

# Random drift and culture change

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

<sup>1</sup>AHRB Centre for the Evolutionary Analysis of Cultural Behaviour, University College London, 31–34 Gordon Square, London WC1H 0PY, UK

<sup>2</sup>Centre for Population Biology, 2320 Storer Hall, University of California, Davis, CA 95616, USA

We show that the frequency distributions of cultural variants, in three different real-world examples—first names, archaeological pottery and applications for technology patents—follow power laws that can be explained by a simple model of random drift. We conclude that cultural and economic choices often reflect a decision process that is value-neutral; this result has far-reaching testable implications for social-science research.

**Keywords:** cultural evolution; random genetic drift; patents; first names; pottery; power laws

## 1. INTRODUCTION

Evolution is the process by which the frequencies of variants in a population change over time. This definition applies to both biological and cultural change: just as with genetic variants, the assemblage of cultural variants may change in frequency over time, whether as a result of selection, drift or some other evolutionary mechanism (Cavalli-Sforza & Feldman 1981; Boyd & Richerson 1985; Neiman 1995; Shennan 2002).

In this paper, we discuss tractable examples of cultural change in situations where, rather than one dominant cultural variant giving way to another over time, there are many different choices available at all times. The situation relates to a classic phenomenon of population genetics, called random genetic drift, which describes how the diversity of variants evolve when the dominant process is one of random copying. A model of genetic drift through random copying is referred to as the neutral model (Kimura & Crow 1964; Crow & Kimura 1970) because the variants are considered to be neutral with respect to the success of the individual.

We argue that the neutral model is a powerful null hypothesis for studying cultural evolution because it allows us to take advantage of the mathematical theory and predictions provided by population genetics (Crow & Kimura 1970; Cavalli-Sforza & Feldman 1981; Boyd & Richerson 1985; Hartl & Clark 1997). Certainly there are cultural phenomena that do not fit the neutral model; in these cases it can be used as a null hypothesis to test against (Neiman 1995; Lipo *et al.* 1997; Shennan & Wilkinson 2001; Hahn & Bentley 2003).

As we have suggested before (Bentley & Shennan 2003; Hahn & Bentley 2003), the neutral model may explain a much-researched property that characterizes evolving systems in biology, ecology, economics and human society: the highly skewed ‘power law’ distribution:

$$P(v) = C/v^\alpha, \quad (1.1)$$

where  $\alpha$  and  $C$  are constants, and  $v > 0$ . In this paper, we use  $P$  to represent the proportion of variants observed with

frequency  $v$  in the population. The power law in equation (1.1) appears as a straight line with a slope of  $-\alpha$  on a graph with logarithmic scales on both axes. A power-law distribution of variant frequencies in a population means that there are many uncommon variants and a very few popular variants that are thousands of times more popular than the majority. In this paper, we show how power laws characterize the distribution of cultural-variant frequencies in a model of random copying of cultural variants in three real-world examples, including first names and patents from the twentieth century and archaeological pottery from the sixth millennium BC. By demonstrating the power of the neutral model in explaining the number and distribution of these cultural variants, we conclude that it can be used to study a wide range of cultural phenomena.

## 2. CULTURAL DRIFT AND THE NEUTRAL MODEL

In the neutral model, there are  $N$  individuals, each characterized by a behavioural or stylistic variant, such as a first name or pottery motif (figure 1). At each time step, each of  $N$  new individuals copies its variant from an individual randomly selected from the previous time step. To this very simple process we add innovation (analogous to genetic mutation) as the continuous introduction of new unique variants over time. The variable  $\mu$  represents the mutation rate ( $\mu \ll 1$ ), which in the case of cultural evolution can represent the innovation rate per individual per time step.

Having defined this simple model, we can predict the effect of random drift on the statistics of the variants, simply by knowing the size of a population,  $N$ , and the innovation (mutation) rate,  $\mu$ , or even just their product,  $N\mu$ . For instance, applying the expression given by Gillespie (1998, p. 27) for a haploid population, the probability that two variants drawn randomly from the population will be the same is  $1/(1 + 2N\mu)$ .

In this paper, we explore what can be predicted about the frequencies of the different variants. For a population at equilibrium at a single moment in time, Kimura & Crow (1964) provided an analytical solution for the frequency distribution of the variants, which we express as the proportion,  $P$ , of variants having a frequency  $v$  (expressed as a proportion of all the variant copies in the

