JOURNAL OF THE ROYAL SOCIETY Interface

J. R. Soc. Interface (2012) **9**, 1956–1964
doi:10.1098/rsif.2011.0846
Published online 15 February 2012

Culturomics meets random fractal theory: insights into long-range correlations of social and natural phenomena over the past two centuries

Jianbo Gao[1,2,*], Jing Hu[3], Xiang Mao[4] and Matjaž Perc[5,*]

[1]*PMB Intelligence, LLC, West Lafayette, IN 47996, USA*
[2]*Department of Mechanical and Materials Engineering, Wright State University, Dayton, OH 45435, USA*
[3]*Affymetrix, Inc., 3380 Central Expressway, Santa Clara, CA 95051, USA*
[4]*Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, USA*
[5]*Department of Physics, Faculty of Natural Sciences and Mathematics, University of Maribor, Koroška cesta 160, SI-2000 Maribor, Slovenia*

Culturomics was recently introduced as the application of high-throughput data collection and analysis to the study of human culture. Here, we make use of these data by investigating fluctuations in yearly usage frequencies of specific words that describe social and natural phenomena, as derived from books that were published over the course of the past two centuries. We show that the determination of the Hurst parameter by means of fractal analysis provides fundamental insights into the nature of long-range correlations contained in the culturomic trajectories, and by doing so offers new interpretations as to what might be the main driving forces behind the examined phenomena. Quite remarkably, we find that social and natural phenomena are governed by fundamentally different processes. While natural phenomena have properties that are typical for processes with persistent long-range correlations, social phenomena are better described as non-stationary, on–off intermittent or Lévy walk processes.

Keywords: culturomics; random fractal theory; Hurst parameter; correlations; society; culture

1. INTRODUCTION

Observational data are often very complex, appearing without any structure or pattern in either time or space. Examples of such observations can be found across the whole spectrum of the social and natural sciences, ranging from economics [1] to physics [2], biology [3] and medicine [4]. The origins of observed irregular behaviour, however, are not always clear. Roughly, five decades ago, deterministic chaos was discovered [5] and quickly rose to prominence as a possible mechanism of inherent unpredictability and complexity [6,7]. Yet, the strict criteria for declaring deterministic chaos in observed data [8], most notably the satisfaction of criteria for stationarity and determinism [2], and the verification of exponential divergence [9,10], are rarely satisfied. In response, attention has begun to shift from chaos to noise and random processes as alternative [11] (or, in many cases, as even more probable) sources of irregularity. While the theory of deterministic chaos relies on nonlinear dynamical systems with typically only a few degrees of freedom, the analysis of stochastic processes, especially those that yield data with scale invariance, relies on random fractal theory [12] or its generalization, multifractal theory [9,13]. Indeed, investigations based on these theoretical foundations may provide an elegant statistical characterization of a broad range of heterogeneous phenomena [14], and in this paper, it is our goal to extend this theory to culturomics, as recently introduced in Michel *et al.* [15].

Culturomics, and the study of human culture in general, seemingly has little to do with deterministic chaos and fractals. However, quantitative analyses of various aspects of human culture have become increasingly popular; examples include the study of human mobility patterns [16–18], the spread of infectious diseases [19–22] and malware [23,24], the dynamics of online popularity [25], social movement [26] and language [27–29], and even tennis [30]. This progress is driven not only by important advances in theory and

*Authors for correspondence (jbgao.pmb@gmail.com; matjaz.perc@uni-mb.si).

Received 3 December 2011
Accepted 27 January 2012

modelling, but also by the increasing availability of vast amounts of data and knowledge, also referred to as metaknowledge [31], which allows scientists to apply advanced methods of analysis on a large scale [32]. The seminal study by Michel *et al.* [15] was accompanied by the release of a vast amount of data comprising metrics derived from approximately 4 per cent of books ever published (over five million in total), and it was this release that made the present study, i.e. the application of random fractal theory, possible. The data are available at ngrams.googlelabs.com as counts of n-grams that appeared in a given corpus of books published in each year. An n-gram is made up of a series of n 1-grams, and a 1-gram is a string of characters uninterrupted by a space. Note that a 1-gram is not necessarily a word, for it may be a number or a typo as well. Besides the counts of individual n-grams, the total counts of n-grams contained in each corpus of books in a given year are also provided, from which yearly usage frequencies can be obtained.

In this paper, we show what new insights are attainable by applying random fractal theory to this vast culturomic dataset. Our goal is to try and go beyond the interpretations of trajectories provided in Michel *et al.* [15] by means of an accurate determination of scaling parameters [33], and in particular, the Hurst parameter H, which enables us to characterize the nature of correlations (memory), if any, contained in the irregular time series. In general, data with long-range correlations are an important subclass of $1/f^{\alpha}$ noise [34–36], which is characterized by a power-law decaying power spectral density, and whose dimensionality cannot be reduced by principal component analysis since the rank-ordered eigenvalue spectrum also decays as a power law [37]. Processes that generate time series with such properties are said to have anti-persistent correlations if $0 < H < 1/2$, are memoryless or have only short-range correlations if $H = 1/2$, and have persistent long-range correlations (long memory) if $1/2 < H < 1$ [12]. Moreover, values of $H > 1$ are possible as well; these values, however, are characteristic of non-stationary processes or rather special stationary processes such as on–off intermittency with power-law distributed on and/or off periods and Lévy walks [10]. (Note that the latter should not be confused with Lévy flights, which are random processes consisting of many independent steps, and are thus memoryless with $H = 1/2$.) Prominent examples where $1/f^{\alpha}$ noise was recently observed and quantified include DNA sequences [38,39], human cognition [40] and coordination [41], posture [42], cardiac dynamics [43–46] and the distribution of prime numbers [47], to name but a few.

