Cross-correlations of American baby names

Paolo Barucca[a,1,2], Jacopo Rocchi[a,1], Enzo Marinari[a,b], Giorgio Parisi[a,b,2], and Federico Ricci-Tersenghi[a,b]

[a]Dipartimento di Fisica, Sapienza Universitá di Roma, I-00185 Rome, Italy; and [b]Sezione di Roma 1, Istituto Nazionale di Fisica Nucleare, I-00185 Rome, Italy

Contributed by Giorgio Parisi, April 27, 2015 (sent for review October 11, 2014; reviewed by R. Alexander Bentley)

The quantitative description of cultural evolution is a challenging task. The most difficult part of the problem is probably to find the appropriate measurable quantities that can make more quantitative such evasive concepts as, for example, dynamics of cultural movements, behavioral patterns, and traditions of the people. A strategy to tackle this issue is to observe particular features of human activities, i.e., cultural traits, such as names given to newborns. We study the names of babies born in the United States from 1910 to 2012. Our analysis shows that groups of different correlated states naturally emerge in different epochs, and we are able to follow and decrypt their evolution. Although these groups of states are stable across many decades, a sudden reorganization occurs in the last part of the 20th century. We unambiguously demonstrate that cultural evolution of society can be observed and quantified by looking at cultural traits. We think that this kind of quantitative analysis can be possibly extended to other cultural traits: Although databases covering more than one century (such as the one we used) are rare, the cultural evolution on shorter timescales can be studied due to the fact that many human activities are usually recorded in the present digital era.

clustering | cultural evolution | cultural traits | complex systems

Cultural traits are behavioral patterns shared by the members of social communities. Traditions, religions, beliefs, language, and values are some examples. Far from being static and isolated, they are continuously evolving and interacting with the external environment, e.g., other communities and mass media, and they can be transmitted among members of communities on timescales that are much shorter than those characterizing cultural movements. Although changes in cultural movements may occur over decades or centuries, changes in cultural traits may be observed from a daily to a yearly basis, depending on the trait. An accurate analysis of existing, public data can teach a lot about their reciprocal influence. A cultural trait may promote or prevent the popularity rise of others, a past cultural trait may have an influence on current and future ones, and finally the rise or fall of a cultural trait in a certain area may influence cultural traits in other areas. Cultural traits can be considered as the fundamental blocks of the culture of communities, and their evolution can be used to describe the evolution of society.

Some of the most important progress in the understanding of the evolutionary process of cultures is described in a number of texts that are at this point classic references: Among them are refs. 1–3. Also many cultural traits have been studied in the past. Among them are those that have negligible differences among each other, in terms of intrinsic costs and benefits, which are usually referred to as neutral traits. They play a special role in our study, as we explain below. Some of these traits are skirt lengths (4), pop songs (5), dog breeds (6), and pottery decorations in the archaeological record (7). Also keywords in academics vocabulary have been the focus of recent interest (8). Data about names given to newborns have been investigated for similar reasons (9, 10), and they are the focus of our investigation.

Names come and go in society, as does any other cultural trait. Most of them have a popularity peak and then disappear. They carry important information on the transformation of the social structure (11, 12). Several quantitative approaches to analyze what can be learned from names have been proposed, and we briefly describe them in the following lines. Compared with other relevant traits, neutral traits and, in particular, names appear very appropriate to study cultural changes, because the success of a name depends mainly on the influence that the surrounding culture wields on the parents of the newborns. Other traits suffer, for example, the influence of external forces, such as that due to the producers, which may artificially shape the tastes of consumers. This is particularly evident in the fashion market and in the music market (13).

The frequency distributions of names given to newborns have fat tails, typical of many complex physical problems (14). Fat-tailed distributions can be generated by different mechanisms (7, 15, 16), and in the case of names they have been given several explanations. A scale-free network was used to study a fashion diffusion process where each node could take one of many values (9), imitating "popular" nodes and avoiding "nonpopular" nodes. A stochastic model for cultural evolution has been proposed recently. Here names were chosen according to both individual preferences and social influence (10). These different mechanisms are all able to reproduce a fat-tailed distribution and were shown to reproduce several features of the real data. The popularity of a name was also shown to be correlated with the popularity of similar names in previous years (13). Furthermore, names were analyzed in terms of activation and inhibition processes (17), to explain their popularity. The rates of the rise and of the fall of the popularity of names were found to be correlated in refs. 11 and 18. The same phenomena are also being studied in the context of collective behavior (19, 20), where limited attention seems to play a crucial role (21, 22), and in the context of citation dynamics (23), where a universal temporal pattern is found.