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

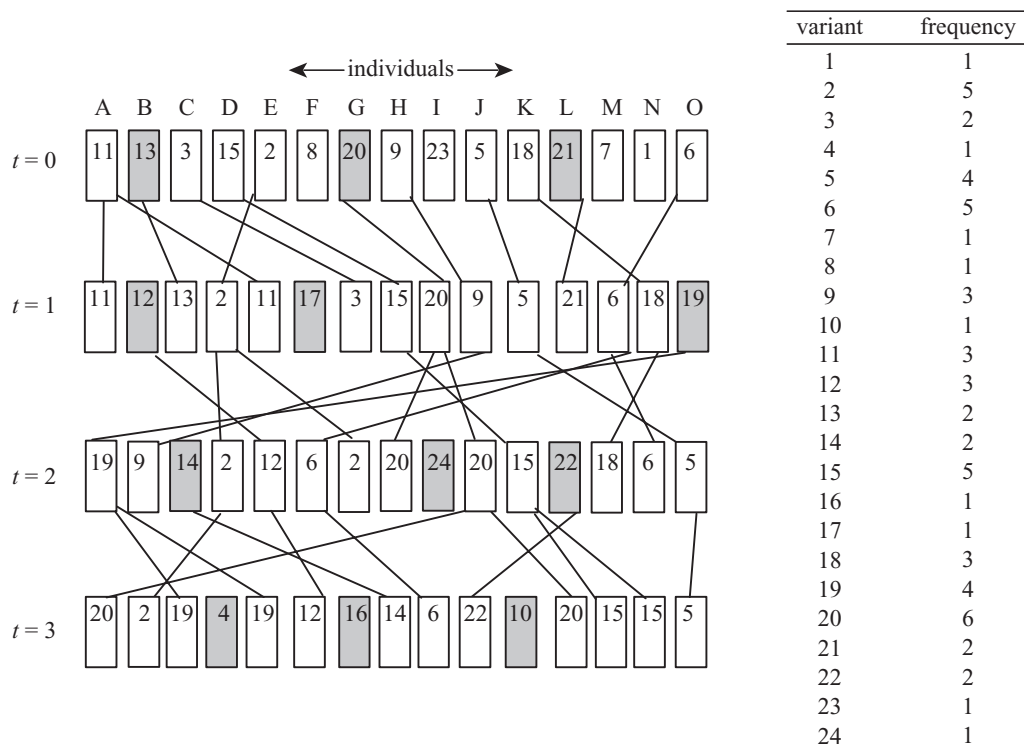


Figure 1. Computer simulation of the neutral model. Each vertical column of squares represents an individual (A, B, ..., O), each row represents a successive time step and different numbers represent different variants. New random numbers represent innovations (in individuals shaded grey). With three mutations per time step in this example,  $\mu = 3/15 = 0.20$ . Each individual that does not mutate copies another individual from the previous time step (black lines). For the frequency distribution, we count all the appearances of each variant over all the time steps, as shown in the columns on the right.

population, the discrete possible values being  $v = 1/N, 2/N, 3/N, \dots, 1$ . This is given by

$$P(v) = 2N\mu^{-1}(1 - v)^{2N\mu-1}. \tag{2.1}$$

Note that the important parameter is the product of  $N$  and  $\mu$ , such that the same equilibrium frequency distribution can result from different population sizes and mutation rates whose products are the same (Kimura & Crow 1964; Ewens 1972). This is somewhat counter-intuitive, but we demonstrate that it is true through computer simulations of the neutral model in § 2a.

Unfortunately, the equilibrium distribution given by Kimura & Crow (1964) is the distribution at a single point in time, whereas archaeologists and anthropologists usually encounter data representing the accumulation of cultural variants over time (e.g. artefacts, yearly record of applications for patents), and this cannot be compared with the instantaneous distribution predicted by Kimura & Crow (1964). Instead, because we are not aware of an analytical solution for the neutral model in terms of the frequency distribution of variants accumulated over time, we use computer simulations to explore the results.

**(a) Computer simulation of the neutral model**

We use a simple computer simulation of the neutral model written in a Java-based simulation package called REPAST v. 2.0 (<http://repast.sourceforge.net/>). As schematically shown in figure 1, the simulation arbitrarily assigns numerical variants to a population of  $N$  individuals, which are then subject to repeated mutation and copying, while the occurrence of every variant to appear in the population throughout the run was kept track of

cumulatively (see Hahn & Bentley (2003) for details). Each individual in the starting population had a unique variant. Every time step, the  $N$  individuals are replaced with  $N$  new individuals, each of which receives either a variant copied at random from the previous time step (with probability  $(1 - \mu)$ ) or a novel variant ('mutation'; with probability  $\mu$ ).

We ran the neutral model using different numbers of individuals,  $N$ , and mutation rates,  $\mu$ . Under random drift, all variants are copied with a probability that is proportional to their frequency at that moment. However, because chance is still quite important, any new mutation in the drift process stands a small chance of rising to a high frequency in future generations. It is actually expected that some of the newly created variants will become highly popular at some point in the future, while most will be lost. The arbitrary nature of the process means that it is impossible to predict which variants in particular will become common and which will be lost. However, if we choose a set of variants that were 'born' during approximately the same time period and follow their changing frequencies through time, the expected average frequency stays the same, even though the standard deviation of variant frequencies increases with age (Wright 1931; Crow & Kimura 1970).

An elegant pattern emerges from this random process over time. Figure 2 shows the variant frequency distributions, counted cumulatively over 1000 time steps, resulting from runs with  $N = 250$  individuals and a range of  $\mu$ -values. The consistent result of each run is very nearly a power-law distribution  $P(v)$  of variant frequencies,  $P(v) \sim C/v^\alpha$ , which in each case appears as an almost straight line, with a slope of  $-\alpha$ , on the log-log

Table 1. Power-law slope,  $\alpha$ , for the frequency distribution resulting from different values of  $N$  and  $\mu$  showing the average of five neutral-model runs ( $\pm 1$  s.d. in parentheses). Diagonally linked cells (shaded) have the same value of  $N\mu$ .

$\mu$	$N = 125$	$N = 250$	$N = 500$	$N = 1000$
0.004	1.46 (0.02)	1.50 (0.02)	1.53 (0.03)	1.52 (0.01)
0.008	1.50 (0.02)	1.52 (0.02)	1.57 (0.01)	1.58 (0.02)
0.016	1.54 (0.04)	1.55 (0.03)	1.61 (0.04)	1.63 (0.03)
0.032	1.64 (0.05)	1.59 (0.06)	1.70 (0.05)	1.73 (0.06)
0.064	1.74 (0.04)	1.78 (0.08)	1.81 (0.10)	1.91 (0.02)
0.128	1.85 (0.12)	1.99 (0.04)	1.98 (0.02)	2.01 (0.11)