Despite the many successful attempts at assessing long-range correlations in complex time series—for example, by means of detrended fluctuation analysis (DFA) [48], as well as many other methods [9,13]—care should be exercised by their interpretation, particularly if one is faced with relatively short time series that contain trends [49], non-stationarity [50] or signs of rhythmic activity [51,52]. Although it is obviously impossible to make general statements concerning these properties for all the n-grams contained in the corpus of the over five million digitized books, which amount roughly to over two billion culturomic trajectories, it is clear that the time series are short, comprising a little more than approximately 200 points corresponding to the two centuries considered (more precisely, from year 1770 to 2007), and that many will inevitably contain strong trends [15]. In order to successfully surpass the difficulties and pitfalls associated with the analysis of such time series [10], besides the traditional DFA, we also use an adaptive fractal analysis (AFA), which is based on nonlinear adaptive multiscale decomposition. We use these methods to determine the Hurst parameter H for several 1-grams that are representative for social and natural phenomena. Examples of words that we focus on include war, unemployment, hurricane and earthquake, and we find that those that describe social phenomena (war, unemployment, etc.) in general have different scaling properties than those describing natural phenomena (hurricane, earthquake, etc.). Our results can be corroborated aptly with arguments from real life, and they fit nicely to the declared goal of culturomics, which is to extend the boundaries of scientific inquiry to a wide array of new phenomena [15].

The remainder of this paper is organized as follows. In the next section, we present the main results, in §3, we summarize them and discuss their potential implications, while in the appendix, we describe the details of fractal analysis.

2. RESULTS

We start by presenting the results of the AFA for natural phenomena. We first plot in figure 1*a* the original time series (thin line) and the estimated trend (thick line) for the 1-gram 'earthquake'. The detrended data are presented in figure 1*b*. It can be observed that overall the trend is very modest and simple, increasing only slightly towards the present day. Using equation (A 4), the Hurst parameter can be estimated from the slope of the $F(w)$ versus w dependence on a double log scale. In figure 1*c*, we show that the analysis of detrended data yields $H = 0.65$, while in figure 1*d* we show that $H = 0.75$ if the original data are used as input. Both calculations produce similar results, showing a very modest slope, and rely on statistically robust scaling. Based on the meaning of the Hurst parameter, the fractal analysis of the culturomic trajectory for 'earthquake' reveals that this phenomenon has persistent long-range correlation.

As another example, we show in figure 2 the same analysis for the 1-gram 'hurricane'. Unlike the 'earthquake' trajectory, the trend for 'hurricane' is more pronounced. It has a strong upwards component, especially in the last couple of decades. Hence, it can be expected that the discrepancy of the two estimated H values for the original and detrended data will be somewhat larger than that for the 1-gram 'earthquake' analysed in figure 1. This expectation is indeed confirmed by comparing figure 2*c* and figure 2*d*, from where it follows that for the detrended data $H = 0.70$ while for original data $H = 0.85$. Still, however, both results robustly classify 'hurricane' as a phenomenon with persistent long-range

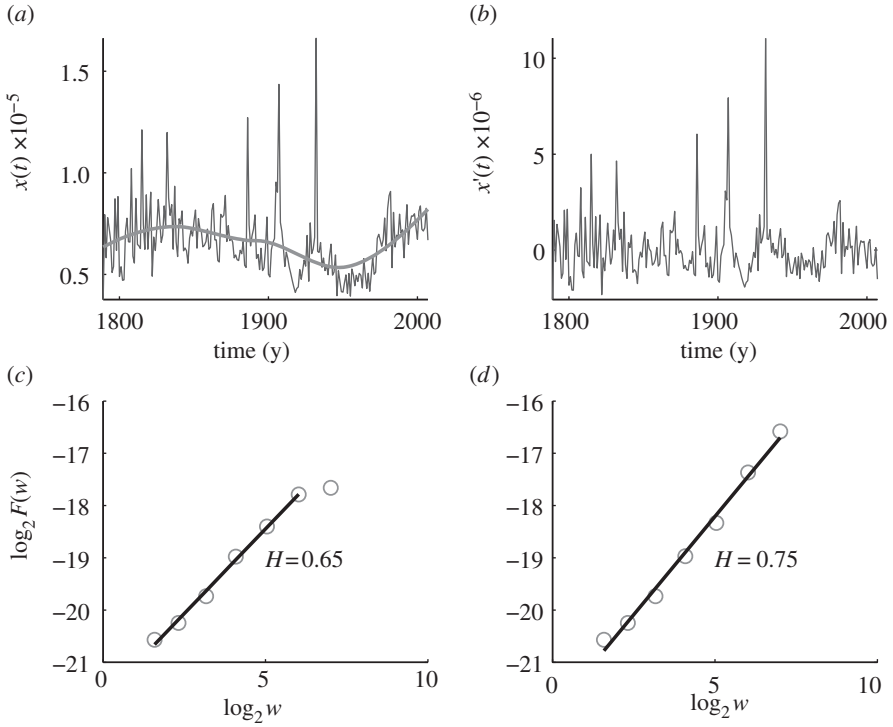

Figure 1. Adaptive fractal analysis (AFA) of the usage frequency of the 1-gram 'earthquake' in the corpus of English books. The Hurst parameter, as obtained from the detrended data, is $H = 0.65$. (*a*) The variation of the usage frequency of 'earthquake' with time. The thin line depicts original data, while the thick line depicts the estimated trend (using a window of length 101). (*b*) Detrended data, i.e. the difference between the thin and thick curves in (*a*). (*c*) Best fit to the $F(w)$ versus w dependence for detrended data on a double log scale yields $H = 0.65$. (*d*) Best fit to the $F(w)$ versus w dependence for original data on a double log scale yields $H = 0.75$.