Some very interesting phenomena on the dynamics of American baby names have been recently discussed in ref. 24, where the authors introduced a new model for the choice of names. The

Significance

Societal and cultural transformations are very general and debated topics, both by scientists (e.g., sociologists) and by public opinion (e.g., artists, music producers, brand manufacturers, and advertising agencies). Although almost everyone would be able to express a position on such arguments, it is much more difficult to support such an opinion based on scientific evidence. In this work we analyze the case of American baby names and describe the evolution of tastes of parents regarding the choice of the name during the years of the last century. Using quantitative methods we find that a deep transformation occurred at the end of the 20th century and suggest that this might be studied from a quantitative sociological point of view.

Author contributions: E.M., G.P., and F.R.-T. designed research; P.B. and J.R. performed research; P.B. and J.R. analyzed data; and P.B., J.R., E.M., G.P., and F.R.-T. wrote the paper.

Reviewers included: R.A.B., Bristol University.

The authors declare no conflict of interest.

[1]P.B. and J.R. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: giorgio.parisi@roma1.infn.it or baruccap@sns.it.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1507143112/-/DCSupplemental.

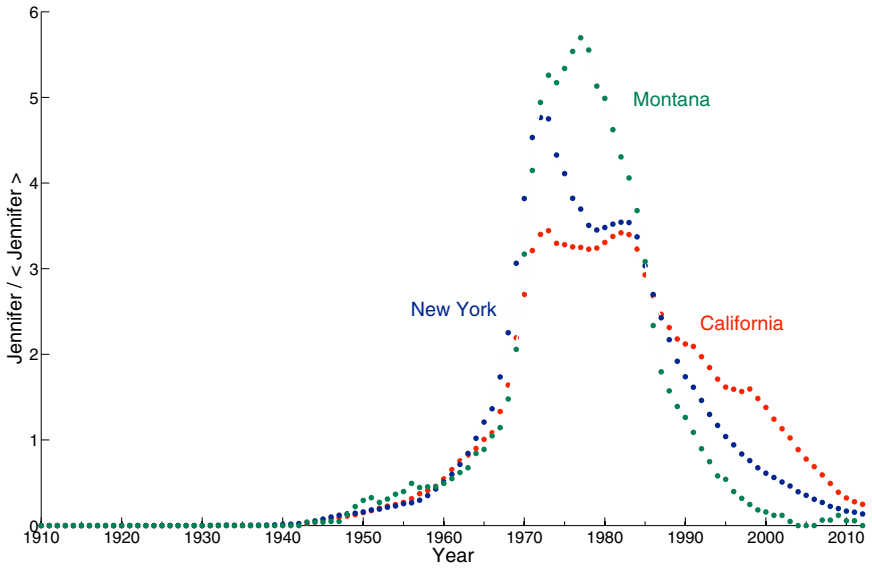

Fig. 1. Number of newborns called Jennifer in the states of California, Montana, and New York, divided by its average values in these three states, as a function of time.

model is defined in terms of a population of agents (babies), each of which holds a single variant (name), and where names of the new generations are given mainly by copying from the last generations but also, sometimes, by inventing new names. Comparing real world data with the model in ref. 24 unveils a considerable increase in the innovation of names in the last part of the century. This is consistent with the main findings of this paper, as we discuss further in the following.

The mechanisms behind the spreading of cultural traits are still debated. The original hypothesis of Simmel (25) was that a fashion arises because individuals of lower social status copy those of perceived higher status. This is the idea used for the analysis done in ref. 9. This approach is different from the neutral model proposed in ref. 26, where naming was considered in close connection with the infinite-allele model of population genetics with a random genetic drift. A preference model of fashion (18) where individuals can copy preferences of other agents, was said to better reproduce the empirical features of American baby names. These studies on names were mainly focused on global distributions, but not on the relations between local distributions of names in different states of the United States (i.e., distributions in single states). We believe that much can be learned, for example, from the relations between local distributions of names in different states. Our main working hypothesis is that local changes of names convey a large body of information on the mutual cultural influences that communities (states) wield on each other.

