Authors for correspondence:
Klaus Desmet
e-mail: kdesmet@smu.edu
Iyad Rahwan
e-mail: rahwan@mpib-berlin.mpg.de
Ángel Cuevas
e-mail: acrumin@it.uc3m.es

THE ROYAL SOCIETY
PUBLISHING

Expanding the measurement of culture with a sample of two billion humans

Nick Obradovich[1], Ömer Özak[2,3,4], Ignacio Martín[5,6], Ignacio Ortuño-Ortín[7], Edmond Awad[8], Manuel Cebrián[1], Rubén Cuevas[9], Klaus Desmet[10,11,12], Iyad Rahwan[1] and Ángel Cuevas[9]

[1]Center for Humans and Machines, Max Planck Institute for Human Development, 14195 Berlin, Germany
[2]Department of Economics and Center for Scientific Computing, Southern Methodist University, Dallas, TX 75205, USA
[3]IZA, Institute of Labor Economics, 53113 Bonn, Germany
[4]GLO, Global Labor Organization, 45141 Essen, Germany
[5]Nommon Solutions and Technologies, 28020 Madrid, Spain
[6]Department of Telematic Engineering, Universidad Carlos III de Madrid, 28911 Leganés (Madrid), spain
[7]Department of Economics,Universidad Carlos III de Madrid, 28903 Getafe (Madrid), Spain
[8]Department of Economics, University of Exeter Business School, Exeter EX4 4PU, UK
[9] Department of Telematic Engineering and UC3M-Santander Big Data Institute, Universidad Carlos III de Madrid, 28911 Leganés (Madrid), Germany
[10]Department of Economics and Cox School of Business, Southern Methodist University, Dallas, TX 75205, USA
[11]NBER, National Bureau of Economic Research, Cambridge, MA 02138, USA
[12]CEPR, Centre for Economic Policy Research, London EC1V 0DX, UK

IR, 0000-0002-1796-4303; ÁC, 0000-0002-5738-0820

Culture has played a pivotal role in human evolution. Yet, the ability of social scientists to study culture is limited by the currently available measurement instruments. Scholars of culture must regularly choose between scalable but sparse survey-based methods or restricted but rich ethnographic methods. Here, we demonstrate that massive online social networks can advance the study of human culture by providing quantitative, scalable and high-resolution measurement of behaviourally revealed cultural values and preferences. We employ data across nearly 60 000 topic dimensions drawn from two billion Facebook users across 225 countries and territories. We first validate that cultural distances calculated from this measurement instrument correspond to traditional survey-based and objective measures of cross-national cultural differences. We then demonstrate that this expanded measure enables rich insight into the cultural landscape globally at previously impossible resolution. We analyse the importance of national borders in shaping culture and compare subnational divisiveness with gender divisiveness across countries. Our measure enables detailed investigation into the geopolitical stability of countries, social cleavages within small- and large-scale human groups, the integration of migrant populations and the disaffection of certain population groups from the political process, among myriad other potential future applications.

1. Introduction

Culture has played a pivotal role in human evolution [1–4]. As a result, the study of human culture is one of the core endeavours of the social sciences. Tens of thousands of scientists around the world study culture [5], with disciplines ranging from anthropology [6,7] to sociology [8], from political science [9–11] to economics [12–14], and from psychology [15–19] to philosophy [20]. Their work has enabled the understanding of many human social, economic and political phenomena [12,21–28], and serves as a bedrock of knowledge in the social sciences.

We focus on a comprehensive and holistic concept of culture. Following a commonly used definition in the field of cultural evolution, we say that

'[c]ulture in its broadest sense is that which is socially rather than genetically transmitted. [· · ·] In its totality, it is that which distinguishes one human group from another' [29]. This is similar to a concept of culture defined as 'that complex whole which includes knowledge, beliefs, arts, morals, law, customs, and any other capabilities and habits acquired by [a human] as a member of society' [30]. Our definition is inclusive, making no explicit value judgments about which traits might be more or less significant.

The traditional quantitative approach to the study of culture has been shaped by the historical availability of data [31], often limiting the ability to measure culture comprehensively. For most of the history of the study of culture, collecting empirical data on humans has been costly, time consuming and in many instances impossible [32]. As a result, scholars often inductively distill the broad theoretical definitions of culture using a top-down approach [31], in the sense that the scholars themselves design the surveys that measure the set of cultural constructs they are interested in. Cultural dimensions that result from this process tend to be a select few salient and measurable features of human groups such as artistic and culinary practices [33], language [34,35] and literature [36], political ideologies [37] and institutions [23], and religions and religious practices [38].

Traditionally, producing a more comprehensive description of a group's culture required ethnographers to observe individuals in the field [39]. Their approach is more bottom-up, in the sense that ethnographers spend long periods of time observing populations, with the aim of describing their culture holistically. More recently, the information age has enabled the emergence of what we might call computational ethnography [32,40]. Humans today spend an ever-increasing amount of time on devices that continuously track and record users' interests, beliefs, preferences, behaviours, locations and interactions. By unobtrusively observing billions of users, social media firms play the role of ethnographers, but on a massive scale.

This changing information environment enables supplementing and expanding the scientific approach to the measurement of culture. Here, we propose new methods for the bottom-up measurement of culture globally. We first explore the strengths and weaknesses of traditional quantitative approaches to the measurement of culture and show their relation to our method. We then measure culture from the bottom up and examine our measure's performance against traditional quantitative measures. Finally, we investigate cultural questions previously impossible to examine without the combined resolution and scope that our method enables. Ultimately, high-resolution granular data are essential for our understanding of many cultural phenomena, ranging from wars and the formation of identity to the integration of immigrants and the fragmentation of societies.

2. Measuring culture from the bottom up

2.1. Traditional quantitative approach to study culture

Traditional quantitative approaches to the study of culture benefit from numerous strengths. For example, these studies of culture are often relatively low in dimension and therefore readily measurable via quantitative surveys. This parsimony has enabled excellent studies of certain cultural features in highly data-constrained settings [41,42]. Traditional approaches

also provide substantial face validity: they focus on many of the concepts typically associated with culture. These traditional approaches, as a result, likely encapsulate many important constituent cultural constructs. A final benefit of the parsimony of traditional approaches is that scholars from a wide variety of fields can measure and study different aspects of culture.

Yet while the traditional approach to culture has a number of benefits, it also has various weaknesses. First, a critical question when attempting to construct a broad measure of human culture from the top down is: where do we draw the line in terms of what to measure? Surely religion should be considered part of culture. But is group-level support for a football team part of culture? What about preferences for video games [43] or television shows? What about group-level appreciation of cat videos or the colour of socks that we choose to wear? A scholar may determine religion—but not preference for a television show—to be the more theoretically important constituent of culture to measure based on the argument that television preferences are not central to the human experience. However, another scholar with a differing opinion could argue that, given the amount of time humans in the developed world spend watching television [44], the choice of a particular show reflects an important implicit value of the humans who watch it. If a line about what to measure must be drawn from the top down, scholars will very likely reasonably disagree on where, precisely, it should fall. However, the very attempt to narrow cultural constructs from the top down highlights the second weakness of this measurement strategy. The justification for excluding any particular construct is necessarily endogenous to the particular culture (and cultural bias) of the scientist(s) doing the theorizing. Crucially, where to draw this line is a direct function of the cultural preferences of each scholar's particular human group. This endogeneity problem arises for every supervised attempt to include or exclude a concept from the measurement of culture.

Third, the parsimonious nature of top-down approaches presents its own limitations. Implicitly, traditional quantitative approaches to the study of culture tend to focus on features that provide insight into differences among human groups [45–51]. Yet, human groups may be similar in many more dimensions than they are dissimilar. Top-down definitions tend to occlude these dimensions of similarity with an implicit focus on those features—the arts, language, politics, religion and distinct traditions—that differentiate human groups.

Finally, traditional quantitative measures of culture commonly rely on self-reported answers to survey questions [52] or subjective evaluations of the particular scientists conducting the study [53]. These pose measurement challenges. Questions such as: 'Is religion important to you?' or 'Do you think adultery is immoral?' pose substantial risks of social desirability bias [54] and direct observation of subjects induces risk of Hawthorne and experimenter demand effects [55], among other related methodological concerns.

2.2. Social media data allow measuring culture from the bottom up

Inspired by ethnographic methods [56], we measure culture from the bottom up, enabling a rich, unobtrusive, quantitative description of global cultural factors. We can conceive of the culture of a human group at a point in time as a complex, high-dimensional hypersurface (the black surface in figure 1*a*). This surface is not

narrowed from the top down. Conceptually, it includes art and Angry Birds appreciation, ethics and email enjoyment, formalities and football fans, language and loungewear likes, religion and running routines, politics and potluck preferences, and social structures and sockwear, among every other feature of human life.

Our approach enables measuring culture from the bottom up (the blue surface in figure 1*a*), providing a complement to traditional top–down approaches (the red surface in figure 1*b*). Importantly, it prioritizes no features over others and enables the data to reveal important dimensions among observed human groups. This allows our method to be general, flexible and unsupervised. And it does not selectively exclude constructs, reducing the biases in the measurement of culture encountered by top-down methods.