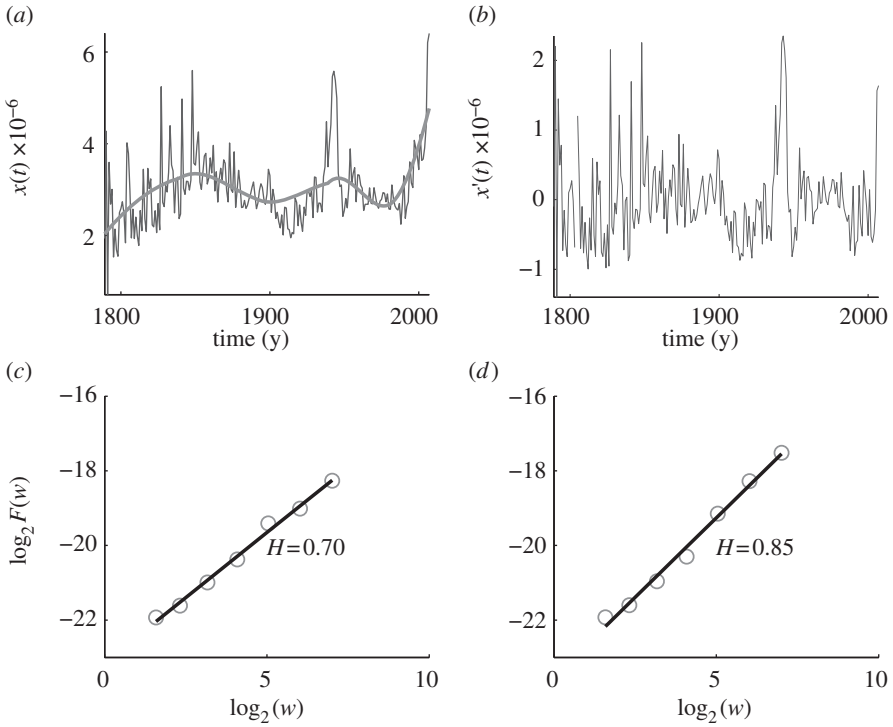

Figure 2. AFA of the usage frequency of the 1-gram 'hurricane' in the corpus of English books. The Hurst parameter, as obtained from the detrended data, is $H = 0.70$. (*a*) The variation of the usage frequency of 'hurricane' with time. The thin line depicts the original data, while the thick line depicts the estimated trend (using a window of length 101). (*b*) Detrended data, i.e. the difference between the thin and thick curves in (*a*). (*c*) Best fit to the $F(w)$ versus w dependence for detrended data on a double log scale yields $H = 0.70$. (*d*) Best fit to the $F(w)$ versus w dependence for original data on a double log scale yields $H = 0.85$.

Table 1. Hurst parameters H, as obtained for the detrended data of all 15 considered 1-grams describing natural phenomena. The left column lists results as obtained with the adaptive fractal analysis (AFA), while the right column lists results as obtained with the detrended fluctuation analysis (DFA). The range of values as obtained by AFA is $0.55 \leq H \leq 0.85$, with an average over all 15 considered 1-grams equalling $\overline{H} = 0.69$. With DFA, we obtain $0.41 \leq H \leq 0.85$ and $\overline{H} = 0.67$.

	Hurst parameter (H)	
1-grams	AFA	DFA
avalanche	0.63 ± 0.06	0.79 ± 0.06
comet	0.60 ± 0.03	0.73 ± 0.04
drought	0.81 ± 0.05	0.69 ± 0.09
earthquake	0.65 ± 0.02	0.72 ± 0.03
erosion	0.85 ± 0.06	0.86 ± 0.08
fire	0.67 ± 0.05	0.70 ± 0.03
flooding	0.85 ± 0.06	0.72 ± 0.08
hurricane	0.70 ± 0.03	0.69 ± 0.08
landslide	0.66 ± 0.05	0.41 ± 0.20
life	0.62 ± 0.03	0.65 ± 0.06
lightning	0.63 ± 0.03	0.70 ± 0.03
mudslide	0.80 ± 0.02	0.58 ± 0.28
tornado	0.59 ± 0.02	0.64 ± 0.06
tsunami	0.81 ± 0.05	0.66 ± 0.03
typhoon	0.55 ± 0.02	0.50 ± 0.09

correlations, thus adding to the evidence that this may be valid, in general, for natural phenomena.

To test this hypothesis more thoroughly, we have performed the same analysis as depicted in figures 1 and 2, along with the DFA, for 13 other phenomena that can be classified as characteristic of natural phenomena. Although there may be some disagreement as to what terms are *characteristic* of natural phenomenon, and other 1-grams as well as n-grams could be suggested as characteristic of natural phenomena and analysed, we consider our selection to be sufficiently representative for this study. Supporting this assumption are the results presented in table 1, which point robustly towards the conclusion that natural phenomena, in general, really can be classified as processes with persistent long-range correlations. More specifically, for detrended data, we find that all estimated Hurst parameters are within the $1/2 < H < 1$ range with an average of $\overline{H} = 0.69$ (AFA), which leads us to the mentioned final conclusion. Results obtained for original data (before detrending, not shown), on the other hand, leave a bit more room for discussion. There, for certain 1-grams, like 'mudslide' and 'flooding', the value of H is larger than one. This suggests that the data would be more appropriately described as being either non-stationary, on–off intermittent or Lévy walk-like. Such a discussion, however, would be to a large degree baseless as the upward trends occurring towards the present time in most n-grams describing natural phenomena must be properly taken into account. The observed trends may be considered as a straightforward consequence of the fact that we have more and more data readily available on natural phenomena, which is due to advancements in measuring techniques as well as the increasingly global reach of the Internet. Modern data collection and telecommunication technologies have raised our awareness, in general, of natural phenomena, and, as a result, it is reasonable to expect this increased awareness to be reflected in an increase of occurrences in the corpus. Note, however, that similar arguments can be raised for other fields and trivia (e.g. celebrity gossip, popular culture) as well, and thus one could argue that relatively, the usage frequencies should not necessarily increase as a result of that.

Turning to social phenomena, we will show that the problems discussed for natural phenomena are in some cases amplified, but, more importantly, that social phenomena, apart from rare exceptions, cannot be classified solely as processes with persistent long-range correlations.

First, we presented the AFA for the 1-gram 'war' in figure 3. The original data depicted by the thin line in figure 3*a* are clearly reminiscent of historical events, as World Wars I and II generate two large peaks that more or less dwarf the usage frequencies reported in other decades. This observation goes hand in hand not just with the magnitude of the two world wars, but also with the increase in the usage frequency of 'war' in the published literature at that time. In agreement with the historical events is the estimated trend line depicted by the thick line in figure 3*a*. However, even after the detrending, the resulting culturomic trajectory still clearly reflects history in that the periods of World Wars I and II stand out from the rest, as can be inferred from the curve depicted in figure 3*b*. The Hurst parameter H determined using the detrended and original data (presented in figure 3*c*,*d*) have similar values to each other ($H = 1.09$ for the detrended data and $H = 1.15$ for the original data). As a result, both classify 'war' as either a non-stationary, on–off intermittent or a Lévy walk-like process.