We focus our correlation analysis on different states of the United States during the 20th century. Statistics on names given to newborns in the United States can be downloaded from the webpage of the US Social Security Administration (SSA) (27). Different states have different popularity spreading curves for each name (Fig. 1) (many of the common names rise and fade with a very similar behavior). The overlap between these curves could be used to describe the similarities between US states. Instead of considering these overlaps in time, we consider the correlations between states on a yearly basis, by studying the whole distribution of baby names in every state (Fig. 2). This analysis gives robust results, as we show in the following.

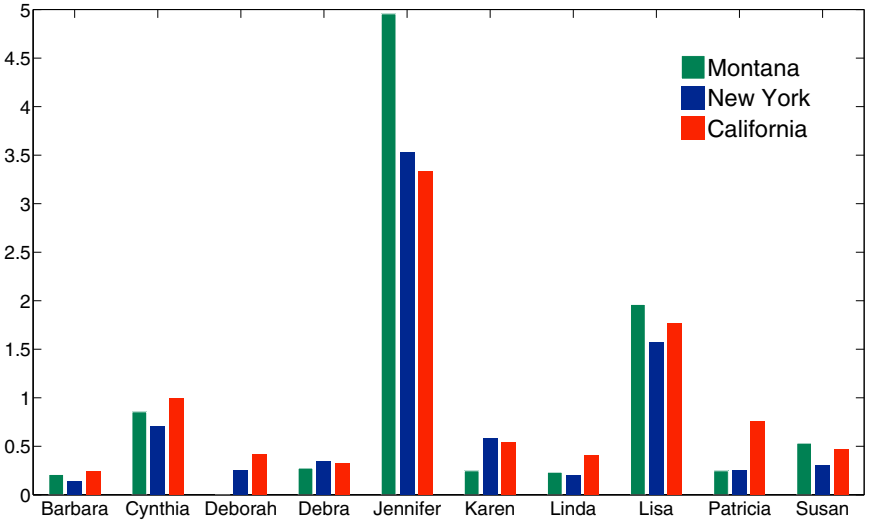

Fig. 2. Histogram of the occurrences of the 10 most popular names given in 1980 in the states of California, Montana, and New York. Normalization is as in Fig. 1.