But how does one go about measuring the high-dimensional hypersurface of culture? Private firms have—perhaps unintentionally—led the way. Effectively and precisely targeting products and services to individuals requires gathering massive amounts of information about those individuals' beliefs, behaviours and preferences [57–59]. For online companies like Google and Facebook, the gathering, storing and selling of this information has itself become a multi-billion dollar enterprise. As a result, the platforms have become adroit at measuring myriad features of human lives via activities that occur during the large amount of time modern humans spend both online and offline in proximity to connected devices [60]. There is a burgeoning literature in the social sciences that relies on social media data [61–65].

Facebook places particular importance in classifying the interests of its users [66]. As a result, the company has inadvertently built the largest platform for the measurement of culture in existence (figure 1*b*). Fortunately for scholars, Facebook makes this information accessible to prospective marketers via a marketing application programming interface (API). Using information drawn from users' self-reported interests, clicking behaviours on Facebook, likes on Facebook, software downloads, GPS location and behaviour on other sites that employ Facebook ads (figure 1*b*), this API provides the ability to create and analyse social groups of interest along hundreds of thousands of interest dimensions and down to very fine spatial and temporal resolution (the zip code-by-day level in the USA). Electronic supplementary material, appendix table B4, illustrates examples of cultural categories along with corresponding Facebook interests both for traditional and non-traditional cultural elements. By making its platform open to those interested in marketing to its users, Facebook has enabled scholars to interrogate its measures of global human interests and construct freely available measures of culture.

We use data gleaned from scraping the Facebook Marketing API to construct a high-dimensional measure of culture. We gathered nearly 60 000 diverse interests by sequentially interrogating Facebook's platform and then constructed—for each administrative unit in our analysis—a vector of the share of individuals in that unit who held each interest. Importantly, each interest on the platform is indexed by a unique identifier, allowing for consistency across languages globally. We use these data to investigate culture at the country, subnational and local levels. Because the data we use are aggregated at the level of population groups, they cannot be used to identify any specific individual and hence do not present privacy concerns. Electronic

supplementary material, appendix A, provides an in-depth discussion of Facebook's algorithm and issues related to representativeness, biases, fake accounts and privacy.

## 3. Facebook-based distances between countries

Employing data on these interest shares drawn from over two billion individual users around the world, we first validate our measure of culture derived from Facebook interests against quantitative measures of country cultural differences taken from prior literature [12,67]. If our Facebook measure captures important components of traditional top-down measures, we expect to observe a positive correspondence between our bottom-up measure and traditional top-down measures (see electronic supplementary material, appendix A, for a description of distance and correlation measures).

Figure 2*a* presents the results of these comparisons. Our bottom-up measure of inter-country cultural distance corresponds positively and significantly to a wide variety of measures of cultural distance between countries. We observe small positive correlations between our measure and measures of linguistic, geographical, religious and genetic distance between country populations [68–77] (see electronic supplementary material, appendix A). However, between more direct measures of traditional notions of culture—provided via the World Values Survey (WVS) [52,78]—we observe a more marked positive correspondence with a correlation coefficient of approximately 0.5 (coefficient: 0.54, *p*-value: 0.0001). Thus our bottom–up measure of cultural distance corresponds positively but imperfectly to traditional measures.

Does this imperfect correspondence result from the measurement of additional components of cultural distance between countries? At face value, figure 2*b* shows that the number of Facebook interests are several orders of magnitude larger than the number of questions in the WVS. To further investigate whether this also translates into capturing more dimensions of culture, we perform principal component analysis on Facebook interests and WVS questions, using the common sample of 69 countries covered by both data sources (see electronic supplementary material, appendix A). Our goal is to reduce the dimensionality of interests and questions and to assess how many unique principal components are able to explain a large share (80%) of the variance in our Facebook measure of culture and in the WVS questions across countries.

Figure 2*c* plots the share of the overall variance in questions and interests that is explained by principal components as a function of their number. Our measure of culture derived from Facebook interests explains 80% of the variance between countries using three times the number of principal components as required to explain 80% of the variance using the WVS. This provides suggestive evidence that the Facebook measure covers a more diverse array of explanatory dimensions of culture as compared to the WVS.

While our Facebook data span a broad variety of interests, do they also capture a broader set of specific cultural traits than those measured by the WVS? To explore this question, we employ a supervised machine learning algorithm that uses all our Facebook interests to predict close to 50 specific cultural attributes, ranging from generosity to gender bias. When comparing the predicted traits to the observed traits, we find an average correlation of 0.6, indicating that the wide array of
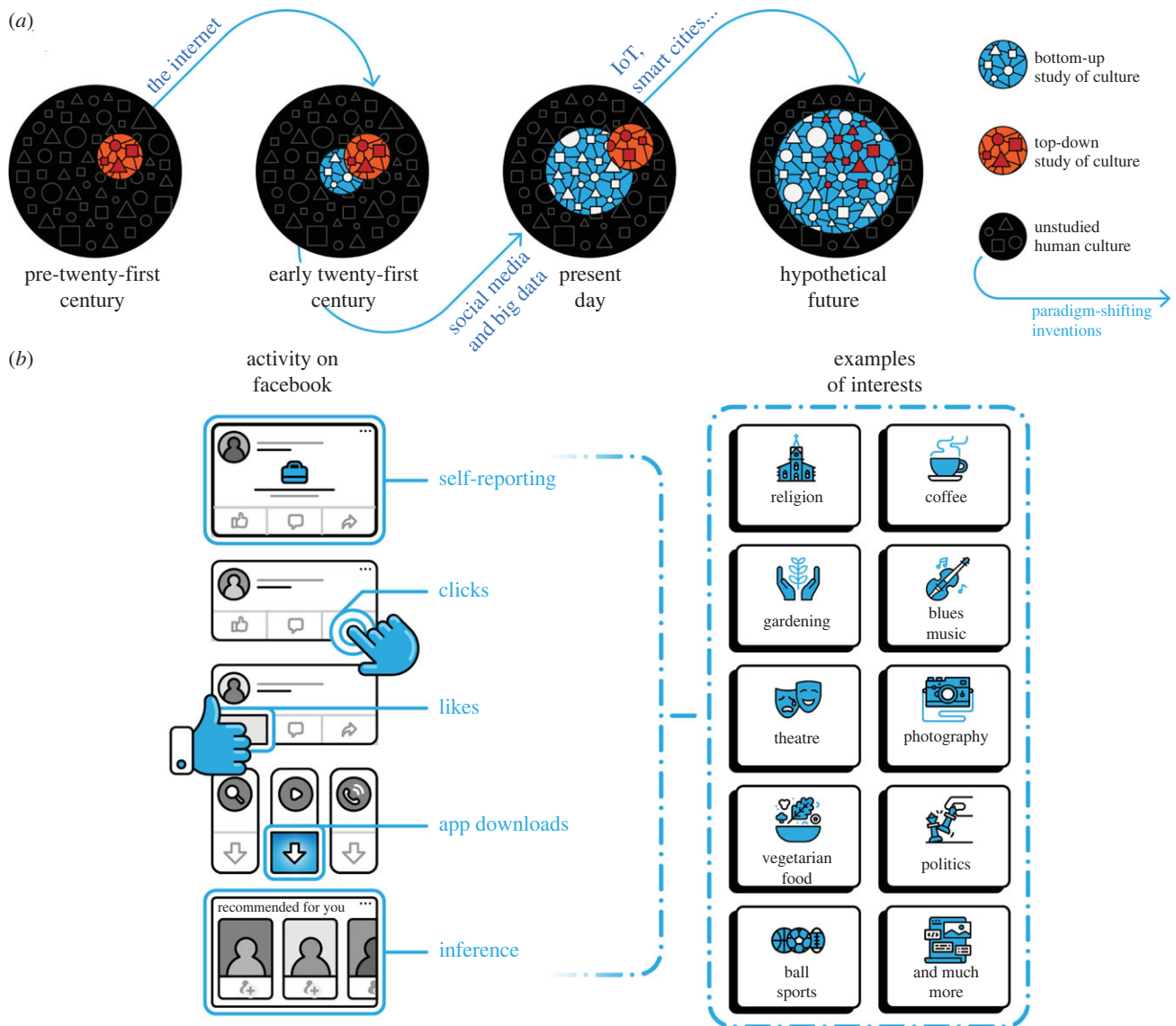
Figure 1. Measurement of the cultural landscape and methods of classification of values, behaviours, preferences and interests on Facebook. Panel (*a*) illustrates how the bottom-up quantitative study of culture is enabled by information technology advances and the broad measurement of humans across the globe. Paradigm shifting technologies such as the Internet, the advent of social media and big data, the introduction of the Internet of Things (IoT) and smart cities all shape the availability of information with which to measure previously unstudied dimensions of culture. We anticipate a hypothetical future in which traditional top-down concepts of culture are encompassed by and integrated into a bottom-up approach to the measurement of culture. Panel (*b*) illustrates how Facebook classifies users' interests via users' self-reporting, via users' observed behaviour which includes the totality of users' clicks on the platform and on ads served elsewhere by the platform, via users' 'likes' and software downloads, and via broad inference based on users' overall behaviour on and off the platform. The interests Facebook infers span hundreds of thousands of dimensions and include topics that both fall within more traditional measures of culture—such as religion, politics and the arts—as well as those that tend to fall outside of traditional measures—such as interests in video games, physical activity preferences and recreational drug interests.