Another illustrative example of fractal analysis is presented in figure 4, where we examine the 1-gram 'unemployment'. A crucial distinction from 'war', as well as all the considered natural phenomena, is that unemployment was non-existent, or at least it was not mentioned, in the literature prior to 1900, which is clearly inferable from the original data depicted with a thin in figure 4*a*. With the coming of age of the industrial revolution, the job market began to take shape, and with it came, rather inevitably it seems, the problem of unemployment. The trend depicted with a thick line in figure 4*a* clearly captures this fact. Moreover, we note that the first broad peak in the plot starts at around 1930, and thus correlates well with the Great Depression, while the second broad peak starts at around 1970, and thus correlates with that period of US economic stagnation and high inflation that was linked with the Middle Eastern oil crisis. After detrending, the situation is of course only marginally improved (in terms of assuring a more stationary record), as can be concluded from the curve depicted in figure 4*b*. The Hurst parameters, equalling $H = 1.32$ for the detrended data (*c*) and $H = 1.39$ for the original data (*d*), both clearly reflect non-stationarity, and accordingly, 'unemployment' can be considered the result of such a process.

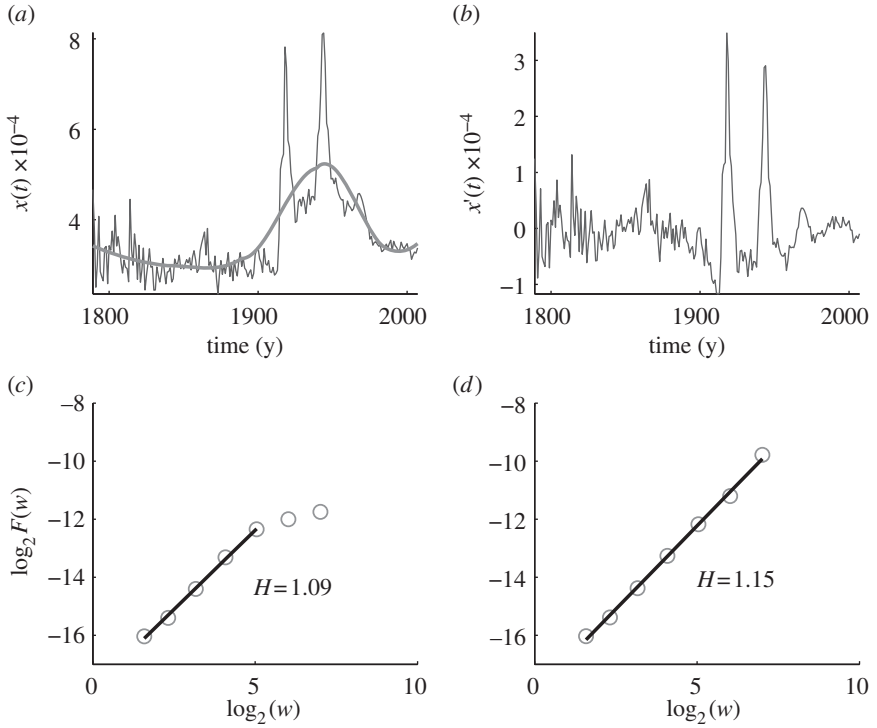

Figure 3. AFA of the usage frequency of the 1-gram 'war' in the corpus of English books. The Hurst parameter, as obtained from the detrended data, is $H = 1.09$. (*a*) The variation of the usage frequency of 'war' with time. The thin line depicts the original data, while the thick line depicts the estimated trend (using a window of length 101). (*b*) The detrended data, i.e. the difference between the thin and thick curves in (*a*). (*c*) Best fit to the $F(w)$ versus w dependence for detrended data on a double log scale yields $H = 1.09$. (*d*) Best fit to the $F(w)$ versus w dependence for original data on a double log scale yields $H = 1.15$.

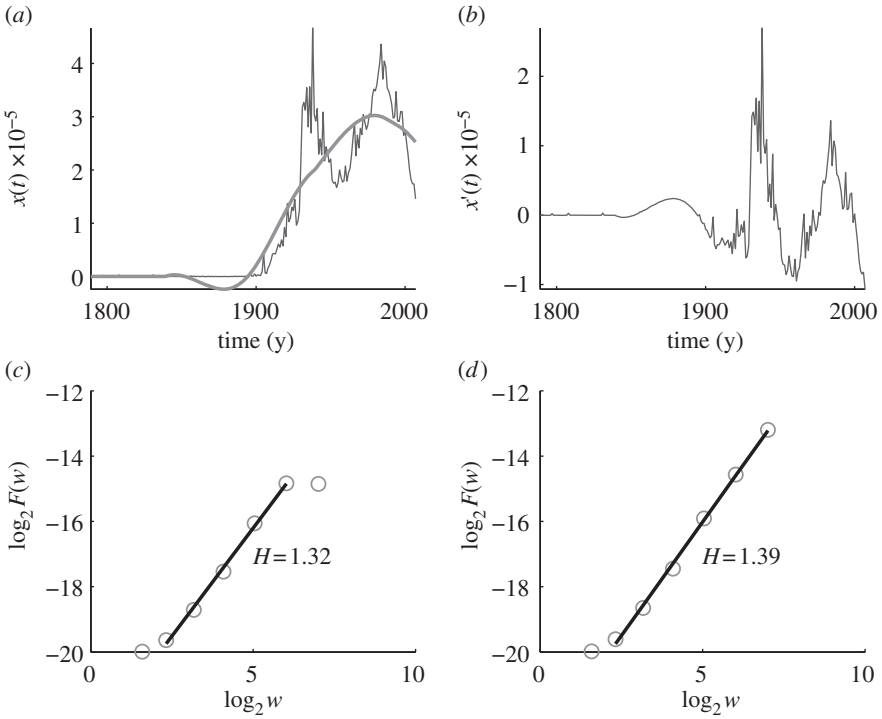

Figure 4. AFA of the usage frequency of the 1-gram 'unemployment' in the corpus of English books. The Hurst parameter, as obtained from the detrended data, is $H = 1.32$. (*a*) The variation of the usage frequency of 'unemployment' with time. The thin line depicts the original data, while the thick line depicts the estimated trend (using a window of length 101). (*b*) Detrended data, i.e. the difference between the thin and thick curves in (*a*). (*c*) Best fit to the $F(w)$ versus w dependence for detrended data on a double log scale yields $H = 1.32$. (*d*) Best fit to the $F(w)$ versus w dependence for original data on a double log scale yields $H = 1.39$.