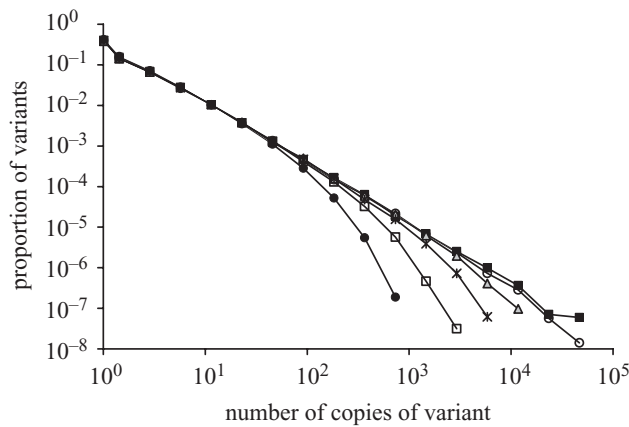


Figure 2. Distributions of the number of variants accumulated during a run of the neutral simulation for 1000 time steps, with  $N = 250$  individuals, and six different values of  $\mu$ . The plot shows the average result from five runs at each value. Filled squares,  $\mu = 0.004$ ; open circles,  $\mu = 0.008$ ; triangles,  $\mu = 0.016$ ; crosses,  $\mu = 0.032$ ; open squares,  $\mu = 0.064$ ; filled circles,  $\mu = 0.128$ .

plot in figure 2. The distribution is a power law for a wide range of mutation rates ( $\mu = 0.004$  to  $\mu = 0.128$ ) and population sizes ( $N = 125$  to  $N = 1000$ ), such that  $N\mu$  is not too large. We characterize each distribution by  $\alpha$ , the best-fit linear slope on a log-log plot (table 1), on which all of the distributions have an  $r^2$  value above 0.93. Table 1 shows the values for  $\alpha$  for 24 different combinations of  $N$  and  $\mu$  drawn from within these ranges. Note that for combinations where  $N\mu$  is the same, e.g.  $N\mu = 4$  ( $N = 125$ ,  $\mu = 0.032$ ;  $N = 250$ ,  $\mu = 0.016$ ;  $N = 500$ ,  $\mu = 0.008$ ;  $N = 1000$ ,  $\mu = 0.004$ ), the results show similar values of  $\alpha$  to within their standard deviations (table 1). This conforms to our expectation (Kimura & Crow 1964; Ewens 1972) that the variant frequency distribution is not determined by  $N$  or  $\mu$  independently but by their product.

The reason that these model results are consistent only for small  $\mu$  is that as  $\mu$  gets large the power-law fit becomes poorer as the tail of the distribution falls off for high frequencies (figure 2). It can easily be seen why the power law must break down as  $\mu$  increases, because in the maximum limit, when  $\mu = 1.0$ , the accumulated frequency of every variant would be one as all would be replaced by mutation in every time step.

We see that as  $N\mu$  increases the exponent  $\alpha$  increases (figure 2), and we plot the relationship between  $N\mu$  and  $\alpha$  in figure 3a, which is well fitted by the function

$$\alpha = 0.1042 \ln(N\mu) + 1.48. \quad (2.2)$$

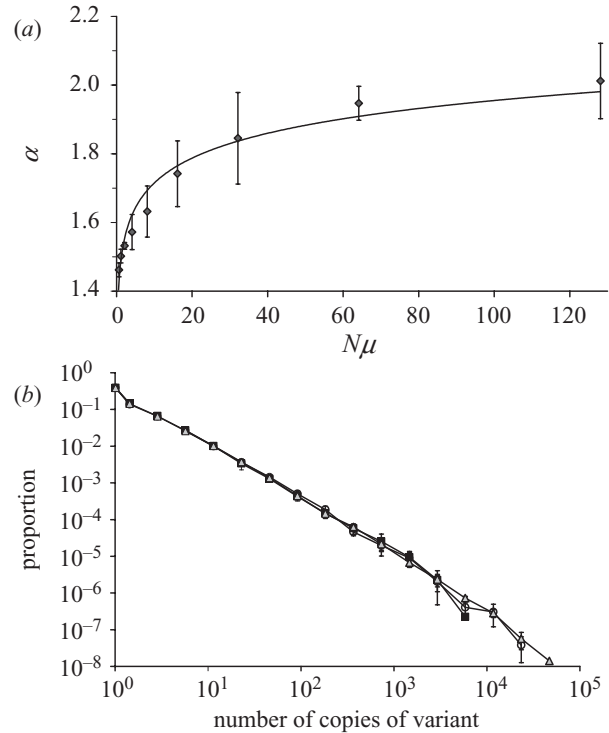


Figure 3. (a) Relationship between the power-law slope,  $\alpha$ , and  $N\mu$ , for the variant frequency distribution resulting from the neutral model run for 1000 time steps. Each point shows the average from all the runs with the same value of  $N\mu$  (table 1);  $y = 0.1042 \ln(x) + 1.4755$ ;  $r^2 = 0.959$ . Error bars show 1 s.d. (b) Relationship between the variant frequency distribution and the duration (in time steps) of the simulation, for the neutral model run with  $N = 250$  and  $\mu = 0.008$  ( $N\mu = 2$ ). Each point represents the average of five runs with error bars showing 1 s.d. Squares, 100 time steps; circles, 500 time steps; triangles, 1000 time steps.

As discussed in § 3, it is extremely difficult to estimate  $N\mu$  empirically for most real-world data, so we do not expect to use equation (2.2) to predict  $\alpha$  from population or mutation data alone. It is, however, straightforward to calculate the ratio of  $N_1\mu_1$  for culture-variant group 1 to  $N_2\mu_2$  for variant group 2. For example, given the frequency distribution of pottery motifs at settlement 1, we might predict the distribution of motifs at settlement 2, knowing that settlement 2 has a population that is five times larger than that of settlement 1 (assuming that the mutation rates are similar) and therefore  $N_2\mu_2/N_1\mu_1 = 5$ . It is therefore useful to write equation (2.2) once for  $\alpha_1$ ,  $N_1$  and  $\mu_1$ , and then again for  $\alpha_2$ ,  $N_2$  and  $\mu_2$ ; the two equations can be combined and rearranged to give

Table 2. Analogies between the neutral model and our three real-world examples.

analogy data	variants	individuals, $N$	mutation rate, $\mu$
baby names	names	new births/time	novel names per birth per time
patents	citations	total citations/time	$1/N \times$ newly cited patents per time
pottery motifs	motifs	pots or households	new motifs per household per time

$$\alpha_1 - \alpha_2 = 0.1042 \ln(N_1\mu_1/N_2\mu_2). \tag{2.3}$$

We can use equation (2.3) to predict the power-law slope,  $\alpha$ , for the distribution of variant frequencies, as long as we are able to determine  $\alpha$  for a reference sample, given the values of  $N\mu$  relative to this reference population. This will become clearer as we analyse our three datasets.

