system (Kimura & Crow 1964; Crow & Kimura 1970). The invention of a new baby name is analogous to mutation, and copying the name of a randomly chosen individual is analogous to sampling.

We ran the simulations across a range of values of  $\theta$  (from  $\theta = 1$  to  $\theta = 7$ ) and found that the power law slope,  $\alpha$ , increases as  $\theta$  increases. The value of  $\theta$  that is equal to  $4N\mu$  is the relevant parameter affecting the shape of the frequency spectrum (Kimura & Crow 1964; Ewens 1972). As long as the product of  $N$  and  $\mu$  remains constant, the same variant distribution will result. Simulations with  $\theta = 4$  fit the data best, with each simulation result fitting a power law distribution with  $r^2$  above 0.97. Figure 1 shows the mean and 95% confidence interval for 20 simulation runs, each for 1000 time-steps with  $\theta = 4$ , plotted together with the name data. Regressions between the model average (solid line in figure 1) and the data yield  $r^2$  above 0.993 for both males and females.

By analogy with the neutral-trait model of population genetics, our random copying model implies that the total number of names maintained in the population is proportional to both the population size and the mutation rate to new names. As the population size gets bigger, more names are found in the population. Likewise, a higher mutation rate should also result in more names. As the population of the United States grew from 76 million in 1900 to 281 million in 2000 (table 1), the percentage of all names represented by the top 1000 dropped from 91% to 75% for females and from 91% to 86% for males. This suggests that the total number of names in the entire United States population increased over time, but at a higher rate for females than for males. We believe that this difference reflects different rates of ‘mutation’ (the probability of giving a child a novel name) between males and females.

As a measure of the mutation rate, we counted the number of novel names appearing each decade (figure 2). The number of new names (controlled for the number of births) is always greater for females than for males (paired  $t$ -test,  $t = 6.59$ ,  $p = 0.0002$ ). In addition, the two mutation rates parallel each other’s fluctuations, which may reflect changing cultural norms or immigration rates affecting the introduction of new names of both sexes. When controlled

Table 1. United States population details, together with the exponent  $\alpha$  and  $r^2$  value for a power law fit to the frequency distribution of the top 1000 baby names.

(Population sizes at the beginning of each decade are from the US Census Bureau at [www.census.gov/prod/2002pubs/01statab/pop.pdf](http://www.census.gov/prod/2002pubs/01statab/pop.pdf). Births per decade for each sex were calculated by multiplying the number of births in the Social Security Administration 5% sample by 20.  $\alpha$  is the exponent in the formula for a power law:  $p = C/\nu^\alpha$ , describing the probability  $p$  that a name is observed with frequency  $\nu$  in the population.)

years	population size (million)	sex	number of births (million)	$\alpha$	$r^2$
1900–1909	76	male	9.6	1.79	0.991
		female	9.6	1.88	0.979
1910–1919	92	male	12.2	1.79	0.992
		female	12.2	1.86	0.985
1920–1929	106	male	13.3	1.69	0.997
		female	13.3	1.86	0.984
1930–1939	123	male	11.9	1.77	0.990
		female	11.6	1.86	0.983
1940–1949	132	male	16.2	1.64	0.995
		female	15.3	1.86	0.983
1950–1959	151	male	21.2	1.63	0.996
		female	20.0	1.71	0.993
1960–1969	179	male	20.2	1.64	0.996
		female	19.4	1.88	0.963
1970–1979	203	male	17.5	1.70	0.994
		female	16.8	1.93	0.970
1980–1989	227	male	19.5	1.71	0.987
		female	18.7	1.84	0.990
1990–1999	249	male	18.3	1.61	0.993
		female	17.5	1.75	0.986

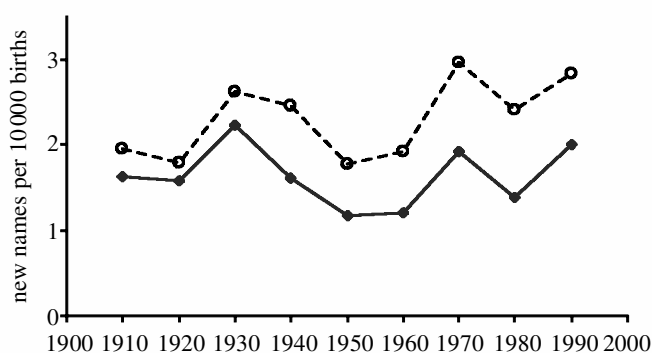


Figure 2. The number of new baby names, per decade, to enter the top 1000 list for both males (filled diamonds) and females (open circles) as a function of the number of male and female births.

for the number births, the mutation rate of female names averages to 2.3 new names per 10 000 births and the mutation rate of male names averages to 1.6 new names per 10 000 births (number of births each decade taken separately for males and females from table 1). This means that girls are, on average, 1.4 times more likely to receive a novel first name than boys, most probably owing to naming customs in a predominantly patriarchal society, and also to the fact that only *ca.* 6% of all the names in Judeo-Christian scriptures are female (Meyers 2000). In any case, a higher mutation rate, producing more novel female than male names in the population, may explain the consistent difference in slopes between male and female name frequency distributions (table 1).