Table 2. Hurst parameters H, as obtained for the detrended data of all 15 considered 1-grams describing social phenomena. The left column lists results as obtained with the AFA, while the right column lists results as obtained with the DFA. The range of values as obtained by AFA is $0.74 \leq H \leq 1.33$, with the average over all 15 considered 1-grams equalling $\overline{H} = 1.11$. With DFA, we obtain $0.66 \leq H \leq 1.44$ and $\overline{H} = 1.08$.

	Hurst parameter (H)	
1-grams	AFA	DFA
Christian	0.85 ± 0.05	0.95 ± 0.08
communism	1.32 ± 0.04	1.44 ± 0.05
crisis	1.15 ± 0.05	1.13 ± 0.08
democracy	1.18 ± 0.02	1.07 ± 0.07
education	1.04 ± 0.05	1.09 ± 0.13
environment	1.13 ± 0.04	1.24 ± 0.08
famine	0.74 ± 0.02	0.66 ± 0.06
malnutrition	1.10 ± 0.07	1.08 ± 0.11
politics	1.14 ± 0.03	0.99 ± 0.06
population	1.01 ± 0.06	0.98 ± 0.10
recession	1.33 ± 0.05	1.06 ± 0.07
socializing	1.28 ± 0.07	1.28 ± 0.09
stock	1.01 ± 0.05	0.99 ± 0.11
unemployment	1.32 ± 0.04	1.28 ± 0.04
war	1.09 ± 0.03	0.99 ± 0.12

As in the case with natural phenomena (table 1), we also performed the same fractal analysis as in 'war' and 'unemployment', along with the DFA, for 13 other social phenomena. The results are presented in table 2. It can be observed that the large majority of considered 1-grams have $H > 1$ (AFA), which indicates that social phenomena are most likely to be either non-stationary, on–off intermittent or Lévy walk-like process. This conclusion is obtained irrespective of whether detrending is performed or not, although the average Hurst parameter for detrended data, equalling $\overline{H} = 1.11$ (AFA), is smaller than that obtained for original data (before detrending, not shown), which is $\overline{H} = 1.26$. This technical discrepancy, however, is probably due to the successful removal of some level of non-stationarity that is in general characteristic of social phenomena (more so than of natural phenomena). We would like to note, however, that in general not all $H > 1$ occurrences should be, by default, attributed to non-stationarity in the trajectories. While visual inspection may lend support to such a conclusion, as was the case for results presented in figure 4, in general the H value alone cannot distinguish between non-stationary, on–off intermittent or Lévy walk-like processes. In fact, the time series are too short for a robust assessment of a more precise nature of the examined social phenomena. At a glance, and since this is indeed most common, it seems convenient to attribute $H > 1$ in social phenomena to non-stationarity, yet only additional future data can enable us to differentiate whether the peaks are part of an on–off intermittent process with power-law distributed on and/or off events, or if they are part of a Lévy walk. Finally, we would also like to point out that of course not *all* phenomena that can be considered as social will have $H > 1$. Examples include 1-grams such as 'famine' or 'Christian', which for the largest parts of the recorded human history were either directly related to natural phenomena (severe droughts, flooding or other phenomena negatively affected that season's yield on vegetables, crops, grass and animal population, hence leading to famine) or have been an integral part of the human culture for a long time (prior to the start of the culturomic trajectories). Moreover, social topics that are of little interest will not garner much attention, and are as such also unlikely to have usage frequencies with $H > 1$. The social phenomena where the human factor has played a key role recently and which are reasonably popular, however, all share features that are characteristic of processes with $H > 1$. In fact, it seems just to conclude that the more the social phenomena can be considered recent (unemployment, recession and democracy), the higher their Hurst parameter is likely to be (table 2). This agrees nicely also with the recent observation of bursts and heavy tails in human dynamics [53].

3. DISCUSSION

By applying fractal analysis based on DFA and AFA to culturomic trajectories of 1-grams describing typical social and natural phenomena over the past two centuries, we have found that they obey different scaling laws. As we will discuss in what follows, our findings agree nicely with existing theory and expectations, as well as offer new interpretations as to what might be the main driving forces behind the examined phenomena.

We find that natural phenomena have properties that are typical of processes that generate persistent long-range correlations, as evidenced by the Hurst parameter being in the range $0.55 \leq H \leq 0.85$, with an average over all 15 considered 1-grams equalling $\overline{H} = 0.69$ (AFA). The prevalence of long-term memory in natural phenomena compels us to conjecture that the long-range correlations in the usage frequency of the corresponding terms is predominantly driven by occurrences in nature of those phenomena. Using data from five million digitized books to arrive at this understanding certainly supports the declared goal of culturomics and lends strong support to its core principles. Owing to this memory, and of course by using statistical data available, we know, based on the Gutenberg–Richter law [54], that in the UK, for example, an earthquake of 3.7–4.6 on the Richter scale is likely to happen every year, an earthquake of 4.7–5.5 is due approximately every 10 years, while an earthquake of 5.6 or larger is bound to happen every 100 years [55]. Similar 'statistical predictions' are available for tsunamis and many other, if not all, natural phenomena. On a more personal level, this also agrees with how we naturally develop an understanding for the weather and related natural phenomena for the region we live in.

Social phenomena, on the other hand, have the Hurst parameter in the range $0.74 \leq H \leq 1.33$, with an average over all 15 considered 1-grams equalling $\overline{H} = 1.11$ (AFA). This is indicative of non-stationary processes, or stationary processes like on–off intermittency with

power-law distributed on and/or off periods or Lévy walks. While our analysis does not allow distinction between these three options, it is clear that all these processes are fundamentally different from those describing natural phenomena. So while it is common to hear speculations about possible average periods regarding social phenomena—for instance, that there may be an average period between major wars or stock market crashes—our analysis suggests this is not the case, and that social phenomena tend to follow different scaling laws from natural phenomena. Such a difference is not unexpected, as social phenomena are, by nature, more complex than natural phenomena; the former depend on political, economic and social forces, as well as on natural phenomena. The results of this additional complexity can be seen in our fractal analysis of a set of culturomic trajectories.