All simulation runs that we discuss from this point forward were for 1000 time steps, unless otherwise stated. We find quite simply that if we plot the variant frequency distribution as a function of the number of copies on the  $x$ -axis, then the duration of the run does not change the distribution except to extend its tail for longer runs (figure 3*b*).

### 3. DATASETS

We now turn to comparisons of the neutral model with datasets on three traits, which are quite different in kind (first names, pottery decoration variants and technological inventions), in time (prehistoric versus modern) and in how they were collected (archaeological excavation versus government recording agencies). These three traits involve variants that are discretely identifiable and that replicate with a high degree of fidelity, such that mutations are obvious, as with patent no. ‘3465784’ or first name ‘Tyler’. While archaeological pottery motifs may not be as discrete, in this particular case they are easily identifiable, faithfully replicated and different from one another (Frirdich 1994), much as the same letters of the alphabet written by different hands are still recognizable.

Table 2 shows how we view these as cultural variants in relation to the neutral model. Different first names are seen as variants, and people given that name are seen as individuals. Decorative motifs, recognizably reproduced on different pottery vessels, are seen as the variants, whereas the pots on which they are found may be seen as the individuals. With patents, we liken each reference to another patent within a patent application to an individual copying a single variant in the neutral model. So we can let the number of individuals,  $N$ , be the total number of new citations per time interval. Then the mutation rate,  $\mu$ , is the chance that one of those new references cites a patent that has never been cited before, which is equivalent to the number of newly cited patents per time step divided by  $N$ . Hence  $N\mu$  is approximately the number of newly cited patents per time step.

After introducing each dataset, we explore it in terms of the overall variant frequency distribution and by using one frequency distribution to predict other distributions, using only equation (2.3).

#### (a) *Dataset 1: first names in the twentieth century USA*

The US Census Bureau ([www.census.gov/genealogy/www/freqnames.html](http://www.census.gov/genealogy/www/freqnames.html)) provides lists of first-

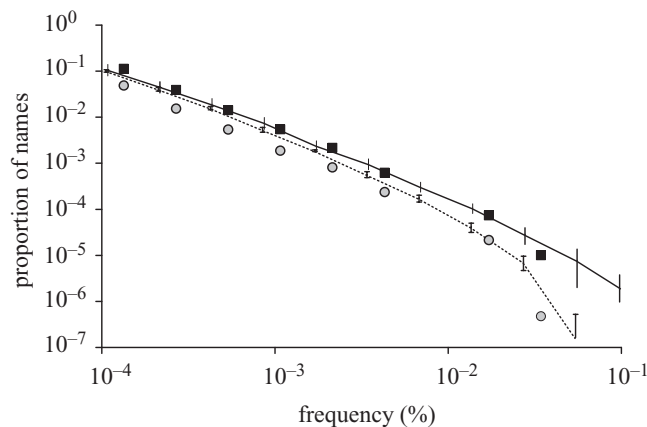


Figure 4. Frequency distribution of first names, from 1990 US Census data. The popularity of a name is measured as a fraction of the US population. The exponent  $\alpha$  for male names (squares) is 1.73 ( $r^2 = 0.990$ ) and for female names (circles) is 1.93 ( $r^2 = 0.968$ ). Dotted line,  $N\mu = 16$  model (100 time steps); solid line,  $N\mu = 8$  model (100 time steps).

name frequencies in a sample of 6.3 million Americans, about one-fortieth of the US population, from the 1990 US Census. These census data include 4275 different female first names and 1219 different male names. An additional dataset was provided by a US Social Security Administration Web site ([www.ssa.gov/OACT/babynames/](http://www.ssa.gov/OACT/babynames/)) that ranks the 1000 most common boys’ and girls’ names in each decade of the twentieth century, from a sample of 5% of the social-security cards issued.

Like that of the variants in our neutral simulation, the observed distribution of US Census names (figure 4) exhibits power laws for both female ( $\alpha = 1.93$ ,  $r^2 = 0.968$ ) and male ( $\alpha = 1.73$ ,  $r^2 = 0.990$ ) names. Similarly, Hahn & Bentley (2003) showed that the distributions of baby-name frequencies are fitted by power laws for each decade of the past century, with all  $r^2$  values above 0.97 for both male ( $\alpha = 1.70 \pm 0.07$  for the 10 decades) and female ( $\alpha = 1.84 \pm 0.06$ ) names. We see that the  $\alpha$ -values for the census and baby-name data are not significantly different. Although the names in the 1990 Census have accumulated over approximately a century (assuming that not many people are over 100 years old), rather than a decade as for the baby names, the census distributions are similar to the decadal baby-name distributions because the shape of the distribution does not change with the duration of the neutral simulation (figure 3*b*), even though the underlying frequencies of individual variants do change. Hence the neutral model fits the frequency distributions from both the census data (figure 4) and the baby-name data (Hahn & Bentley 2003) quite well.