Facebook data are also able to capture specific cultural traits (see electronic supplementary material, appendix table B3).

Next, we examine whether clusters of countries returned via our Facebook measure of culture mirror common conceptions of cultural similarity, providing a measure of face validity. Figure 3 presents a dendrogram of countries, based on the cosine distance between culture vectors constructed from our Facebook data employing the Ward linkage method. The sample of countries consists of those that overlap with the WVS, have a population of more than 300 000, and have a Facebook penetration rate of more than 5% (see electronic supplementary material, appendix A). As can be seen, the unsupervised clustering of countries within our sample provides substantial validity to our measure. Countries that typically are culturally or historically associated with one another—the USA and Canada, India and Bangladesh, Germany and Austria—are placed directly

next to one another in the clusters. Our approach also reveals novel features in the data that go beyond obvious geographical clustering. For example, Puerto Rico is closer to the Latin American cluster than it is to the USA, despite being a US territory. Furthermore, linguistically similar but geographically disparate countries—such as the US and Australia, and Brazil and Portugal—cluster together. It is important to mention that clustering algorithms have some difficulty in dealing with 'outliers'. This explains the maybe surprising location of Japan in figure 3. However, when we extend the dendrogram to include all 225 countries, Japan is no longer special, appearing alongside China (see electronic supplementary material, appendix figure B16). Electronic supplementary material, appendix B, provides extensive robustness checks, exploring different ways of measuring distances and analysing different samples of countries and interests.
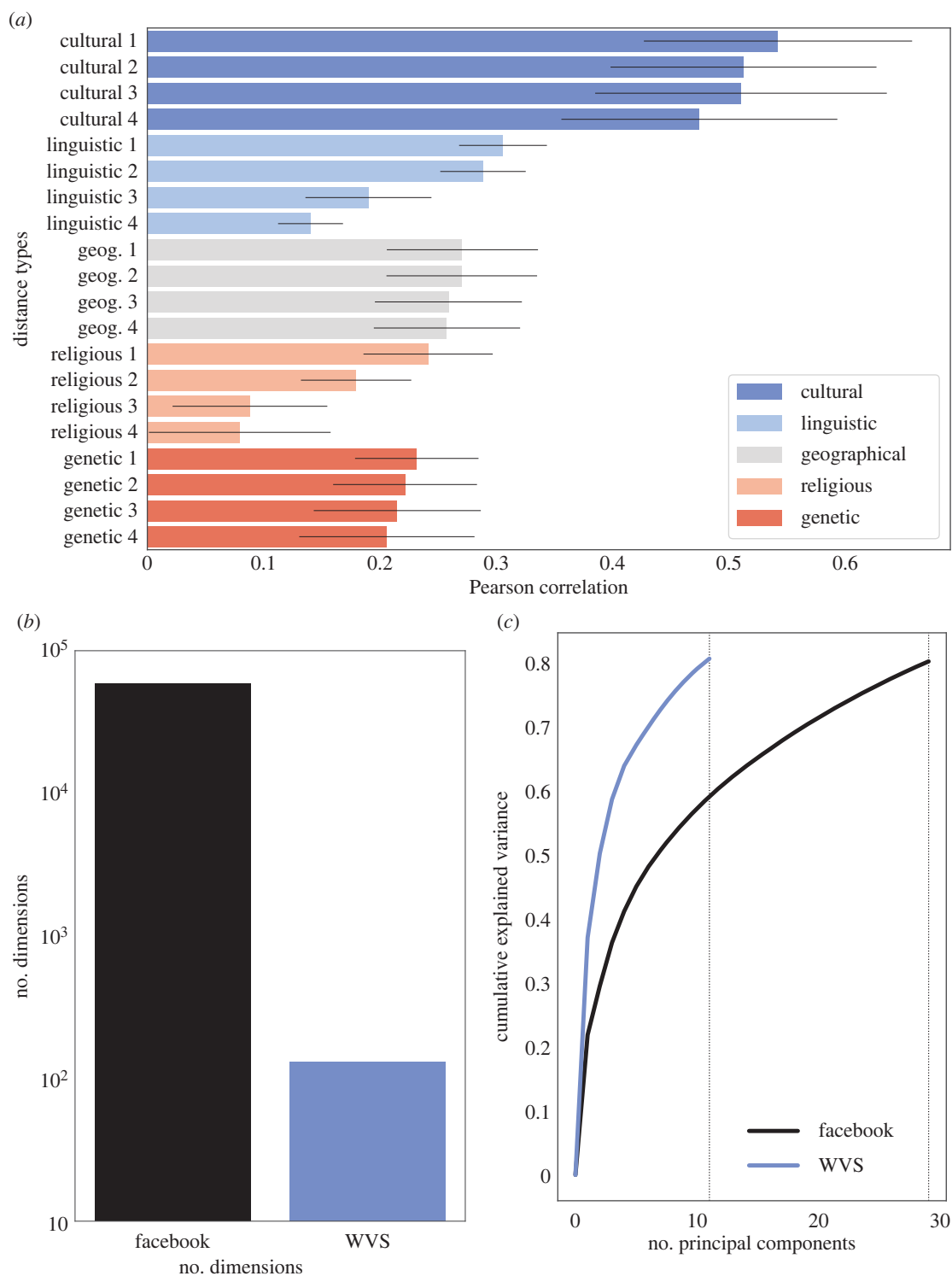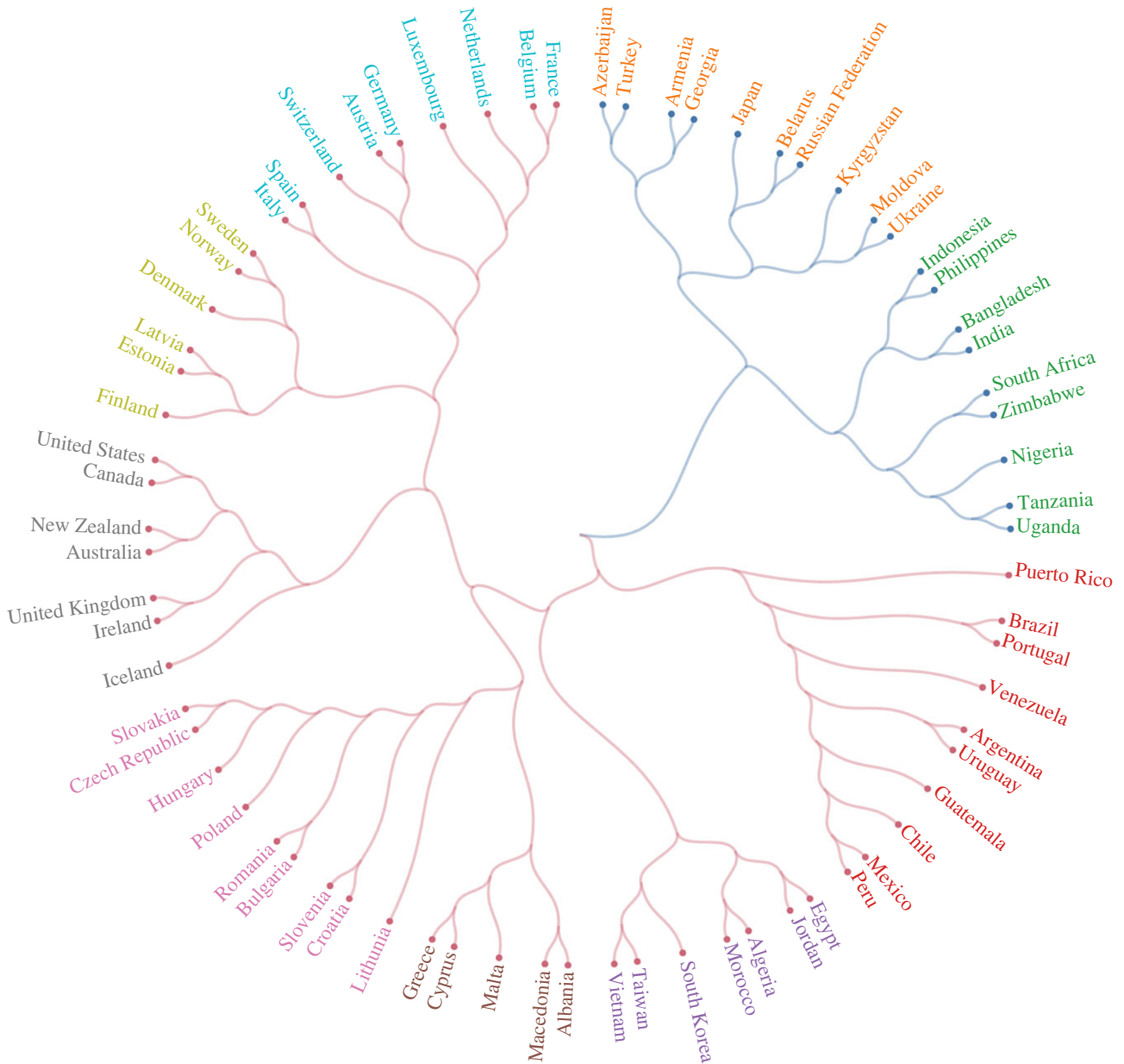
Figure 2. Bottom-up measurement of culture correlates with traditional top-down measures, enables the use of orders of magnitude more features, and explains additional variance. (a) Correlation between our bottom-up Facebook inter-country cultural distance measure and inter-country cultural distances based on traditional proxies (see electronic supplementary material, appendix A, for full list). Bars show the Pearson correlation coefficient between our measure and the four most highly correlated measures within each type of conventional proxy (cultural, linguistic, geographic, religious, and genetic). Black lines depict the 95% confidence intervals for the correlation coefficients based on Mantel tests. Results are based on the common sample of 69 countries for which all measures are available. (b) Bars show the number of Facebook interests and the number of common questions included in all waves of the World Values Survey (WVS). (c) Lines depict the number of principal components and their cumulative explained variance for each measure. To explain 80% of the variation underlying the Facebook interests data across countries 29 principal components are required, as opposed to the 11 principal components necessary to explain the same level of variation in the WVS.