In summary, we hope to have successfully demonstrated that the data made available through the Culturomics project [15], when coupled with advanced methods of analysis, offer fascinating opportunities to explore human culture in the broadest possible sense.

This research was supported in part by the US NSF grant CMMI-0825311 to Jianbo Gao, and by the Slovenian Research Agency's grant J1-4055 to Matjaž Perc. The authors are grateful to Prof. Johnny Lin of North Park University for many useful discussions.

APPENDIX A. METHODS

Nonlinear adaptive multiscale decomposition starts by partitioning a time series into segments of length $w = 2n + 1$, where neighbouring segments overlap by $n + 1$ points, thus introducing a time scale of $((w + 1)/2)\tau = (n + 1)\tau$, where τ is the sampling time. Each segment is then fitted with the best polynomial of order M. Note that $M = 0$ and 1 correspond to piece-wise constant and linear fitting, respectively. We denote the fitted polynomials for the ith and $(i + 1)$th segments by $y^{(i)}(l_1)$ and $y(i + 1)(l_2)$, respectively, where $l_1, l_2 = 1, \ldots, 2n + 1$. We then define the fitting for the overlapped region as

$$y^{(c)}(l) = w_1 y^{(i)}(l + n) + w_2 y^{(i+1)}(l), \quad l = 1, 2, \ldots, n + 1, \tag{A 1}$$

where $w_1 = (1 - (l - 1)/n)$ and $w_2 = (l - 1)/n$ can be written as $(1 - d_j/n)$ for $j = 1$, 2, and where d_j denotes the distances between the point and the centres of $y^{(i)}$ and $y^{(i+1)}$, respectively. This means that the weights decrease linearly with the distance between the point and the centre of the segment. Such a weighting ensures symmetry and effectively eliminates any jumps or discontinuities around the boundaries of neighbouring segments. In fact, the scheme ensures that the fitting is continuous everywhere, is smooth at the non-boundary points, and has the right- and left-derivatives at the boundary. Moreover, since it can deal with an arbitrary trend without *a priori* knowledge, it can remove non-stationarity, including baseline drifts and motion artefacts [56], and the procedure may also be used as either high-pass or low-pass filter with superior noise-removal properties than linear filters, wavelet shrinkage or chaos-based noise reduction schemes [57].

Based on the described adaptive decomposition, a fractal analysis can be conducted as follows. Let $\{x_1, x_2, \ldots, x_n\}$ be a stationary stochastic process with mean $\bar{x}$ and autocorrelation function of type

$$r(k) \sim k^{2H-2} \quad \text{as } k \to \infty, \tag{A 2}$$

where H is the Hurst parameter. This is often called an increment process, and its power spectral density is $1/f^{2H-1}$. The integral of the increment process

$$u(i) = \sum_{k=1}^{i} (x_k - \bar{x}), \quad i = 1, 2, \ldots, n, \tag{A 3}$$

on the other hand, is called a random walk process, and its power spectral density is $1/f^{2H+1}$. Starting from an increment process, similarly to DFA [48], we first construct a random walk process using equation (A 3). If, however, the original data can already be classified as a random walk-like process, then this step is not necessary, although for ideal fractal processes, there is no penalty even if this step is done. Next, for a window size w, we determine, for the random walk process $u(i)$ (or the original process if it is already a random walk-like process), a global trend $v(i)$, $i = 1$, $2, \ldots, N$, where N is the length of the walk. The residual, $u(i) - v(i)$, characterizes fluctuations around the global trend, and its variance yields the Hurst parameter H according to

$$F(w) = \left[\frac{1}{N}\sum_{i=1}^{N} (u(i) - v(i))^2\right]^{1/2} \sim w^H. \tag{A 4}$$

The validity of equation (A 4) can be proved if one starts from an increment process with the Hurst parameter equal to H. Using Parseval's theorem [9], the variance of the residual data corresponding to a window size w may be equated to the total power in the frequency range (f_w, f_{cutoff}) as

$$\text{total power} \sim \int_{f_w}^{f_{cutoff}} \frac{1}{f^{2H+1}}\,\mathrm{d}f \sim \frac{1}{2H}(w^{2H} - f_{cutoff}^{-2H}), \tag{A 5}$$

where $f_w = 1/w$, and f_{cutoff} is the highest frequency of the data. When $f_w \ll f_{cutoff}$, we see that equation (A 4) has to be valid. In fact, the above treatment makes it clear that even if we start from a random walk process with the Hurst exponent equal to H, integration will give the process a spectrum of $1/f^{2H+1+2} = 1/f^{2(H+1)+1}$, and, therefore, the final Hurst parameter will be simply $H + 1$. This in turn indicates that there is indeed no penalty if one uses equation (A 3) when the data are already a random walk-like process. Note that the proposed approach, if needed, can be readily extended and applied successfully to multifractal as well as higher dimensional data.

The described fractal analysis approach, which we will refer to as AFA, in general yields results that are consistent with the traditionally used DFA [48], as can be concluded from results presented in figure 5. Nevertheless, especially for processes having $H > 1$ [10], AFA may yield better scaling, which is why, although we

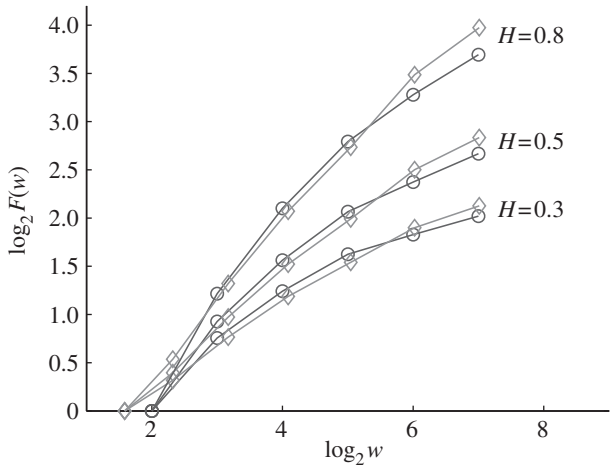

Figure 5. Scaling analysis of fractional Gaussian noise processes of the same length as the 1-gram data (240 points). Circles and diamonds depict results as obtained by means of the DFA and AFA, respectively, for three different values of H. It can be observed that both methods yield consistent results, regardless of the shortage of the examined time series.

analyse the culturomic trajectories with both methods, we rely on the results of AFA for final interpretation. The potential advantage of AFA over DFA is due to the fact that the trend for each window of size w obtained by AFA is smooth, while that obtained by DFA may change abruptly at the boundary of neighbouring segments. For short non-stationary time series, this may prove favourable for obtaining better scaling in the $F(w)$ versus w dependence.

