Appendix to "Expanding the Measurement of Culture with a Sample of Two Billion Humans"

Nick Obradovich, Ömer Özak, Ignacio Martín, Ignacio Ortuño-Ortín, Edmond Awad, Manuel Cebrián, Rubén Cuevas, Klaus Desmet, Iyad Rahwan and Ángel Cuevas*

April 13, 2022

Abstract

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 behaviorally 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 analyze the importance of national borders in shaping culture and compare subnational divisiveness to 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, the disaffection of certain population groups from the political process, among myriad other potential future applications.

Keywords: Culture, Cultural Distance, Identity, Regional Culture, Gender Differences.

JEL Classification: C80, F1, J1, O10, R10, Z10

*Obradovich: Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, obradovich@mpib-berlin.mpg.de; Özak: Department of Economics and Center for Scientific Computing, Southern Methodist University, Dallas, TX, IZA and GLO, ozak@smu.edu; Martín: Nommon Solutions and Technologies, Madrid, and Department of Telematic Engineering, Universidad Carlos III, Madrid, ignacio.martin@nommon.es; Ortuño-Ortín: Department of Economics, Universidad Carlos III, Madrid, iortuno@eco.uc3m.es; Awad: Department of Economics, University of Exeter Business School, Exeter, e.awad@exeter.ac.uk; Cebrián: Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, cebrian@mpib-berlin.mpg.de; Rubén Cuevas: Department of Telematic Engineering and UC3M-Santander Big Data institute, Universidad Carlos III, Madrid, rcuevas@it.uc3m.es; Desmet: Department of Economics and Cox School of Business, Southern Methodist University, Dallas, TX, NBER and CEPR, kdesmet@smu.edu; Rahwan: Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, rahwan@mpib-berlin.mpg.de; Ángel Cuevas: Department of Telematic Engineering and UC3M-Santander Big Data institute, Universidad Carlos III, Madrid, acrumin@it.uc3m.es.

A Data appendix

This section starts by providing details on the process of collecting Facebook interests and on the calculation of Facebook distances. We then discuss the data sources for our other distance measures (genetic, linguistic, geographic, religious, WVS). In addition, we explain the methodology for creating dendrograms, for conducting principal component analysis, and for calculating and analyzing our regional divisiveness.

Facebook Marketing API

We collect data on nearly 60,000 Facebook interests across countries and territories, European subnational regions, U.S. states, California counties, and various demographic subgroups between 2017 and 2018 using Facebook's Marketing API (see `https://developers.facebook.com/docs/marketing-apis`, last accessed: April 2019). Note that for some experiments, such as the gender-region analysis or the US counties analysis, we used a subset of interests, since some interests in the original set did not have enough users when considering smaller user groups. This collection process yields a vector of the number of Facebook users in each entity that holds each interest, which in turn serves to create a vector containing entity-interest shares. Using this vector of interest shares, we compute distances between each group of interest.

Advertisers configure their ad campaigns on Facebook through Facebook Ads Manager which can be accessed through a dashboard that queries the Facebook Marketing API. This interface, which is also accessible to the public, allows advertisers to define the group they want to target with their advertising campaigns, i.e., the population of interest. The group specifications can include geographic location (country, region, city, zip code, latitude/longitude...), demographics (gender, age, language, education,...), behaviors (mobile device, operating system, browser,...), and interests (sports, food, cars, art,...). For a given set of group specifications, the Facebook Marketing API provides the number of monthly active users (MAU), daily active users (DAU), and different advertising costs (per click, per thousand visualizations, etc.).

For our analysis, we gather information on the number of users in each group who hold different interests. It is therefore important to clarify how Facebook assigns interests to individuals. A user is identified as having a particular interest based on their data and activity on Facebook, as well as on external websites, apps, and online services where Facebook has a presence. (Facebook has been estimated to have a presence on over 30% of popular websites)[1]. As individuals spend an increasing share of their time in the proximity of devices that track their location, Facebook also knows many dimensions of its users' offline activity, such as whether they go to football games, spend times downtown, or attend religious services.

Facebook's business model is based on identifying its users' true interests. Facebook's revenues depend crucially on the time its users spend on the platform, as this is what allows the company to show them relevant ads. Showing ads to users who are uninterested would negatively affect Facebook's bottomline. Advertisers are drawn to Facebook because of its ability to correctly identify groups of people interested in the products and services they wish to promote.

Once a user is identified as being interested in "cars", she will be included as a potential target of any advertising campaign configured to reach users interested in cars. There are hundreds of thousands of interests, spanning huge swaths of human preferences. To give an idea of the breadth and comprehensiveness, Facebook organizes interests in a multi-level hierarchical structure, with 14 categories in the first level: business and industry, education, family and relationships, fitness and wellness, food and drink, hobbies and activities,

lifestyle and culture, news and entertainment, people, shopping and fashion, sports and outdoors, technology, travel places and events, and null.