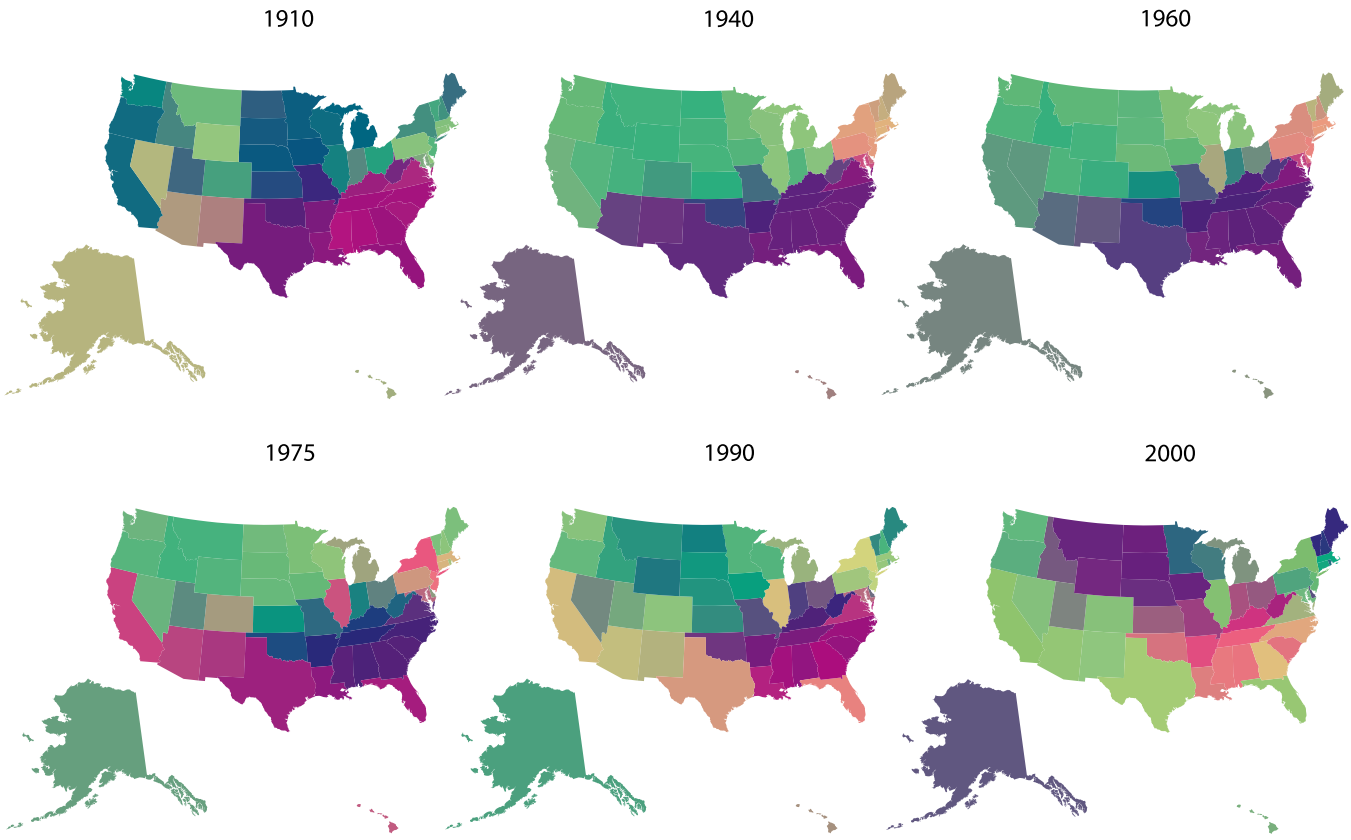

Fig. 3. The colors assigned to the states reflect the similarity in their distributions of names. We use a notion of similarity based on the principal component analysis of the matrices of states correlations C_{ij}, defined in the text. Details are provided in *Methods*. A difference between the central decades of the 20th century and its last decades is clearly visible. Northern and southern states were very correlated among them and were forming two separated, uncorrelated entities until 1960. A new configuration emerges at the end of the 20th century and eventually ends up with a patchy situation where central and coastal states are correlated among them (with very long-range correlations covering thousands of miles of physical distance) and are roughly part of two different cultural areas. Data shown in this figure have been obtained looking at baby girl names. Movies provide an animated visualization of the cultural transition both for girls (Movie S1) and boys (Movie S2).

Methods

For all available years [that range, in the SSA archives (27), from 1910 to 2012] we study how names given in a state i are correlated to names given in a state j. The distribution of these names has already been analyzed (28) and it is further described in *SI Text* (Figs. S1 and S2). For each pair of states i and j, with $i, j = 1, \ldots, M = 51$ (the federal district of Washington, DC is considered by itself), we evaluate a correlation coefficient C_{ij}, computed as follows. Let us consider a generic year y and let $n_S(q)$ be the number of girls named q born in the state S in the year y (we limit ourselves to describing the girls' case as we have verified that analyzing baby boys' names leads to the same conclusions). In each year, we have a $N_f \times M$ rectangular matrix, where $N_f = 19{,}492$ is the total number of different girl names present in the database, and $M = 51$ the number of US states. The entries of the matrix, $n_S(q)$, are the occurrences of the baby girl names, with $q = 1, \ldots, N_f$ and $S = 1, \ldots, M$. Because the information provided by the SSA includes only names that occur at least five times, if the name q has been used less than five times in the state S in the year y, then we have $n_S(q) = 0$. These matrices are sparse, and only an average of 3% of the entries are different from zero. The frequency $f_S(q)$ of the name q in the state S is given by

$$f_S(q) = \frac{n_S(q)}{\sum_{q=1}^{N_f} n_S(q)}. \quad [1]$$

The average frequency of the name q over the states is

$$\bar{f}(q) = \frac{1}{M} \sum_{S=1}^{M} f_S(q). \quad [2]$$

It is useful to define the quantities

$$\tilde{f}_S(q) = f_S(q) - \bar{f}(q), \quad [3]$$

which are related to fluctuations of the frequencies of the names over the states. The average of $\tilde{f}_S(q)$ over all of the names is zero in each state S, as can be explicitly seen from their definition