4. Facebook-based distances between subnational regions

Figures 2 and 3 provide evidence for the validity of our method. However, cultural variation is not relegated solely to nation-level groupings. Cultural differences at the subnational level are essential to understand nation-building efforts as well as geopolitical and secessionist threats around the globe. Unfortunately, traditional quantitative measures are highly costly to construct and thus provide
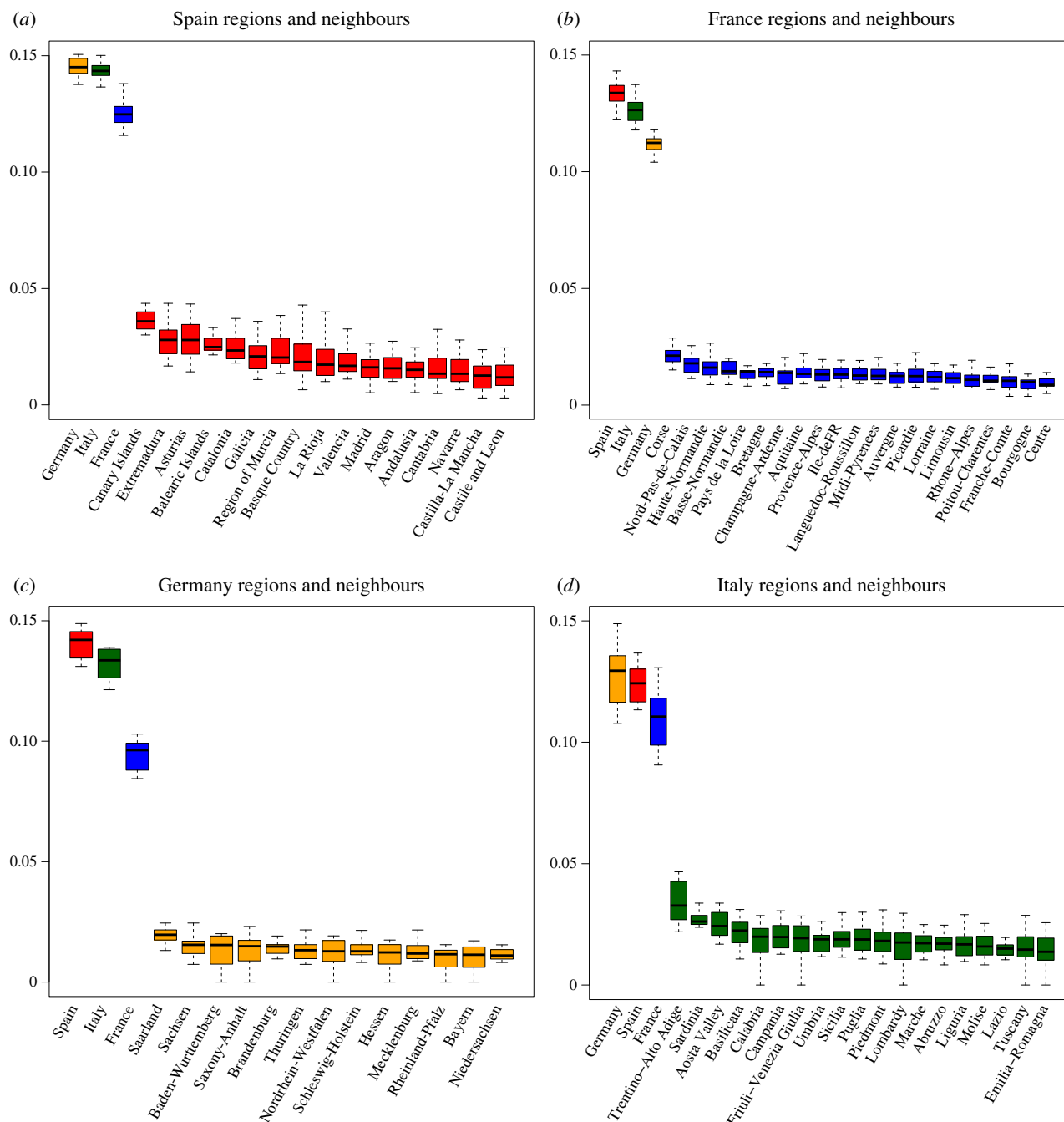
Figure 3. Bottom-up measurement of culture returns sensible clusters of countries. Dendrogram is generated using cosine distance and the Ward linkage method (see electronic supplementary material, appendix A). Countries and territories with at least 300 000 persons, a Facebook penetration rate of more than 5%, and representation in both the Facebook and WVS data were selected. The colour of a country's link represents its membership of a main cluster, while the colour of its name represents its membership of a sub-cluster. Two countries of the same colour are closer to each other than to a country of a different colour.

little representative insight into subnational cultures. Conducting representative surveys at high resolution globally would be cost-prohibitive (in the limit, the costs would approach those of the Facebook platform itself). Might our Facebook measure be able to provide improved, scalable insight into novel subnational cultural variations?

To investigate this question, we gather vectors of Facebook interests for subnational regions in the USA and Europe and compute cosine distances between each region within the country and all the other regions within the same country (see electronic supplementary material, appendix A). One question is whether subnational regions are less distant from other regions within their own country than from closeby countries. Figure 4 investigates this for the regions of Spain (red), France (blue), Germany (yellow) and Italy (green). For example, panel (a) depicts the distribution of the cultural distances between Spanish regions, and between these regions and other European countries. The subnational regions within Spain are much less distant from one other than from

neighbouring European countries. For example, Catalonia is culturally markedly closer to any other Spanish region than to either Italy or France. This same pattern holds for regions in the other countries (figure 4b–d).