Having shown that the neutral model fits the distribution of first names, we would also like to use this model to predict the frequency distribution of data using

equation (2.3). In both the census and baby-name cases,  $\alpha$  is higher for female than for male names. Under the neutral model, this higher  $\alpha$ -value can be explained by a higher mutation rate for females: there is an average of 2.3 new female names versus 1.6 new male names in the top 1000 list per 10 000 births in the population (Hahn & Bentley 2003). Although we cannot determine  $N\mu$  absolutely because the mutation rate depends on our arbitrary definition of a 'time step', we can obtain an estimate of  $(N\mu)_{\text{female}}/(N\mu)_{\text{male}}$  from the census data, which include 4275 different female first names and 1219 different male names. Plugging  $(N\mu)_{\text{female}}/(N\mu)_{\text{male}} = 4275/1219$  into equation (2.3), we get  $\alpha_{\text{female}} - \alpha_{\text{male}} = 0.1042\ln(3.5)$  or  $\alpha_{\text{female}} - \alpha_{\text{male}} = 0.13$ . With the values that Hahn & Bentley (2003) determined for  $\alpha_{\text{female}}$  and  $\alpha_{\text{male}}$ , we find that  $\alpha_{\text{female}} - \alpha_{\text{male}} = 0.14$  for the decadal baby-name data. Using the simple relationship in equation (2.3), therefore, we are able to predict the shape of the distribution of baby names from the census data.

### (b) Dataset 2: pottery motifs from Neolithic Germany

Our second dataset includes decorative motifs on archaeological pottery excavated from early farming settlements along a 1.3 km stretch of the Merzbach river, near Bonn in western Germany. Here, at least 160 long rectangular houses were occupied during a span of over 400 years, *ca.* 5300 BC–4850 BC (Lüning & Stehli 1994; Stehli 1989). From the recovered pottery, Frirdich (1994) identified 35 types of distinctive decorative motif, as well as a chronology for them (see Shennan & Wilkinson 2001, fig. 3).

As the prehistoric population changed over these 400 years, there was an approximately one-to-one correlation between the number of occupied houses and the concurrent number of different decorative motifs being used (Bentley & Shennan 2003, fig. 9). This linear correlation is expected under the neutral model, which predicts that the effective number of variants will be proportional to  $N\mu$ .

Figure 5*a* shows the frequency distributions for motifs whose first appearances were in different chronological phases (each phase is approximately a generation, perhaps 25 years). By analogy with the neutral model, the motifs that first appeared in phases of different ages are analogous to runs of different lengths (figure 3*b*). As with the neutral model, the shape of the power-law distribution for the pottery motifs does not change, even though the motifs from phase 6, having been around longest for people to copy, show a power law covering more orders of magnitude (figure 5*a*). Allowing for the limits of archaeological data, the pottery motifs from each of the different chronological phases are fitted well by the neutral model.

There were several distinct settlements within the Merzbach Valley during its 400 year occupation, including a founding settlement (Langweiler 8) and subsequent foundations that probably split off from the initial settlement. Hence we have a chance to test the effect of different values of  $N$  on the distributions of motif frequencies. When the motifs are sorted according to settlement, the frequency distributions are also power laws, with almost identical values of  $\alpha$  (figure 5*b*; table 3). Figure 5*b* shows that simulations of the neutral model with  $N\mu = 1$  recapitulate the observed distributions. According to

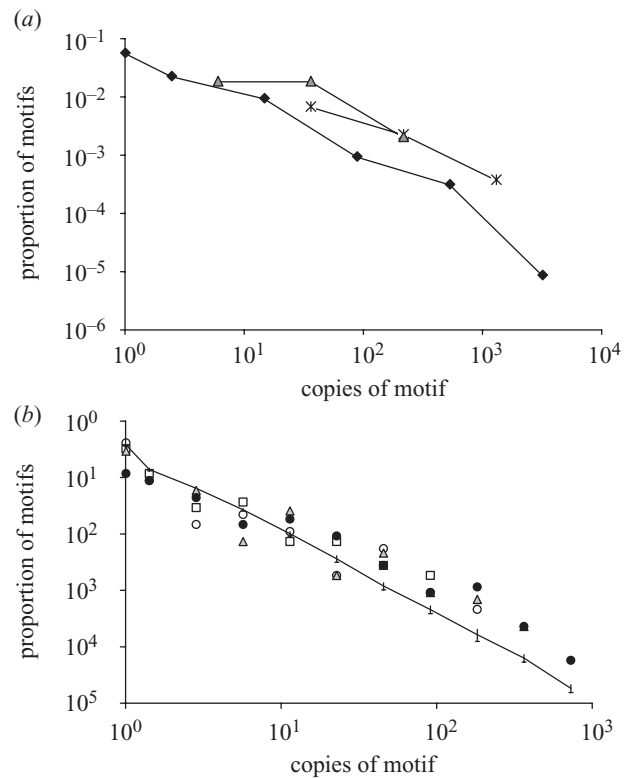


Figure 5. (a) Frequency distributions of Merzbach pottery motifs for three different time spans during its 400 year occupation: diamonds, phases 1–6; crosses, phase 7; triangles, phases 9–11 (because the data are limited, phases 1–6 and 9–11 were combined). After Bentley & Shennan (2003, fig. 15). (b) Frequency distributions of pottery motifs (cumulative over 400 years) for different settlements within the Merzbach valley (open circles, site 1; squares, site 2; triangles, site 3; filled circles, site 8) along with the results from the neutral model (solid line) with  $N\mu = 1$ .

equation (2.3), we should be able to use the  $\alpha$ -value for one settlement to predict the  $\alpha$ -values at the other settlements if we know how the different values of  $N\mu$  for the settlements compare with one another. We can characterize the average values of  $N\mu$  relative to each other by assuming that  $\mu$  was approximately the same for each settlement and estimating  $N$  from the number of recovered pottery samples per phase (Shennan & Wilkinson 2001, p. 587). Table 3 shows that, even though  $N$  varies widely, with  $\alpha$ -values predicted to be smaller for the smaller settlements, the observed  $\alpha$ -values are all about the same. The simplest explanation for this is that these settlements, which lay along only a few kilometres of the Merzbach River, were not distinct when it came to the way in which potters copied decorations from one another's pots. In this case, there is only one effective population of individuals, rather than several distinct smaller populations, and the highly similar power laws in figure 5*b* are simply samples from a single power law for the valley as a whole. This is supported by evidence that the site LB3 had pottery but not house remains, and yet the mutation rates for sites LW8 and LB3 paralleled each other over time (Shennan & Wilkinson 2001, tables 3 and 4). The predictions of the neutral model can thus inform us about cultural exchange between these prehistoric settlements.