$$\left\langle \tilde{f}_S(q) \right\rangle_q = \frac{1}{N_f} \sum_{q=1}^{N_f} \tilde{f}_S(q) = \frac{1}{N_f} - \frac{1}{N_f} = 0, \quad [4]$$

given that in each state S, $\sum_q f_S(q) = 1$. To compute how the names in the state i are correlated with the names in the state j, we analyzed, year after year, the Pearson correlation between the variables $\tilde{f}_i$ and $\tilde{f}_j$, which is the $M \times M$ square matrix

$$C_{ij} = \frac{\sum_{n=1}^{N_f} \tilde{f}_i(n) \tilde{f}_j(n)}{\sqrt{\sum_{n=1}^{N_f} \tilde{f}_i^2(n)} \sqrt{\sum_{n=1}^{N_f} \tilde{f}_j^2(n)}}. \quad [5]$$

This matrix can be used to capture the emergence of complex correlations between clusters of states and to study their evolution in time. However, separating the interesting information from the underlying noise is a nontrivial problem. Similar issues have already been faced when analyzing biological problems (29–32) and financial stock markets (33–36) as well as in Internet traffic analysis (37) and in the statistics of atmospheric correlations (38). The main point is that even though the empirical correlation matrix is noisy, it does have stable properties (Figs. S3 and S4). We checked these properties, such as the eigenvalue spectrum and eigenvector localization, and compared them to the ones implied by a null hypothesis, i.e., to the properties of random matrices (Fig. S5).

Here we apply two general methods for the analysis of correlation matrices, i.e., principal component analysis (PCA) and hierarchical clustering