Before concluding that national borders demarcate sharp cultural boundaries, we also compare cultural distances between sub-national regions in different countries. While sub-national regions are culturally closer to each other than to other countries (Paris is closer to other regions of France than to Spain), we might expect capital cities to resemble each other (Paris might be closer to Madrid than to rural regions of France). We observe the opposite. Almost all sub-regions in our data are closest to sub-regions within their own national borders. Only two sub-regions in our European data—Flanders in Belgium and Donegal County in Ireland—are closer to regions in a separate country than they are to other regions in their own nation. Both exceptions can be traced back to fairly recent changes in country borders: the splitting of the province of Limburg between Belgium
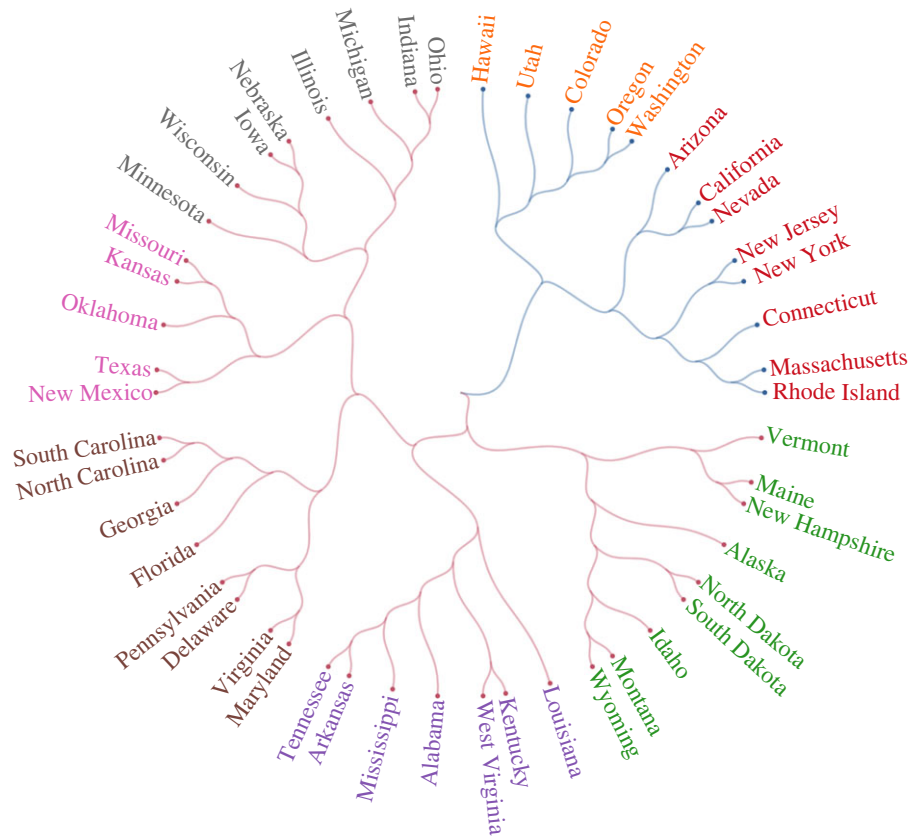
Figure 4. Bottom-up measurement of subnational culture indicates substantial within-country cultural similarity. (*a*) Subnational units in Spain are less distant from one another than from European cultural neighbours of Germany, Italy or France. (*b*) French regions are culturally closer to one another than to European neighbours. (*c*) German regions demonstrate greater similarity to one another than to European neighbours. (*d*) Italian regions are culturally less distant from one another than from their European neighbours. Subnational units in (*a–d*) are represented with boxplots that show the distance distribution between a given region and all other regions within the country under analysis. In the case of neighbouring countries, boxplots show the distance distribution between that neighbouring country and all the regions within the country under analysis. Colour coding: distances to Spain or Spanish regions (red), to France or French regions (blue), to Germany or German regions (yellow), to Italy or Italian regions (green).

and the Netherlands in the 1830s and the Partition of Ireland in the 1920s. Overall, this suggests the importance of national boundaries in shaping cultural distances. It also suggests that our measure captures deep cultural elements that persist over longer periods of time.

Do subnational cultures as measured via our Facebook data cluster together in a sensible manner? To examine this question, we calculate cosine distances for US states and perform unsupervised clustering using the Ward linkage method (see electronic supplementary material, appendix A). Figure 5 presents the resulting dendrogram. The clusters return traditional regional and cultural groupings. For

example, states in the US Midwest are placed in proximity to one another, as are the states in the US South. Mountainous and more rural states also cluster together, with Alaska being closest to states like North Dakota, Idaho and New Hampshire, despite the substantial geographical distances between them.

Figure 4 demonstrates that regions within countries can bear substantial similarity to one another. However, not all countries are likely to have the same amount of within-nation cultural similarity. Do some countries have more regional cultural variation within them than do others? Figure 6 examines countries according to their interregional

Figure 5. Bottom-up measurement of subnational culture returns sensible clustering of subnational units. Facebook culture vectors drawn from the states of the USA cluster together in an unsupervised manner into traditional regional and cultural groupings. The colour of a state's link represents its membership of a main cluster, while the colour of its name represents its membership of a sub-cluster. Two states of the same colour are closer to each other than to a country of a different colour. South Carolina bears similarity to North Carolina, Montana to Wyoming, New Jersey to New York and West Virginia to Kentucky. Similarities are not strictly geographical, with Alaska bearing similarity to New Hampshire and North Dakota, for example.

cultural divisiveness, or the average cultural distance between regions within a country (see electronic supplementary material, appendix A). Figure 6a ranks 18 selected countries in increasing order of interregional divisiveness. Two findings stand out in figure 6a. First, developed countries exhibit smaller interregional divisiveness when compared with developing countries, suggesting that they benefit from greater cohesiveness between regions. Second, within the group of developed countries, the three countries with the least interregional cohesiveness are Belgium, Spain and the USA. The first two have well-known regional issues, with threats of secession, whereas the third is a large geographical nation.

Figure 6b,c displays the geographical maps and networks of regions in Germany (13 regions) and regions in India (34 regions), respectively (see electronic supplementary material, appendix A). We detect two communities of regions in Germany that map closely to the historical east–west divide in the country, suggesting this cultural divide still persists to this day. We detect three regional communities in India that correspond roughly geographically with linguistic regions defined by the language families spoken within the country.

5. Other applications of Facebook-based distances

Figures 4–6 highlight the utility of our measure in assessing subnational questions that are simply too expensive to

measure with traditional quantitative approaches. Yet differences in a society are not limited only to subnational differences. Societies can also differ along other identity cleavages, such as age, gender or race. Our measure enables us to also delve into the nature of cultural differences that vary according to such demographic groups and into differences that occur at even finer degrees of spatial resolution. Figure 7a,b explore whether countries that exhibit more divisiveness in one dimension also do so in other dimensions. It shows that age divisions and gender divisions have a weak positive association (Pearson correlation coefficient: 0.234, p: 0.146), whereas countries that suffer from greater regional divisions have smaller differences between men and women (Pearson correlation coefficient: −0.702, p: 0.001). Many developing countries exhibit more cohesiveness between genders, although they experience larger regional divides, compared to many developed countries.