The neutral model also predicts that the variance in variant frequencies increases with age (Wright 1931; Crow & Kimura 1970). Throughout the century, both

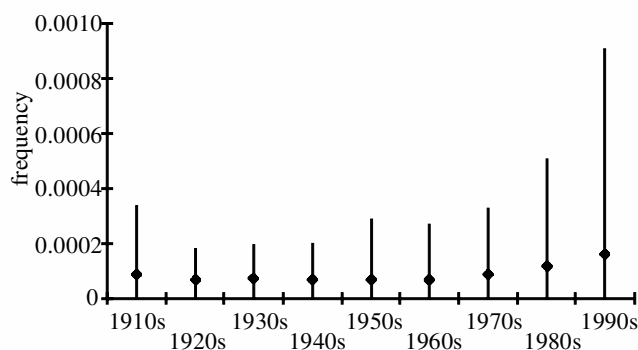


Figure 3. Frequency and variance over time of 99 male names that were new to the top 1000 in the 1910s. Filled diamonds, mean frequency; lines, one standard deviation.

male and female names have had large increases or decreases in frequency. For example, the names Tyler and Ashley, ranked number 11 for boys and 1 for girls, respectively, in the 1990s, were not on the list before the 1950s (we note that names not on the list are undetectable, but not necessarily absent: they may exist at extremely low frequencies or in literature). By contrast, the names Clarence and Mildred, numbers 20 and 9 respectively in the 1900s, are number 491 and not found in the 1990s. As expected (Wright 1931), the variance in the frequency of new names grows over time even as the mean frequency stays the same (figure 3). As new names enter the population, random drift causes changes in frequencies as these names are repeatedly sampled; some are lost and some drift to higher frequencies.

The distribution of baby names in the United States and their turnover in the twentieth century can be explained by

a completely value-neutral process: no preference, fitness, or selection is needed, only proportional sampling. The congruence of many predictions of the neutral model with the observed data leads us to favour it over alternatives such as frequency-dependent or balancing selection. For neutral objects of human choice, the random-drift model provides theoretical predictions that include the frequency distribution, number of variants expected and relationship between frequency and age (Crow & Kimura 1970; Cavalli-Sforza & Feldman 1981; Hartl & Clark 1997). The multiplicative nature of repeated random copying may explain why a few cultural elements inevitably become highly popular, without necessarily being intrinsically *better*. Certainly, not all cultural and economic elements are value-neutral; in these cases, random drift serves as a null model to test against (Boyd & Richerson 1985; Neiman 1995).

#### Acknowledgements

The authors thank A. Lakhina and P. E. Fernandez for computational assistance and comments. They also thank M. Hauber, S. Miller, L. Moyle, M. Rausher, J. Shapiro, and two anonymous reviewers for helpful suggestions. We also thank the Central European University and the Santa Fe Institute for initiating this collaboration.

Albert, R. & Barabasi, A. L. 2002 Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97.

- Bentley, R. A. & Shennan, S. J. 2003 Cultural transmission and stochastic network growth. *Am. Antiq.* **68**. (In the press.)
- Boyd, R. & Richerson, P. J. 1985 *Culture and the evolutionary process*. University of Chicago Press.
- Cavalli-Sforza, L. L. & Feldman, M. W. 1981 *Cultural transmission and evolution*. Princeton University Press.
- Crow, J. F. & Kimura, M. 1970 *An introduction to population genetics theory*. New York: Harper Row.
- Ewens, W. J. 1972 The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3**, 87–112.
- Hartl, D. L. & Clark, A. G. 1997 *Principles of population genetics*. Sunderland, MA: Sinauer.
- Henrich, J. 2001 Cultural transmission and the diffusion of innovations. *Am. Anthropol.* **103**, 992–1013.
- Jensen, H. J. 1998 *Self-organized criticality: emergent complex behavior in physical and biological systems*. Cambridge University Press.
- Kimura, M. & Crow, J. F. 1964 The number of alleles that can be maintained in a finite population. *Genetics* **49**, 725–738.
- Laherrère, J. & Sornette, D. 1998 Stretched exponential distributions in nature and economy: ‘fat tails’ with characteristic scales. *Eur. Phys. J. B* **2**, 525–539.
- Meyers, C. L. (ed.) 2000 *Women in scripture: a dictionary of named and unnamed women in the Hebrew bible, the Apocryphal/Deuterocanonical books, and the New Testament*. Boston, MA: Houghton Mifflin.
- Neiman, F. D. 1995 Stylistic variation in evolutionary perspective. *Am. Antiq.* **60**, 7–36.
- Rogers, E. 1995 *Diffusion of innovations*. New York: The Free Press.
- Simon, H. A. 1955 On a class of skew distribution functions. *Biometrika* **42**, 425–440.
- Wright, S. 1931 Evolution in mendelian populations. *Genetics* **16**, 97–159.
- Zanette, D. H. & Manrubia, S. C. 2001 Vertical transmission of culture and the distribution of family names. *Physica A* **295**, 1–8.