Table 3. Estimates of  $N$  for different Merzbach sites, and  $\alpha$ -values for the power-law distribution of motif frequencies at each site.

site	$N$ (pots per phase)	observed $\alpha$	$r^2$	predicted $\alpha$ based on $\alpha_{LWS}$ and pots per phase
site 8 (LW8)	375	1.08	0.97	—
site 3 (LB3)	190	1.12	0.90	1.01
site 2	99	1.11	0.94	0.94
site 1	74	1.08	0.94	0.91

**(c) Dataset 3: US patents and their citations**

The US Patent and Trademark Office (USPTO) provides an online database of patents and their citations ([www.uspto.gov](http://www.uspto.gov)), which includes text-searchable records with titles, abstracts, citations, dates filed and dates issued for patents granted since 1976. In addition to the searchable online data, the National Bureau of Economic Research Web site ([www.nber.org/patents](http://www.nber.org/patents)) provides downloadable files listing all US patents filed since 1963 ( $n = 2.92$  million) and the patents cited by each ( $n = 16.5$  million citations in total). There is no imposed limit on how old the cited patents can be: some even go a hundred years into the past (Hall *et al.* 2002). During the 1980s, when the patent database was computerized, the average number of citations made per patent increased, from about five in 1975 to over 10 by the late 1990s.

For all US patents since 1963, the distribution of the number of times patents have been cited is not completely described by a power law (figure 6a). The distribution starts off flatly at the low end, but becomes a power law for citations greater than *ca.* 10, with the exponent  $\alpha$  fitted to this power-law tail equal to 3.75. The flatter low end reflects the fact that many patents are cited at least a few times. The absence of a power law at the low end of the distribution may demonstrate that patents do not completely conform to the neutral model. This may be because patent-office regulations enforce the citation of relevant patents (Hall *et al.* 2002), leading to patents being cited that would otherwise be lost by sampling in the population.

However, the patent data largely suggest properties of random drift. As figure 7 shows, the distribution of patent citations conforms to the neutral model in appearing to remain constant over different periods of time.

To acquire citation data for particular technological ‘niches’, we did title-word searches on the USPTO Web site for ‘compact disc’ (CD) and ‘automobile’, downloaded the reference list from each patent returned in the search and compiled all references together for each sample (‘CD’, 6932 total references; ‘automobile’, 16 383 references). We then counted the number of times that each patent had been cited within the list. Determined in this way, the ‘CD’ citations show a power law, with  $\alpha = 2.96$ , over two orders of magnitude (figure 6b), and the ‘automobile’ citations show a slightly steeper power law, with  $\alpha = 3.23$ , over two orders of magnitude (figure 6c). These values of  $\alpha$  are too high for us to reproduce with our neutral simulation: by equation (2.3), we would have to use an  $N\mu$  value of *ca.* 1.4 million to reproduce a power-law distribution with a slope of 2.96. Again, this may reflect the influence of legal regulations on patent citations (Hall *et al.* 2002).

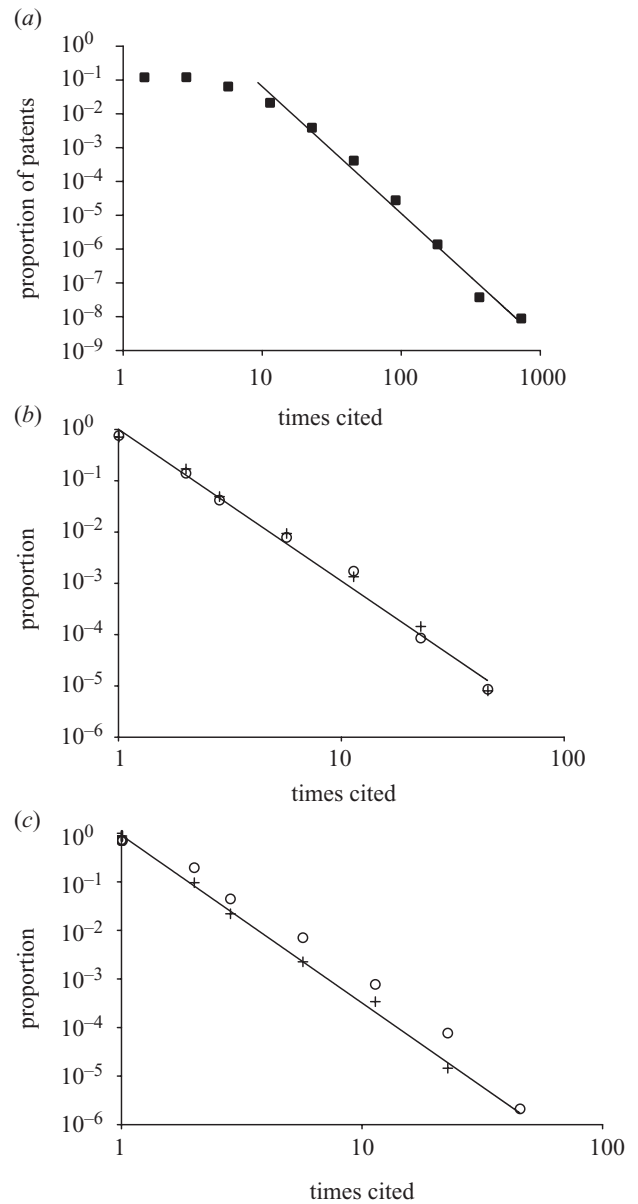


Figure 6. (a) The frequency distribution of number of times cited for all US patents since 1963;  $\alpha = 3.75$ ,  $r^2 = 0.99$ . (b) The frequency distribution of number of times cited for all US patents with ‘CD’ in the title (plus symbols), or anywhere in the text (circles) of the application;  $\alpha = 2.96$ ,  $r^2 = 0.99$ . (c) The frequency distribution of number of times cited for all US patents with ‘automobile’ in the title (plus symbols), or anywhere in the text (circles) of the application;  $\alpha = 3.45$ ,  $r^2 = 0.996$ .