Furthermore, our data enable us to investigate cultural similarities and differences at even higher spatial resolution. Figure 8 depicts the dendrogram of the cultural clustering of the most populous California counties. Geographically disparate but culturally similar counties—such as coastal surfing communities of San Luis Obispo and Santa Cruz as well as the rural inland counties of Imperial and Butte—are located next to one another in the dendrogram. Generalizing this approach would allow us to identify which local areas culturally diverge from the rest of the nation in which they are located, a phenomenon that might provide insight into regional political disaffection.
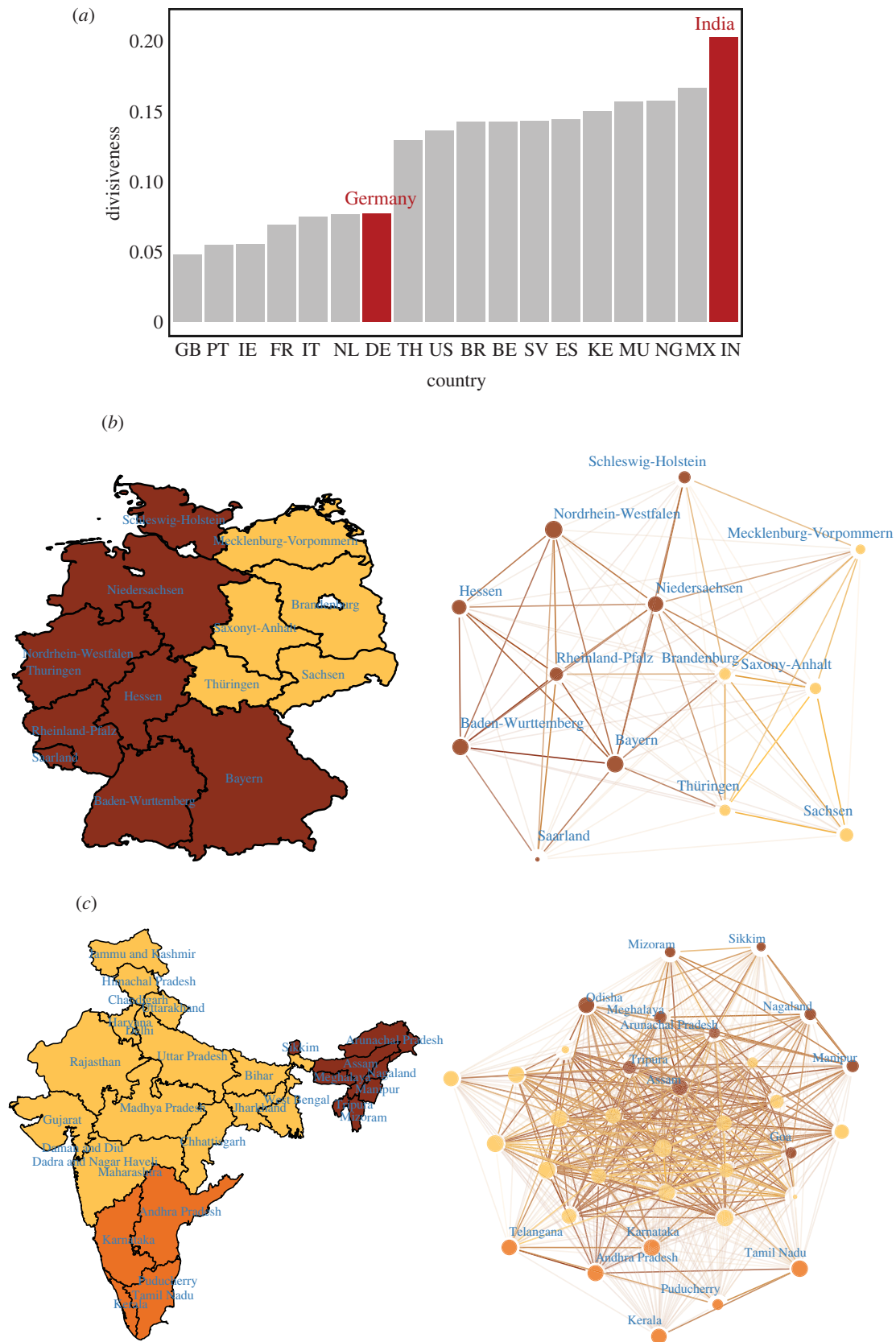
Figure 6. Regional divisiveness. (*a*) Population-weighted regional divisiveness for 18 countries (range for number of regions: [3,76]). Countries included are: BE, Belgium; BR, Brazil; DE, Germany; ES, Spain; FR: France; GB, Great Britain; IE, Ireland; IN, India; IT, Italy; KE, Kenya; MU, Mauritius; MX, Mexico; NG, Nigeria; NL, Netherlands; PT, Portugal; SV, El Salvador; TH, Thailand; US, United States. We chose these 18 countries to have a mix of developed countries and developing countries, as well as a mix of large and small countries. (*b*) Geographical map and network of regions in Germany (13 regions). Two communities of regions are detected. (*c*) Geographical map and network of regions in India (34 regions). Three communities of regions are detected. Networks are constructed from nodes as regions, and links are weighted by standardized cosine similarity between regions. Communities are detected using a multi-level modularity optimization algorithm (Louvain method) [79]. Nodes are resized proportionally to region population, and are coloured according to community affiliation. Links are coloured according to adjacent nodes, with lower transparency applied to higher weight links. Map regions are coloured according to communities calculated from the corresponding network. (*a*) Population-weighted regional divisiveness, (*b*) Germany, (*c*) India.
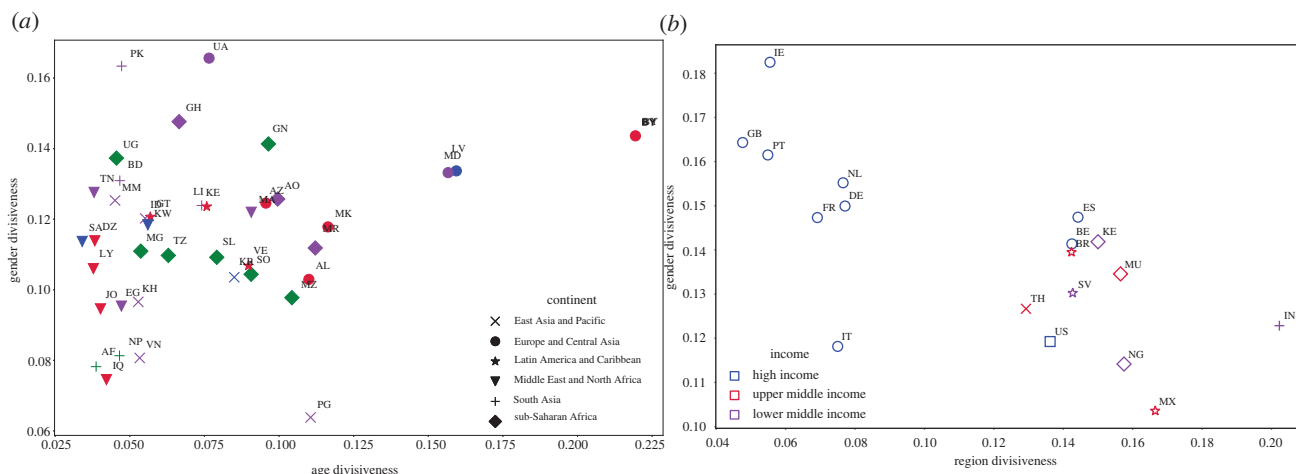
Figure 7. Subnational age, gender and regional cultural differences. (*a*) A scatter plot of age and gender divisiveness for 40 countries. The gender divisiveness is computed as the cosine similarity between the FB penetration vector for women and men using our sample of Facebook interests. The age divisiveness is computed as the median cosine similarity between the Facebook culture vector of three different age groups 15–29, 30–64 and 65+. (AF, Afghanistan; AL, Albania; AO, Angola; AZ, Azerbaijan; BD, Bangladesh; BY, Belarus; DZ, Algeria; EG, Egypt; GH, Ghana; GN, Guinea; GT, Guatemala; ID, Indonesia; IQ- Iraq; JO, Jordan; KH, Cambodia; KR, South Korea; KW, Kuwait; LK, Sri Lanka; LV, Latvia; LY, Libya; MA, Morocco; MD, Moldova; MG, Madagascar; MK, Macedonia; MM, Myanmar; MR, Mauritania; MZ, Mozambique; NP, Nepal; PE, Peru; PG, Papua New Guinea; PK, Pakistan; SA, Saudi Arabia; SL, Sierra Leone; SO, Somalia; TN, Tunisia; TZ, Tanzania; UA, Ukraine; UG, Uganda; VE, Venezuela; VN, Vietnam). Panel B shows a scatter plot of regional and gender divisiveness for 18 countries using (BE, Belgium; BR, Brazil; DE, Germany; ES, Spain; FR, France; GB, Great Britain; IE, Ireland; IN, India; IT, Italy; KE, Kenya; MU, Mauritius; MX, Mexico; NG, Nigeria; NL, Netherlands; PT, Portugal; SV, El Salvador; TH, Thailand; US, United States). (*a*) Divisiveness by age versus gender, (*b*) divisiveness by regions versus gender.

6. Further discussion and concluding remarks

Our method lays out a complementary approach to the study of culture that can increasingly be measured via the application of computational social science to the ever larger portion of human lives that are unobtrusively and observationally measured online and offline. Doing so with our sample of Facebook interests for countries and subnational and local regions around the world indicates that our bottom-up measure of culture corresponds positively to traditional quantitative measures, contains a greater number of explanatory dimensions, enables the clustering of countries, subnational units, and localities into sensible groupings, and provides insight into cultural variation at unprecedented spatial, demographic and topic-based resolution. A further strength of our approach is its ability to answer questions about human culture that have been—up to this point— impossible to investigate at scale using traditional quantitative methods. For example, our method and data can enable us to investigate questions such as, 'Which country is the cultural centre of the world?' and 'Which is the global 'sister region' of a particular region within a country?' (see electronic supplementary material, appendix B4 and B6).

The high spatial (zip code-level) and temporal (daily) resolution of Facebook's available data, coupled with the more than 200 million individuals on the platform in the United States and the more than two billion on the platform around the world, enable the measuring of cultural differences with remarkable precision. Computing cultural differences between subnational regions (figures 4–6), cities, counties (figure 8), or any different subgroups of any country (figure 7), which is cost-prohibitive when using traditional surveys, now becomes a straightforward endeavour. These studies can be conducted freely via the publicly available data provided by the Facebook Marketing API. Even so, this API is limited relative to what is theoretically possible, given the magnitude of human behavioural data that firms are currently collecting globally.

While we believe our conception and measurement of culture provide numerous complementary benefits to traditional measures, a number of considerations are worth noting. For instance, the fact that our approach does not inductively distill culture into parsimonious concepts means that the constellations of interests and behaviours that might diverge between two cultures may not always lend themselves to ease of conceptual interpretation. This is one drawback of our methodology.

Furthermore, while our Facebook measure of culture represents a marked improvement in terms of its ability to measure the surface of culture as compared to traditional surveys with high rates of non-response and relatively few questions [54], it is still far from perfect. Not all individuals in every country around the world are on Facebook. Our analysis only generalizes to differences among those who use Facebook (however, our validity results persist even when looking at countries with lower penetration of Facebook use; see electronic supplementary material, appendix B2). Additionally, while nearly 60 000 dimensions represent a dramatic increase over traditional top-down methods (of three orders of magnitude; see figure 2*a*), they still measure only a fraction of the full dimensionality of culture. Moreover, these dimensions themselves suffer from a certain endogeneity of measurement: Facebook does not have an interest listed for every possible feature of culture. Those it chooses to classify are endogenous to the platform itself. For example, these interests exclude certain topics—like sex and hate speech—which are banned from the platform. A more ideal computational system would classify all interests at an even finer scale of demarcation.

Finally, while we believe that our measure possesses numerous complementary benefits to traditional quantitative and qualitative approaches to culture—unobtrusiveness, scale, resolution, richness and breadth of constructs, and the ability to freely peer into the lives of billions of people—our measure emphatically cannot substitute for traditionally employed approaches to culture [80]. If we

Figure 8. Local cultural similarities of California counties. The figure represents a dendrogram for California counties, based on cosine distances and Ward linkage method. In this figure, we examine interests with at least 20 000 users in the state, and in counties with a population of more than 75 000 users. Our unsupervised clustering returns clusters that are culturally sensible. The colour of a county's link represents its membership of a main cluster, while the colour of its name represents its membership of a sub-cluster. Los Angeles County is closest to Orange County, Tulare County is most similar to Stanislaus County, Santa Cruz County is proximal to San Luis Obispo Countyand San Francisco County is nearest Marin County.

observe that individuals in a place spend substantial time looking at religious websites, we cannot know to what degree they personally hold religion as important: traditional approaches are needed to investigate further. And while our method is useful in providing a culturomic [81,82] barcoding [83,84] of global culture, the classification and interpretation of culturally important factors still necessitate in-depth and careful study of features uncovered by our approach—as well as of those features of culture that our measure omits.