REFERENCES

1 Mantegna, R. N. & Stanley, H. E. 2000 *Introduction to econophysics: correlations and complexity in finance*. Cambridge, UK: Cambridge University Press.
2 Kantz, H. & Schreiber, T. 1998 *Nonlinear time series analysis*. Cambridge, UK: Cambridge University Press.
3 Glass, L. & Mackey, M. C. 1988 *From clocks to chaos: the rhythms of life*. Princeton, NJ: Princeton University Press.
4 Goldberger, A. L. *et al.* 2000 Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation* **101**, 215.
5 Lorenz, E. N. 1963 Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**, 130–141. (doi:10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2)
6 Crutchfield, J. P., Farmer, J. D. & Huberman, B. A. 1982 Fluctuations and simple chaotic dynamics. *Phys. Rep.* **92**, 45–82. (doi:10.1016/0370-1573(82)90089-8)
7 Eckmann, J. P. & Ruelle, D. 1985 Ergodic theory of strange attractors. *Rev. Mod. Phys.* **57**, 617. (doi:10.1103/RevModPhys.57.617)
8 Abarbanel, H. D. I. 1996 *Analysis of observed chaotic data*. New York, NY: Springer.
9 Gao, J. B., Cao, Y. H., Tung, W. W. & Hu, J. 2007 *Multiscale analysis of complex time series: integration of chaos and random fractal theory, and beyond*. Hoboken, NJ: Wiley-Interscience.
10 Gao, J. B., Hu, J., Tung, W. W., Cao, Y. H., Sarshar, N. & Roychowdhury, V. P. 2006 Assessment of long-range correlation in time series: how to avoid pitfalls. *Phys. Rev. E* **73**, 016117. (doi:10.1103/PhysRevE.73.016117)
11 Stratonovich, R. L. 1963 *Topics in the theory of random noise*. New York, NY: Gordon and Breach.
12 Mandelbrot, B. B. 1982 *The fractal geometry of nature*. San Francisco, CA: Freeman.
13 Bunde, A. & Havlin, S. 1996 *Fractals and disordered systems*. New York, NY: Springer.
14 Stanley, H. E. & Meakin, P. 1988 Multifractal phenomena in physics and chemistry. *Nature* **335**, 405–409. (doi:10.1038/335405a0)
15 Michel, J. B. 2011 Quantitative analysis of culture using millions of digitized books. *Science* **331**, 176–182. (doi:10.1126/science.1199644)
16 González, M. C., Hidalgo, C. A. & Barabási, A.-L. 2008 Understanding individual human mobility patterns. *Nature* **453**, 779–782. (doi:10.1038/nature06958)
17 Song, C., Koren, T., Wang, P. & Barabási, A.-L. 2010 Modelling the scaling properties of human mobility. *Nat. Phys.* **6**, 818–823. (doi:10.1038/nphys1760)
18 Song, C., Qu, Z., Blumm, N. & Barabási, A.-L. 2010 Limits of predictability in human mobility. *Science* **327**, 1018–1021. (doi:10.1126/science.1177170)
19 Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J. & Vespignani, A. 2009 Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl Acad. Sci. USA* **106**, 21 484–21 489. (doi:10.1073/pnas.0906910106)
20 Meloni, S., Arenas, A. & Moreno, Y. 2009 Traffic-driven epidemic spreading in finite-size scale-free networks. *Proc. Natl Acad. Sci. USA* **106**, 16 897–16 902. (doi:10.1073/pnas.0907121106)
21 Sanz, J., Floría, L. M. & Moreno, Y. 2010 Spreading of persistent infections in heterogeneous populations. *Phys. Rev. E* **81**, 056108. (doi:10.1103/PhysRevE.81.056108)
22 Meloni, S., Perra, N., Arenas, A., Gómez, S., Moreno, Y. & Vespignani, A. 2011 Modeling human mobility responses to the large-scale spreading of infectious diseases. *PLoS ONE* **1**, 62.
23 Hu, H., Myers, S., Colizza, V. & Vespignani, A. 2009 WiFi networks and malware epidemiology. *Proc. Natl Acad. Sci. USA* **106**, 1318–1323. (doi:10.1073/pnas.0811973106)
24 Wang, P., González, M., Hidalgo, C. A. & Barabási, A. L. 2009 Understanding the spreading patterns of mobile phone viruses. *Science* **324**, 1071–1076. (doi:10.1126/science.1167053)
25 Ratkiewicz, J., Fortunato, S., Flammini, A., Menczer, F. & Vespignani, A. 2010 Characterizing and modeling the dynamics of online popularity. *Phys. Rev. Lett.* **105**, 158701. (doi:10.1103/PhysRevLett.105.158701)
26 Borge-Holthoefer, J. *et al.* 2011 Structural and dynamical patterns on online social networks: the Spanish May 15th Movement as a case study. *PLoS ONE* **6**, e23883. (doi:10.1371/journal.pone.0023883)
27 Lieberman, E., Michel, J. B., Jackson, J., Tang, T. & Nowak, M. A. 2007 Quantifying the evolutionary dynamics of language. *Nature* **449**, 713–716. (doi:10.1038/nature06137)
28 Puglisi, A., Baronchelli, A. & Loreto, V. 2008 Cultural route to the emergence of linguistic categories. *Proc. Natl Acad. Sci. USA* **105**, 7936–7940. (doi:10.1073/pnas.0802485105)
29 Loreto, V., Baronchelli, A., Mukherjee, A., Puglisi, A. & Tria, F. 2011 Statistical physics of language dynamics. *J. Stat. Mech.* P04006. (doi:10.1088/1742-5468/2011/04/P04006)
30 Radicchi, F. 2011 Who is the best player ever? A complex network analysis of the history of professional tennis. *PLoS ONE* **6**, e17249. (doi:10.1371/journal.pone.0017249)
31 Evans, J. A. & Foster, J. G. 2011 Metaknowledge. *Science* **331**, 721–725. (doi:10.1126/science.1201765)
32 Lazer, D. *et al.* 2009 Computational social science. *Science* **323**, 721–723. (doi:10.1126/science.1167742)