The interest extraction process done by Facebook relies on the information each user directly reveals as well as on proprietary inference on and off the platform. Users' interests are highly influenced by page likes and interests from other users with similar interests. In previous work, we have found six reasons for the assignment of interests and ad preferences[2]: (i) "This is a preference you added", (ii) "You have this preference because we think it may be relevant to you based on what you do on Facebook, such as pages you've liked or ads you've clicked", (iii) "You have this preference because you clicked on an ad related to...", (iv) "You have this preference because you installed the app...", (v) "You have this preference because you liked a page related to...", (vi) "You have this preference because of comments, posts, shares or reactions you made related to...".

In order to define and query groups, location is the only field mandatory in the API. Other parameters are accepted and may be combined to obtain a more specific group. Facebook offers in all their APIs (including the Marketing API) unique identifiers for most parameter values, so values are language-independent. To help advertisers, the Marketing API also provides a search function to retrieve interests by string matching: given some string, it returns a list of full and partial matching name items, typically containing the unique interest ID, name, total number of users worldwide (reach within Facebook) and, sometimes, a brief description.

Defining and downloading interests by populations

To collect the data for our analysis we queried the Facebook Marketing API more than 75M times across the following geographical areas: 225 countries and territories, 413 subnational regions and 58 California counties. In addition, for some countries and regions we also obtained information by gender and age groups. When querying the Facebook Marketing API, we obtain the number of Monthly Active Users (MAU) and the number of Daily Active Users (DAU) for each interest and geographical location. To avoid daily fluctuations, we prefer to employ MAU over DAU. However, for privacy reasons, Facebook imposes a lower bound of 1,000 users for its MAU measure, so any unit with less than 1,000 users will be reported as 1,000. To mitigate this problem, whenever we have a unit of interest that is at its lower bound, we substitute MAU by DAU.

To obtain a comprehensive list of interests, we use the Marketing API's targeting search function that returns partially or fully matched interests by name when given a string query. To feed this function with meaningful and representative interests, we use all article titles in the English Wikipedia and all entries in the English dictionary. The Wikipedia titles are obtained from the DBpedia project (see `https://wiki.dbpedia.org`, last accessed August 1, 2017), and the dictionary is the one contained in Ubuntu OS (Linux). This process yields 200 million records, including duplicates and non-interest parameters, such as demographics. After cleaning, the final number of unique interests is 399,182. From that collection, we select all those with a potential reach of at least 500,000 users worldwide. While we employ this process for purposes of cross-national comparability, there is no theoretical reason why other scholars couldn't employ the list of all 400,000 interests, if they were interested in examining interests that were unique to particular populations. Our method leaves us with nearly 60,000 unique interest identifiers.

Privacy

All data are aggregated at the level of population groups — countries or subnational regions. Thus, the data cannot be used to identify any specific individual. To give an example, we have information on the number of

users interested in cross-country skiing in Canada, but not on whether a specific individual user is interested in cross-country skiing. Since our dataset does not meet the definition of "personal data", informed consent is not required. Said differently, the dataset is anonymous and confidential with respect to the capacity of identifying an individual, and we never access any individual-specific information via our methods.

Fake accounts, bots and biases

Although there has been growing concern about the number of fake accounts, the vast majority of these are deleted by Facebook within minutes. More importantly, these deleted accounts do not count towards the Monthly Active Users (MAU) metric, so they do not affect our data. Overall, Facebook estimates that undetected fake accounts represent around 5% of its worldwide monthly active users in 2020.[3] This number is too small to create significant distortions to our measures.

Some papers have pointed out certain gender or racial biases in Facebook's ad delivery algorithm.[4] For example, Facebook might deliver an ad more frequently to women, even if the advertiser did not use gender in defining its target audience. However, these findings do not apply to the separate algorithm that assigns interests to users. Our data are solely based on users' interests, which are not determined by the ad delivery algorithm. That said, Facebook's ultimate goal in identifying its users' interests is for the purpose of advertising, rather than for the purpose of measuring culture. As such, we cannot completely rule out that such biases are also present in the algorithm that identifies users' interests.

Representativeness

While most countries impose almost no restrictions on what content is available nor on how each citizen relates online, a few countries do limit the use of Facebook or the Internet. In the cases of Iran, Sudan and Cuba, the Marketing API does not provide information on people residing in these countries, returning a 2641 error with the following message: "By law, we cannot serve ads to the following countries." Another special case is China. Although it is possible to obtain user counts for China, Internet access to sites outside the country is severely restricted, making the figures reported by Facebook unrealistic. Hence, we do not include these countries in our analysis.

We also collect age and gender data from the World Bank's World Development Indicators (WDI) and compare them to the age and gender structure of Facebook users by country. In *Appendix B* we show that our qualitative findings are robust to limiting the analysis to either more representative or less representative subsamples, both in terms of gender and age structure and in terms of Facebook penetration.