Although we cannot simulate the neutral model for this dataset, we can estimate the relative differences in  $N\mu$  between CD, automobile and all patents by comparing

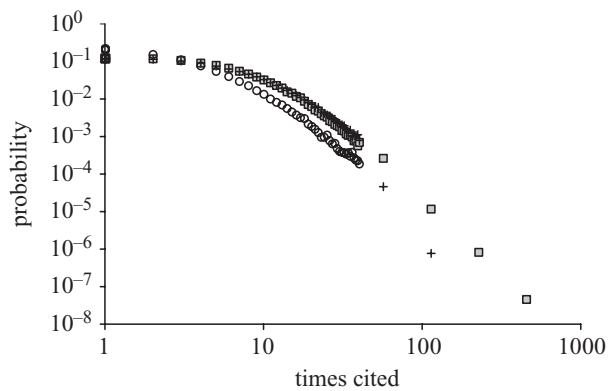


Figure 7. Distribution of number of times cited for all the patents, in yearly time slices: squares, 1975; plus symbols, 1985; circles, 1995.

the different rates at which patents have appeared in these categories. Up to 1983, the rate of all US patent applications was constant, at *ca.*  $65\,000 \pm 10\,000 \text{ yr}^{-1}$ , after which there was a sharp increase as the rate reached almost  $140\,000 \text{ yr}^{-1}$  by the mid-1990s (Hall *et al.* 2002). The first patent with ‘CD’ in the title was filed in 1984, after which the rate of new CD patents increased until it reached a steady rate of *ca.*  $55 \text{ patents yr}^{-1}$  by about 1991. The number of patents with ‘automobile’ in the title since 1976 ( $n = 5358$ ) has averaged *ca.*  $222 \text{ patents yr}^{-1}$ , also at a steady rate.

Given the  $\alpha$ -value for one of the power-law distributions, we can infer  $N\mu$  from the relation in equation (2.3), and then use that to calculate the other values of  $N\mu$ . If  $t$  is the number of time steps per year, then  $N\mu = 65\,000/t$  for all patents,  $222/t$  for ‘automobile’ and  $55/t$  for ‘CD’. We calculate  $(N\mu)_{\text{aut}}/(N\mu)_{\text{CD}} = 222/55 = 4.04$  and  $(N\mu)_{\text{all}}/(N\mu)_{\text{CD}} = 65\,000/55 = 1182$ , giving  $\alpha_{\text{aut}} - \alpha_{\text{CD}} = 0.15$  and  $\alpha_{\text{all}} - \alpha_{\text{CD}} = 0.74$ , from which we would predict  $\alpha_{\text{aut}} = 3.11$  and  $\alpha_{\text{all}} = 3.70$ . These predictions compare well with the values observed from the data of  $\alpha_{\text{aut}} = 3.23$  and  $\alpha_{\text{all}} = 3.75$ . The predictions seem particularly good because we have used so little to make them: given the citation distribution for CDs, we can predict the distributions for two unrelated samples of patents, using only the relative ratios of their filing rates.

It appears from the distributions in figure 6, representing sample sizes ranging from the small set of ‘CD’ patents to the set of ‘all’ patents, that the power-law exponent may increase with sample size. As a test, we also searched for ‘CD’ and ‘automobile’ anywhere in the patent, rather than just in the title. While this returned about an order of magnitude more patents and references in each case (CD, 6197 patents and 24 780 references; automobile, 29 483 patents and 98 352 references), the frequency distribution is almost identical in the CD case, and only slightly different in the automobile case (figure 6); the power-law distributions did not change simply because we increased our sample sizes.

#### 4. DISCUSSION AND CONCLUSIONS

We have shown that the frequency distributions of cultural variants can be explained by the neutral model—a simple model of random copying—for three real-world

examples: first names, archaeological pottery motifs and citations of patented inventions.

While the neutral model is unarguably the simplest model that can explain the observed patterns of cultural variants, population geneticists have often had a difficult time distinguishing between predictions of the neutral model and those made by models with natural selection. For instance, Gillespie (1977) has shown that models of selection in fluctuating symmetric environments are often indistinguishable from a neutral model. Indeed, a protracted debate over the merits of neutral and selective models in population genetics has carried on for many years (e.g. Kimura 1983; Gillespie 1991). One has only to look to the field of ecology to see the controversy that a neutral model of biodiversity has sparked (Hubbell 2001; McGill 2003; Volkov *et al.* 2003).

Additional complexities of applying the neutral model to cultural change are the meaning and detection of neutrality. For population geneticists, finding a population of variants that are neutral does not necessarily imply that the trait is unconstrained: neutrality merely means that the variants of the trait that are observed in the population are neutral. It does not mean that any imagined variant will also be value-neutral if introduced into the population (Kimura 1983). We suggest that this meaning be carried over into the social sciences, as it helps to clarify the interpretation of results and correctly distinguishes between variants and the traits that vary. Identifying truly unconstrained cultural traits or characters may be quite difficult. Only by careful examination of the data and comparison with theoretical predictions can one ascribe ‘neutrality’ to a population of variants or the more restrictive designation ‘unconstrained’ to a trait. In this paper, for instance, we find baby names to be evolving neutrally, but we have not shown that the name trait itself is ‘neutral’ or ‘unconstrained’. In fact, the lower mutation rate in male names is probably a result of this trait being more constrained than female names, i.e. there are proportionally fewer variants that are neutral for males than for females. Similarly, it may be that specific pottery motifs, perhaps representing prominent households, were not neutral (Shennan & Wilkinson 2001; Bentley & Shennan 2003).