Humans around the world share many cultural similarities but also have many differences. Until very recently, quantitatively measuring vast portions of culture was impossible. However, as we collect data on humanity it becomes increasingly possible to measure the surface of human culture in manners that approach cultures' full underlying dimensionality. Doing so—in addition to furthering traditional quantitative and qualitative approaches to culture—will aid in a richer understanding of global human culture.

This improved ability to measure cultural differences between population groups at a more granular level enables a dramatic advance in the evaluation of some of the most pressing questions in the social sciences, such as: do national borders shape cultures? Are societies more likely to fracture along gender lines, racial lines, or regional lines? Which specific locations on the globe are more prone to civil conflict and violence? Does a lack of cultural cohesiveness contribute to political extremism? Are certain immigrant cultures more adept at integrating than others? Some of these questions have been touched upon in this paper; others have not. All have one element in common: answering them requires cultural measurement at previously unavailable scales and resolutions.

Authors' contributions. N.O.: conceptualization, formal analysis, investigation, methodology, supervision, visualization, writing—original draft, writing—review and editing; Ö.Ö.: conceptualization, data curation, formal analysis, investigation, methodology, resources, software, visualization, writing—original draft, writing—review and editing; I.M.: data curation, investigation, methodology, resources, software, visualization, writing—original draft; I.O.O.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, resources, validation, visualization, writing—original draft, writing—review and editing; E.A.: data curation, investigation, visualization, writing—review and editing; M.C.: conceptualization, investigation; R.C.: conceptualization, data curation,

References

1. Boyd R, Richerson PJ. 2005 The origin and evolution of cultures. Oxford, UK: Oxford University Press.

2. Kempe M, Lycett SJ, Mesoudi A. 2014 From cultural traditions to cumulative culture: parameterizing the differences between human and nonhuman culture. J. Theor. Biol. 359, 29–36. (doi:10.1016/j.jtbi.2014.05.046)

3. Mesoudi A, Thornton A. 2018 What is cumulative cultural evolution? Proc. R. Soc. B 285, 20180712. (doi:10.1098/rspb.2018.0712)

4. Dean LG, Kendal RL, Schapiro SJ, Thierry B, Laland KN. 2012 Identification of the social and cognitive processes underlying human cumulative culture. Science 335, 1114–1118. (doi:10.1126/science.1213969)

5. Baldwin JR, Faulkner SL, Hecht ML, Lindsley SL. 2006 Redefining culture: perspectives across the disciplines. New York, NY: Taylor & Francis.

6. Harris M . 2001 Cultural materialism: the struggle for a science of culture. Walnut Creek, CA: AltaMira Press.

7. Goldstein LJ. 1957 On defining culture. Am. Anthropol. 59, 1075–1081. (doi:10.1525/aa.1957.59.6.02a00110)

8. Swidler A. 1986 Culture in action: symbols and strategies. Am. Sociol. Rev. 51, 273–286. (doi:10.2307/2095521)

9. Chilton S. 1988 Defining political culture. West. Polit. Q. 41, 419–445. (doi:10.1177/106591298804100303)

10. Ross MH. 2009 Culture in comparative political analysis. In Comparative politics: rationality, culture, and structure, pp. 134–161. Cambridge, UK: Cambridge University Press. (doi:10.1017/CBO9780511804007.007).

11. Huntington SP. 1997 The clash of civilizations and the remaking of world order. New York, NY: Simon & Schuster.

12. Spolaore E, Wacziarg R. 2016 Ancestry, language and culture. In The Palgrave handbook of economics and language (eds V Ginsburgh , S Weber), pp. 174–211. New York, NY: Springer.

13. Oosterbeek H, Sloof R, van de Kuilen G. 2001 Cultural differences in ultimatum game experiments: evidence from a meta-analysis. (doi:10.2139/ssrn.286428)

14. Ottaviano GIP, Peri G. 2004 The economic value of cultural diversity: evidence from US cities. NBER WP 10904. (doi:10.3386/w10904)

15. Tomasello M. 2009 The cultural origins of human cognition. Cambridge, MA: Harvard University Press.

16. Cohen AB. 2009 Many forms of culture. Am. Psychol. 64, 194–204. (doi:10.1037/a0015308)

17. Betancourt H, López SR. 1993 The study of culture, ethnicity, and race in American psychology. Am. Psychol. 48, 629–637. (doi:10.1037/0003-066X.48.6.629)

18. Mesoudi A. 2016 Cultural evolution: integrating psychology, evolution and culture. Curr. Opin. Psychol. 7, 17–22. (doi:10.1016/j.copsyc.2015.07.001)

19. Ekman P, Friesen WV, O'Sullivan M, Chan A. 1987 Universals and cultural differences in the judgments of facial expressions of emotion. J. Pers. Soc. Psychol. 53, 712–717. (doi:10.1037/0022-3514.53.4.712)

20. Moore OK. 1952 Nominal definitions of 'Culture'. Phil. Sci. 19, 245–256. (doi:10.1086/287209)

21. Alesina A, Spolaore E. 1997 On the number and size of nations. Q. J. Econ. 112, 1027–1056. (doi:10.1162/003355300555411)

22. Alesina A, Spolaore E, Wacziarg R. 2000 Economic integration and political disintegration. Am. Econ. Rev. 90, 1276–1296. (doi:10.1257/aer.90.5.1276)

23. Desmet K, Le Breton M, Ortuño-Ortín I, Weber S. 2011 The stability and breakup of nations: a quantitative analysis. J. Econ. Growth 16, 183–213. (doi:10.1007/s10887-011-9068-z)

24. Fearon JD, Laitin DD. 2003 Ethnicity, insurgency, and civil war. Am. Polit. Sci. Rev. 97, 75–90. (doi:10.1017/S0003055403000534)

25. Fearon JD. 2003 Ethnic and cultural diversity by country. J. Econ. Growth 8, 195–222. (doi:10.1023/A:1024419522867)

26. Desmet K, Ortuño-Ortín I, Wacziarg R. 2012 The political economy of linguistic cleavages. J. Dev. Ec. 97, 322–338. (doi:10.1016/j.jdeveco.2011.02.003)

27. Alesina A, Ferrara EL. 2005 Ethnic diversity and economic performance. J. Econ. Lit. 43, 762–800. (doi:10.1257/002205105774431243)

28. Thomson R et al. 2018 Relational mobility predicts social behaviors in 39 countries and is tied to historical farming and threat. Proc. Natl. Acad. Sci. USA 115, 7521–7526. (doi:10.1073/pnas.1713191115)

29. O'Hear A. 1998 Routledge encyclopedia of philosophy. London, UK: Taylor and Francis.

30. Tylor EB. 1871 Primitive culture: researches into the development of mythology, philosophy, religion, art and custom. London, UK: J. Murray.

31. Taras V, Rowney J, Steel P. 2009 Half a century of measuring culture: review of approaches, challenges, and limitations based on the analysis of 121 instruments for quantifying culture. J. Int. Manage. 15, 357–373. (doi:10.1016/j.intman.2008.08.005)

32. Lazer D et al. 2009 Life in the network: the coming age of computational social science. Science 323, 721. (doi:10.1126/science.1167742)

33. van der Post DJ, Hogeweg P. 2008 Diet traditions and cumulative cultural processes as side-effects of grouping. Anim. Behav. 75, 133–144. (doi:10.1016/j.anbehav.2007.04.021)

34. Tamariz M, Kirby S. 2016 The cultural evolution of language. Curr. Opin. Psychol. 8, 37–43. (doi:10.1016/j.copsyc.2015.09.003)

35. Kirby S, Dowman M, Griffiths TL. 2007 Innateness and culture in the evolution of language. Proc. Natl Acad. Sci. USA 104, 5241–5245. (doi:10.1073/pnas.0608222104)

36. Arendt H. 2007 Reflections on literature and culture. Stanford, CA: Stanford University Press.

37. Acerbi A, Enquist M, Ghirlanda S. 2009 Cultural evolution and individual development of openness and conservatism. Proc. Natl Acad. Sci. USA 106, 18 931–18 935. (doi:10.1073/pnas.0908889106)

38. Atran S, Henrich J. 2010 The evolution of religion: how cognitive by-products, adaptive learning heuristics, ritual displays, and group competition generate deep commitments to prosocial religions. Biol. Theory 5, 18–30. (doi:10.1162/BIOT_a_00018)

39. Geertz C. 1973 The interpretation of cultures. New York, NY: Basic Books.

40. Murthy D. 2008 Digital ethnography: an examination of the use of new technologies for social research. Sociology 42, 837–855. (doi:10.1177/0038038508094565)

41. Henrich J. 2004 Demography and cultural evolution: how adaptive cultural processes can produce

maladaptive losses—the Tasmanian case. *Am. Antiq.* **69**, 197–214. (doi:10.2307/4128416)

42. Posner DN. 2004 The political salience of cultural difference: why Chewas and Tumbukas are allies in Zambia and adversaries in Malawi. *Am. Polit. Sci. Rev.* **98**, 529–545. (doi:10.1017/S0003055404041334)

43. Murray JH. 2006 Toward a cultural theory of gaming: digital games and the co-evolution of media, mind, and culture. *Popular Commun.* **4**, 185–202. (doi:10.1207/s15405710pc0403_3)

44. Kottak CP. 2016 *Prime-time society: an anthropological analysis of television and culture*. New York, NY: Routledge.