33 Stanley, H. E. 1987 *Introduction to phase transitions and critical phenomena*. Oxford, UK: Oxford University Press.
34 Press, W. H. 1978 Flicker noises in astronomy and elsewhere. *Comments Astrophys.* **7**, 103.
35 Bak, P., Tang, C. & Wiesenfeld, K. 1987 Self-organized criticality: an explanation of $1/f$ noise. *Phys. Rev. Lett.* **59**, 381–384. (doi:10.1103/PhysRevLett.59.381)
36 Bak, P. 1996 *How nature works: the science of self-organized criticality*. New York, NY: Copernicus.
37 Gao, J. B., Cao, Y. H. & Lee, J. M. 2003 Principal component analysis of $1/f$ noise. *Phys. Lett. A* **314**, 392–400. (doi:10.1016/S0375-9601(03)00938-1)
38 Voss, R. F. 1992 Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences. *Phys. Rev. Lett.* **68**, 3805–3808. (doi:10.1103/PhysRevLett.68.3805)
39 Peng, C. K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M. & Stanley, H. E. 1992 Long-range correlations in nucleotide sequences. *Nature* **356**, 168–170. (doi:10.1038/356168a0)
40 Gilden, D. L., Thornton, T. & Mallon, M. W. 19951/f noise in human cognition. *Science* **267**, 1837–1839. (doi:10.1126/science.7892611)
41 Chen, Y., Ding, M. & Scott Kelso, J. A. 1997 Long memory processes ($1/f^{\alpha}$ type) in human coordination. *Phys. Rev. Lett.* **79**, 4501–4504. (doi:10.1103/PhysRev Lett.79.4501)
42 Collins, J. J. & De Luca, C. J. 1994 Random walking during quiet standing. *Phys. Rev. Lett.* **73**, 764–767. (doi:10.1103/PhysRevLett.73.764)
43 Ivanov, P. Ch., Rosenblum, M. G., Peng, C. K., Mietus, J., Havlin, S., Stanley, H. E. & Goldberger, A. L. 1996 Scaling behaviour of heartbeat intervals obtained by wavelet-based time-series analysis. *Nature* **383**, 323–327. (doi:10.1038/383323a0)
44 Amaral, L. A. N., Goldberger, A. L., Ivanov, P. Ch. & Stanley, H. E. 1998 Scale-independent measures and pathologic cardiac dynamics. *Phys. Rev. Lett.* **81**, 2388–2391. (doi:10.1103/PhysRevLett.81.2388)
45 Ivanov, P. Ch., Rosenblum, M. G., Amaral, L. A. N., Struzik, Z. R., Havlin, S., Goldberger, A. L. & Stanley, H. E. 1999 Multifractality in human heartbeat dynamics. *Nature* **399**, 461–465. (doi:10.1038/20924)
46 Bernaola-Galvan, P., Ivanov, P. Ch., Amaral, L. A. N. & Stanley, H. E. 2001 Scale invariance in the nonstationarity of human heart rate. *Phys. Rev. Lett.* **87**, 168105. (doi:10.1103/PhysRevLett.87.168105)
47 Wolf, M. 1997 $1/f$ noise in the distribution of prime numbers. *Phys. A* **241**, 493–499. (doi:10.1016/S0378-4371(97)00251-3)
48 Peng, C. K., Buldyrev, S. V., Havlin, S., Simons, M. & Stanley, H. E. 1994 Mosaic organization of DNA nucleotides. *Phys. Rev. E* **49**, 1685–1689. (doi:10.1103/PhysRevE.49.1685)
49 Hu, K., Ivanov, P. Ch., Chen, Z., Carpena, P. & Stanley, H. E. 2001 Effect of trends on detrended fluctuation analysis. *Phys. Rev. E* **64**, 011114. (doi:10.1103/PhysRevE.64.011114)
50 Stanley, H. E., Kantelhardt, J. W., Zschiegner, S. A., Koscielny-Bunde, E., Havlin, S. & Bunde, A. 2002 Multifractal detrended fluctuation analysis of nonstationary time series. *Phys. A* **316**, 87–114. (doi:10.1016/S0378-4371(02)01383-3)
51 Chen, Z., Hu, K., Carpena, P., Bernaola-Galvan, P., Stanley, H. E. & Ivanov, P. Ch. 2005 Effect of nonlinear filters on detrended fluctuation analysis. *Phys. Rev. E* **71**, 011104. (doi:10.1103/PhysRevE.71.011104)
52 Hu, J., Gao, J. & Wang, X. 2009 Multifractal analysis of sunspot time series: the effects of the 11-year cycle and Fourier truncation. *J. Stat. Mech.* P02066. (doi:10.1088/1742-5468/2009/02/P02066)
53 Barabási, A. L. 2005 The origin of bursts and heavy tails in humans dynamics. *Nature* **435**, 207–211. (doi:10.1038/nature03459)
54 Gutenberg, B. & Richter, C. 1954 *Seismicity of the earth and associated phenomena*. Princeton, NJ: Princeton University Press.
55 Musson, R. 2011 Seismicity and earthquake hazard in the UK. See http://www.quakes.bgs.ac.uk/ (retrieved 23 October 2011).
56 Gao, J. B., Hu, J. & Tung, W. W. 2011 Facilitating joint chaos and fractal analysis of biosignals through nonlinear adaptive filtering. *PLoS ONE* **6**, e24331. (doi:10.1371/journal.pone.0024331)
57 Tung, W. W., Gao, J., Hu, J. & Yang, L. 2011 Recovering chaotic signals in heavy noise environments. *Phys. Rev. E* **83**, 046210. (doi:10.1103/PhysRevE.83.046210)