Even considering these details, however, we, like many before us (Dunnell 1978; Neiman 1995; Lipo *et al.* 1997; Bentley & Shennan 2003; Hahn & Bentley 2003), favour the neutral model for studying cultural change because it is the simplest null model, with a minimum of free parameters, against which hypotheses can be quantitatively tested using the many tools provided by population genetics (Crow & Kimura 1970; Hartl & Clark 1997). The frequency distribution of variants is a power law whose slope can be predicted from the neutral model if just two quantities are known, the effective population size,  $N$ , and the mutation rate,  $\mu$ . In fact, because what actually determines the slope is the product  $N\mu$ , only one parameter need be known, which is the number of new variants appearing per ‘time step’. Given the slope,  $\alpha$ , of the power-law distribution of a reference sample of variant frequencies, the slope of another sample can be remarkably well predicted by an empirical relation (equation (2.3)) and the ratio of the new  $N\mu$  value to the  $N\mu$  of the reference sample. We found that this prediction method

worked well with baby names, pottery decorations and patents, giving further evidence that these are changing neutrally. It may be somewhat surprising that, for oft-cited patents, the way patents are cited reflects a neutral model. But if scientists can copy each other's papers in a seemingly random way (Simkin & Roychowdhury 2003), perhaps so can inventors or users of the World Wide Web, because links to Web sites are power-law distributed (cf. Barabási & Albert 1999; Albert & Barabási 2002).

We thank A. Diaz-Guilera for comments in the early stages of writing and assistance in retrieving patent data from the Internet. We also thank J. Gillespie for valuable comments and criticism.

## REFERENCES

- Albert, R. & Barabási, A.-L. 2002 Statistical mechanics of complex networks. *Rev. Mod. Physics* **74**, 47–97.
- Barabási, A.-L. & Albert, R. 1999 Emergence of scaling in random networks. *Science* **286**, 509–512.
- Bentley, R. A. & Shennan, S. J. 2003 Cultural transmission and stochastic network growth. *Am. Antiquity* **68**, 459–485.
- 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.
- Dunnell, R. C. 1978 Style and function: a fundamental dichotomy. *Am. Antiquity* **43**, 192–202.
- Ewens, W. J. 1972 The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3**, 87–112.
- Firdich, C. 1994 Kulturgeschichtliche Betrachtungen zur Bandkeramik im Merzbachtal. In *Die bandkeramik im merzbachtal auf der aldenhovener platte* (ed. J. Lüning & P. Stehli), pp. 207–394. Bonn, Germany: Habelt.
- Gillespie, J. H. 1977 Sampling theory for alleles in a random environment. *Nature* **266**, 443–445.
- Gillespie, J. H. 1991 *The causes of molecular evolution*. Oxford University Press.
- Gillespie, J. H. 1998 *Population genetics: a concise guide*. Baltimore, MD: Johns Hopkins University Press.
- Hahn, M. W. & Bentley, R. A. 2003 Drift as a mechanism for cultural change: an example from baby names. *Proc. R. Soc. Lond. B* **270**(Suppl. 1), S120–S123. (DOI 10.1098/rsbl.2003.0045.)
- Hall, B. H., Jaffe, A. B. & Trajtenberg, M. 2002 The NBER patent citations data file: lessons, insights and methodological tools. In *Patents, citations, and innovations: a window on the knowledge economy* (ed. M. Trajtenberg), pp. 403–460. Cambridge, MA: MIT Press.
- Hartl, D. L. & Clark, A. G. 1997 *Principles of population genetics*. Sunderland, MA: Sinauer Associates.
- Hubbell, S. P. 2001 *The unified neutral theory of biodiversity and biogeography*. Princeton University Press.
- Kimura, M. 1983 *The neutral theory of molecular evolution*. 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.
- Lipo, C. P., Madsen, M. E., Dunnell, R. C. & Hunt, T. 1997 Population structure, cultural transmission and frequency seriation. *J. Anthropol. Archaeol.* **16**, 301–333.
- Lüning, J. & Stehli, P. (eds) 1994 *Die bandkeramik im merzbachtal auf der aldenhovener platte*. Bonn, Germany: Habelt.
- McGill, B. J. 2003 A test of the unified neutral theory of biodiversity. *Nature* **422**, 881–885.
- Neiman, F. D. 1995 Stylistic variation in evolutionary perspective. *Am. Antiquity* **60**, 7–36.
- Shennan, S. J. 2002 *Genes, memes and human history*. London: Thames and Hudson.
- Shennan, S. J. & Wilkinson, J. R. 2001 Ceramic style change and neutral evolution: a case study from Neolithic Europe. *Am. Antiquity* **66**, 577–594.
- Simkin, M. V. & Roychowdhury, V. P. 2003 Read before you cite! *Complex Systems* **14**, 269.
- Stehli, P. 1989 Merzbachtal–Umwelt und Geschichte Einer Bandkeramischen Siedlungskammer. *Germania* **67**, 51–76.
- Volkov, I., Banavar, J. R., Hubbell, S. P. & Maritan, A. 2003 Neutral theory and the relative species abundance in ecology. *Nature* **424**, 1035–1037.
- Wright, S. 1931 Evolution in Mendelian populations. *Genetics* **16**, 97–159.

As this paper exceeds the maximum length normally permitted, the authors have agreed to contribute to production costs.