45. Bontempo RN, Bottom WP, Weber EU. 1997 Cross-cultural differences in risk perception: a model-based approach. *Risk Anal.* **17**, 479–488. (doi:10.1111/j.1539-6924.1997.tb00888.x)

46. Yuki M, Maddux WW, Brewer MB, Takemura K. 2005 Cross-cultural differences in relationship- and group-based trust. *Pers. Soc. Psychol. Bull.* **31**, 48–62. (doi:10.1177/0146167204271305)

47. Schimmack U, Oishi S, Diener E. 2005 Individualism: a valid and important dimension of cultural differences between nations. *Pers. Soc. Psychol. Rev.* **9**, 17–31. (doi:10.1207/s15327957pspr0901_2)

48. Archer J. 2006 Cross-cultural differences in physical aggression between partners: a social-role analysis. *Pers. Soc. Psychol. Rev.* **10**, 133–153. (doi:10.1207/s15327957pspr1002_3)

49. Segall MH, Campbell DT, Herskovits MJ. 1963 Cultural differences in the perception of geometric illusions. *Science* **139**, 769–771. (doi:10.1126/science.139.3556.769)

50. Shenkar O. 2001 Cultural distance revisited: towards a more rigorous conceptualization and measurement of cultural differences. *J. Int. Bus. Stud.* **32**, 519–535. (doi:10.1057/palgrave.jibs.8490982)

51. Hofstede G. 1983 National cultures in four dimensions: a research-based theory of cultural differences among nations. *Int. Stud. Manage. Organ.* **13**, 46–74. (doi:10.1080/00208825.1983.11656358)

52. Inglehart R *et al.* 2014 *World Values Survey: All Rounds - Country-Pooled Datafile Version*: http://www.worldvaluessurvey.org/WVSDocumentationWVL.jsp. Madrid, Spain and Vienna, Austria: JD Systems Institute.

53. Geertz C. 2008 Thick description: toward an interpretive theory of culture. In *The Cultural Geography Reader* (eds Oakes T, Price PL), pp. 41–51. London, UK: Routledge.

54. Fisher RJ. 1993 Social desirability bias and the validity of indirect questioning. *J. Consum. Res.* **20**, 303–315. (doi:10.1086/209351)

55. McCambridge J, Witton J, Elbourne DR. 2014 Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *J. Clin. Epidemiol.* **67**, 267–277. (doi:10.1016/j.jclinepi.2013.08.015)

56. Brewer J. 2000 *Ethnography*. Milton Keynes, UK: McGraw-Hill Education.

57. Zhang Y, Neelankavil JP. 1997 The influence of culture on advertising effectiveness in China and the USA. *Eur. J. Mark.* **31**, 134–149. (doi:10.1108/03090569710157106)

58. Hong JW, Muderrisoglu A, Zinkhan GM. 1987 Cultural differences and advertising expression: a comparative content analysis of Japanese and U.S. magazine advertising. *J. Advert.* **16**, 55–68. (doi:10.1080/00913367.1987.10673061)

59. Chau PYK, Cole M, Massey AP, Montoya-Weiss M, O'Keefe RM. 2002 Cultural differences in the online behavior of consumers. *Commun. ACM* **45**, 138–143. (doi:10.1145/570907.570911)

60. Kosinski M, Stillwell D, Graepel T. 2013 Private traits and attributes are predictable from digital records of human behavior. *Proc. Natl Acad. Sci. USA* **110**, 5802–5805. (doi:10.1073/pnas.1218772110)

61. Brady WJ, Wills JA, Jost JT, Tucker JA, Van Bavel JJ. 2017 Emotion shapes the diffusion of moralized content in social networks. *Proc. Natl Acad. Sci. USA* **114**, 7313–7318. (doi:10.1073/pnas.1618923114)

62. Carrignon S, Bentley RA, Ruck D. 2019 Modelling rapid online cultural transmission: evaluating neutral models on Twitter data with approximate Bayesian computation. *Palgrave Commun.* **5**, 83. (doi:10.1057/s41599-019-0295-9)

63. Matz SC, Kosinski M, Nave G, Stillwell DJ. 2017 Psychological targeting as an effective approach to digital mass persuasion. *Proc. Natl Acad. Sci. USA* **114**, 12 714–12 719. (doi:10.1073/pnas.1710966114)

64. Cabañas JG, Cuevas R. 2021 Does Facebook use sensitive data for advertising purposes? *Commun. ACM* **64**, 62–69. (doi:10.1145/3426361)

65. Rampazzo F, Bijak J, Vitali A, Weber I, Zagheni E. 2021 A framework for estimating migrant stocks using digital traces and survey data: an application in the United Kingdom. *Demography* **58**, 2193–2218. (doi:10.1215/00703370-9578562)

66. Kosinski M, Matz SC, Gosling SD, Popov V, Stillwell D. 2015 Facebook as a research tool for the social sciences: opportunities, challenges, ethical considerations, and practical guidelines. *Am. Psychol.* **70**, 543–556. (doi:10.1037/a0039210)

67. Muthukrishna M, Bell A, Henrich J, Curtin C, Gedranovich A, McInerney J, Thue B. 2020 Beyond WEIRD psychology: measuring and mapping scales of cultural and psychological distance. *Psychol. Sci* **31**, 678–701. (doi:10.2139/ssrn.3259613)

68. Cavalli-Sforza LL, Feldman MW. 1981 *Cultural transmission and evolution: a quantitative approach*. Princeton, NJ: Princeton University Press.

69. Cavalli-Sforza LL, Menozzi P, Piazza A. 1994 *The history and geography of human genes*. Princeton, NJ: Princeton University Press.

70. Cavalli-Sforza LL. 2000 *Genes, peoples, and languages*. New York, NY: North Point Press.

71. Atkinson QD. 2011 Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science* **332**, 346–349. (doi:10.1126/science.1199295)

72. Rogers DS, Feldman MW, Ehrlich PR. 2009 Inferring population histories using cultural data. *Proc. R. Soc. B* **276**, 3835–3843. (doi:10.1098/rspb.2009.1088)

73. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. 2005 Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc. Natl Acad. Sci. USA* **102**, 15 942–15 947. (doi:10.1073/pnas.0507611102)

74. Diamond JM. 1997 *Guns, germs, and steel: the fates of human societies*. New York, NY: W.W. Norton & Co.

75. Henrich JP. 2016 *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*. Princeton, NJ: Princeton University Press.

76. Richerson PJ, Christiansen MH. 2013 *Cultural evolution: society, technology, language, and religion*. Cambridge, MA: MIT Press.

77. Galor O, Özak Ö. 2016 The agricultural origins of time preference. *Am. Econ. Rev.* **106**, 3064–3103. (doi:10.1257/aer.20150020)

78. Inglehart R *et al.* 2014 *World Values Survey: Round Six - Country-Pooled Datafile Version*: http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp. JD Systems Institute.

79. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. 2008 Fast unfolding of communities in large networks. *J. Stat. Mech: Theory Exp.* **2008**, P10008. (doi:10.1088/1742-5468/2008/10/P10008)

80. Bail CA. 2014 The cultural environment: measuring culture with big data. *Theory Soc.* **43**, 465–482. (doi:10.1007/s11186-014-9216-5)

81. Michel J-B *et al.* 2011 Quantitative analysis of culture using millions of digitized books. *Science* **331**, 176–182. (doi:10.1126/science.1199644)

82. Reeves B, Robinson T, Ram N. 2020 Time for the human screenome project. *Nature* **577**, 314–317. (doi:10.1038/d41586-020-00032-5)

83. Mesoudi A. 2017 Pursuing Darwin's curious parallel: prospects for a science of cultural evolution. *Proc. Natl Acad. Sci. USA* **114**, 7853–7860. (doi:10.1073/pnas.1620741114)

84. Hebert PDN, Cywinska A, Ball SL. 2003 Biological identifications through DNA barcodes. *Proc. R. Soc. B* **270**, 313–321. (doi:10.1098/rspb.2002.2218)

85. Obradovich N *et al.* 2022 Expanding the measurement of culture with a sample of two billion humans. Figshare. (doi:10.6084/m9.figshare.c.5978508)