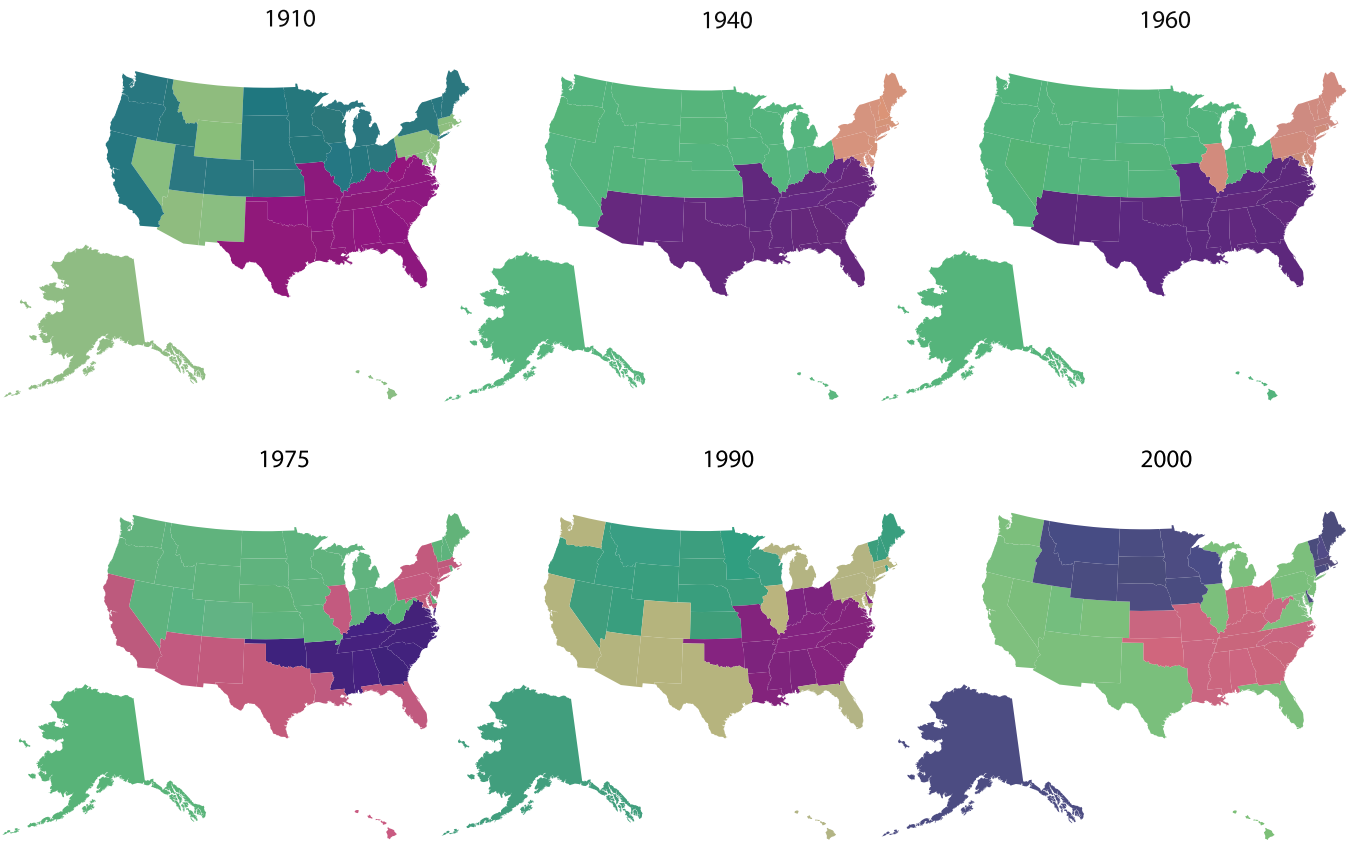

Fig. 4. Hierarchical clustering of the states. This analysis clearly shows that groups of states form well-separated clusters. Details on the clustering method are provided in *SI Text*. The social evolution of geographical correlation is very clear. This figure has been obtained looking at baby girl names. Movies provide an animated visualization of the cultural transition both for girls (Movie S3) and for boys (Movie S4).

(HC). PCA is based on the selection of the eigenvectors corresponding to the largest eigenvalues of the cross-correlation matrix. This choice relies on the hypothesis that smaller eigenvalues are related to noise whereas larger ones are related to the true system dynamics. HC, on the other hand, starts from M clusters formed of one state each and allows one to set up a hierarchy of clusters by merging clusters according to their distances, which can be defined in several ways from their mutual correlations. These two methods give very similar results, both for male and for female names, year after year and for different choices of the metrics in the HC algorithm. These methods are further described in *SI Text*.

Results

Both our algorithms show a clear division of the different states in a number of homogeneous groups. A group of states is qualitatively defined as states that share some level of similarity in their distributions of names, and it is natural to associate them to a common cultural area. In Fig. 3, states in the same group are assigned similar colors. This group structure is robust over timescales of the order of a few years: It is thus worth looking at their evolution over larger timescales. In the beginning of the 20th century states were divided into a group of northern states and a group of southern ones, and this separation remains stable across many years. This structure suddenly breaks down in the last decades of the 20th century, and a new configuration of groups emerges. The evolution of these groups of states is clear in Fig. 3.

In the new stable configuration that emerges at the end of the 20th century, some states of the Atlantic and of the Pacific coasts share common features and belong to the same group, different from that to which many of the central states belong. To better identify these groups of states we used a hierarchical clustering algorithm. A better visualization of this transition can be observed in Movies S1–S4. This method allows a precise and quantitative definition of the groups mentioned above and leads to the formation of clusters, identified by different colors in Fig. 4 (see also Fig. S6). The two different methods give the same answer and make manifest a very interesting social cross-fertilization. This approach is able to describe the emergence of clusters of states through the analysis of the mutual correlations of their newborns names and extracts interesting information on the evolution of these clusters (Figs. S7 and S8).

Discussion

The study of name distributions at the state level enabled us to avoid the effects of strong fluctuations on smaller scales due to local socioeconomic factors, such as for instance economic segregation or ethnicity (39), and to capture macroscopic changes in the structure of mutual correlations. We do not discuss here the origin of these correlations or the mechanisms according to which names are given to newborn babies. Some steps in this direction have been taken in refs. 10 and 24. We also do not study the reasons why there is a reorganization of clusters in the last decades of the 20th century, compared with the relatively stable situation of the first half of the century. These are two very interesting issues that deserves a more specific study. Irregularities in the retarded cross-correlations between the total distributions of the American names were found in ref. 10 to appear in the 1970s; this effect is very probably connected to the reorganization of the clustering of the states that we unveil here. As suggested by the authors in ref. 10, this effect is probably due to the deep cultural transformation that occurred in the United States after the Vietnam War. The authors also proposed a model for the generation of names, showing, very interestingly, that in recent years the inequality between names has been decreasing. The decrease in time

of the probability that two individuals chosen at random from a set of newborn babies share the same name (and the corresponding increase in the diversity of names) is shown in ref. 9 to be related to the shorter timescale of the name's lifetime, whereas the authors of ref. 10 explained how the weakening of social influence could also be a relevant factor. This effect is very closely related to the increase of innovation found by Bentley and Ormerod (24), which detected a steep increase in the coefficient that regulates the amount of invention of new names, compared with the number of names copied by previous generations. One of the possible explanations of this fact is the much deeper interconnectedness of the society of the last decades, compared with that of the last century, whose culture sources were mainly represented by mass media. Much more should be said on this, but this would take us far from the main aims of this work. We mention only that our analysis can be adapted to capture which states influence a given state and which ones are influenced by the same state. This amounts to identifying a directed network, which may be relevant for studying time correlations and culture propagation. In the Internet era one can get high statistical data about an extremely large number of behaviors and cultural traits: Using our approach on the combined ensemble of such abundant datasets should allow one to organize a precise quantitative understanding about the functioning of cultural influences and their evolution.

ACKNOWLEDGMENTS. The research leading to these results has received funding from the Italian Research Ministry through the Futuro in Ricerca Project RBFR086NN1.

1. Feldman MW, Cavalli-Sforza LL (1981) *Cultural Transmission and Evolution: A Quantitative Approach* (Princeton Univ Press, Princeton, NJ), Vol 16.
2. Boyd R, Richerson PJ (1985) *Culture and the Evolutionary Process* (Univ of Chicago Press, Chicago).
3. Mesoudi A (2011) *Cultural Evolution: How Darwinian Theory Can Explain Human Culture and Synthesize the Social Sciences* (Univ of Chicago Press, Chicago).
4. Belleau BD (1987) Cyclical fashion movement: Women's day dresses: 1860-1980. *Cloth Text Res J* 5(2):15–20.
5. Bentley RA, Lipo CP, Herzog HA, Hahn MW (2007) Regular rates of popular culture change reflect random copying. *Evol Hum Behav* 28(3):151–158.
6. Herzog HA, Bentley RA, Hahn MW (2004) Random drift and large shifts in popularity of dog breeds. *Proc Biol Sci* 271(Suppl 5):S353–S356.
7. Neiman FD (1995) Stylistic variation in evolutionary perspective: Inferences from decorative diversity and interassemblage distance in Illinois woodland ceramic assemblages. *Am Antiq* •••:7–36.
8. Bentley RA (2008) Random drift versus selection in academic vocabulary: An evolutionary analysis of published keywords. *PLoS ONE* 3(8):e3057.
9. Krawczyk MJ, Dydejczyk A, Kulakowski K (2014) The Simmel effect and babies' names. *Physica A Stat Mech Appl* 395:384–391.
10. Xi N, et al. (2014) Cultural evolution: The case of babies' first names. *Physica A Stat Mech Appl* 406:139–144.
11. Berger J, Le Mens G (2009) How adoption speed affects the abandonment of cultural tastes. *Proc Natl Acad Sci USA* 106(20):8146–8150.
12. Goldstone RL, Gureckis TM (2009) Collective behavior. *Top Cogn Sci* 1(3):412–438.
13. Berger J, Bradlow ET, Braunstein A, Zhang Y (2012) From Karen to Katie: Using baby names to understand cultural evolution. *Psychol Sci* 23(10):1067–1073.
14. Jensen HJ (1998) *Self-Organized Criticality: Emergent Complex Behavior in Physical and Biological Systems* (Cambridge Univ Press, Cambridge, UK), Vol 10.
15. Albert R, Barabási AL (2002) Statistical mechanics of complex networks. *Rev Mod Phys* 74(1):47.
16. Feldman MW, Cavalli-Sforza LL (1981) Further remarks on Darwinian selection and "altruism". *Theor Popul Biol* 19(2):251–260.
17. Zanette DH (2012) Dynamics of fashion: The case of given names. arXiv:1208.0576.
18. Acerbi A, Ghirlanda S, Enquist M (2012) The logic of fashion cycles. *PLoS ONE* 7(3): e32541.
19. Onnela JP, Reed-Tsochas F (2010) Spontaneous emergence of social influence in online systems. *Proc Natl Acad Sci USA* 107(43):18375–18380.
20. Gleeson JP, Cellai D, Onnela JP, Porter MA, Reed-Tsochas F (2014) A simple generative model of collective online behavior. *Proc Natl Acad Sci USA* 111(29):10411–10415.
21. Wu L, Flammini A, Vespignani A, Menczer F (2012) Competition among memes in a world with limited attention. *Sci Rep* 2:335.
22. Wu F, Huberman BA (2007) Novelty and collective attention. *Proc Natl Acad Sci USA* 104(45):17599–17601.
23. Wang D, Song C, Barabási AL (2013) Quantifying long-term scientific impact. *Science* 342(6154):127–132.
24. Bentley RA, Ormerod P (2012) Accelerated innovation and increased spatial diversity of us popular culture. *Adv Complex Syst* 15:1150011.
25. Simmel G (1904) Fashion. *Am J Sociol* 62(6):541–558.
26. Hahn MW, Bentley RA (2003) Drift as a mechanism for cultural change: An example from baby names. *Proc Biol Sci* 270(Suppl 1):S120–S123.
27. US Social Security Administration, The Official Website of the US Social Security Administration. Available at www.ssa.gov.
28. Li W (2012) Analyses of baby name popularity distribution in us for the last 131 years. *Complexity* 18(1):44–50.
29. Sporns O, Chialvo DR, Kaiser M, Hilgetag CC (2004) Organization, development and function of complex brain networks. *Trends Cogn Sci* 8(9):418–425.
30. Luo F, Zhong J, Yang Y, Scheuermann RH, Zhou J (2006) Application of random matrix theory to biological networks. *Phys Lett A* 357(6):420–423.
31. Luo F, et al. (2007) Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. *BMC Bioinformatics* 8(1):299.
32. Jalan S, Solymosi N, Vattay G, Li B (2010) Random matrix analysis of localization properties of gene coexpression network. *Phys Rev E Stat Nonlin Soft Matter Phys* 81(4 Pt 2):046118.
33. Laloux L, Cizeau P, Bouchaud JP, Potters M (1999) Noise dressing of financial correlation matrices. *Phys Rev Lett* 83(7):1467.
34. Plerou V, Gopikrishnan P, Rosenow B, Amaral LAN, Stanley HE (1999) Universal and non-universal properties of cross correlations in financial time series. *Phys Rev Lett* 83(7):1471.
35. Plerou V, et al. (2002) Random matrix approach to cross correlations in financial data. *Phys Rev E Stat Nonlin Soft Matter Phys* 65(6 Pt 2):066126.
36. Conlon T, Ruskin HJ, Crane M (2009) Cross-correlation dynamics in financial time series. *Physica A Stat Mech Appl* 388(5):705–714.
37. Barthélemy M, Gondran B, Guichard E (2002) Large scale cross-correlations in Internet traffic. *Phys Rev E Stat Nonlin Soft Matter Phys* 66(5 Pt 2):056110.
38. Santhanam MS, Patra PK (2001) Statistics of atmospheric correlations. *Phys Rev E Stat Nonlin Soft Matter Phys* 64(1 Pt 2):016102.
39. Fryer RG, Levitt SD (2004) The causes and consequences of distinctively Black names. *The Quarterly Journal of Economics* 119(3):767–805.
40. Edelman A (1988) Eigenvalues and condition numbers of random matrices. *SIAM J Matrix Anal Appl* 9(4):543–560.
41. Cizeau P, Bouchaud JP (1994) Theory of Lévy matrices. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 50(3):1810–1822.
42. Johnstone IM (2007) High dimensional statistical inference and random matrices. *Proceedings of the International Congress of Mathematicians Madrid, August 22–30, 2006*, eds Sanz-Solé M, Soria J, Varona JL, Verdera J (EMS Publishing House, Zürich) pp 307–333.