Facebook cultural distances

Given data on the number of Facebook users and the number of users interested in each of the 59,763 Facebook interests in each of the 225 countries and territories, we construct a matrix of size (225 x 59,763) with the share of Facebook users that have a given interest in each country. Thus, the element in row k and column i of the matrix gives the share of Facebook users in location k that have interest i, and each row vector has the shares of users for all interests for a given location.

The Facebook distance between two populations is computed as the cosine distance between the vectors of Facebook interest shares of the two populations. Consider two population groups, k and l. The interests of population group k can be represented by an n-dimensional vector S_k with components s_{ik} that measure the share of population k that holds a particular interest i, where $i = 1, ..., n$. Similarly, vector S_l represents the

interests of population group l. Denote the angle between S_k and S_l by θ. The cosine distance between the interests of groups k and l is then:

$$\cos \text{dist}\,(k,l) = 1 - \cos(\theta) = 1 - \frac{S_k \cdot S_l}{\|S_k\| \, \|S_l\|} \tag{1}$$

Since the cosine distance is based on the angle between two vectors, it does not depend on differences in the lengths of the vectors. In our context this is an advantage, because the norms of the vectors differ substantially across countries in a non-systematic way. For example, Spain and Italy, two countries that are similar in size, economic development and Facebook penetration, have vectors of very different lengths. There are many reasons why such differences may arise. For example, Facebook may be more easily able to identify interests in one country than in another, or people's intensity of using Facebook may differ across countries. Since we would not want such differences to be driving our measure of cultural distances, we want a distance measure that does not depend on the norms of the interest vectors. The cosine distance achieves this goal.

As an alternative, we could use the Euclidean distance, based on interest shares that are normalized by the length of the interest vectors. After defining the normalized vector as $S_k^{`} = \frac{S_k}{\|S_k\|}$, the normalized Euclidean distance is a simple transformation of the cosine distance:

$$\textit{norm euc}\ \text{dist}\,(k,l) = \left\|S_k' - S_l'\right\| = \sqrt{\left\|S_k'\right\|^2 + \left\|S_l'\right\|^2 - 2\left\|S_k'\right\|\left\|S_l'\right\|\cos(\theta)} = \sqrt{2\cos \text{dist}\,(k,l)} \tag{2}$$

The correlation between cosine distance and normalized Euclidean distance for the countries in our sample is 0.97.

For the reasons mentioned above, other distance measures that do depend on differences between the lengths of the vectors may be less appropriate. However, reassuringly, even if we use Facebook distances based on other metrics (non-normalized Euclidean, Manhattan), these are highly correlated with cosine distances in our dataset ($r > 0.73$, $p < 0.0001$). Figure B13 reports our main results based on Euclidean distances. Qualitatively, the results are unchanged.

As is usual with correlations between distance matrices, all confidence intervals in this paper are based on Mantel tests.[5] All computations were done in Python 3.5 using distance measures implemented in *scikits-learn* version 0.19.2.[6] In addition to calculating distances between countries, we also build matrices of Facebook distances for sub-national regions, age and gender groups within countries, as well as California counties.

Other distance measures

Cultural distances based on the World Values Survey are derived from 98 questions across 74 countries, spanning the period 1981-2010. Specifically, we follow Spolaore and Wacziarg[7] and use the set of questions that is most common across countries and years, but unlike these authors we use cosine distance as the benchmark distance measure (the correlations among the various distances measures based on this set of questions for different metrics – cosine, Euclidean, Manhattan – is also very high: $r > 0.97$, $p < 0.0001$). In addition to distances based on all 98 questions, we also include a number of alternative measures of cultural distances, based on subsets of questions. These include measures for different macro-categories of questions (such as perceptions of life, family, work, etc.), as well as for binary and non-binary questions.

Genetic distances come from Spolaore and Wacziarg[8] and are based on genetic data from Creanza et al.[9]

They measure population-weighted F_{ST} genetic distances between countries. As additional measures, we also include F_{ST} genetic distances between the plurality groups and genetic distances as they were in the year 1500, as well as additional distance measures based on Spoloare and Wacziarg[7].

Geographic distances measure distances between capitals. We also include a number of alternative measures: distances between the most populated cities, population-weighted distances, and population-weighted distances that account for the sensitivity of trade flows to distance. All these measures come from the CEPII GeoDist database.[10]

Linguistic distances measure the distance between two randomly drawn individuals of two different countries, where the linguistic distance is based on the Ethnologue language tree and uses the formula by Desmet, Ortuño-Ortín and Wacziarg[11]. We also include alternative measures of linguistic distances taken from Spolaore and Wacziarg:[8] 15 additional measures where languages are defined at 15 different levels of aggregation (e.g., at aggregation level 1 all Indo-European languages are taken to be the same), and two distance measures based on cognates, one for plurality languages and another weighted by the language shares.

The measure for religious distances is based on population-weighted distances of religions using a religion tree from the World Christian Database[7]. We also include alternative measures: a distance measure based on plurality groups, and two measures based on an alternative religion tree from Mecham, Fearon and Laitin.[12]

Principal component analysis

We perform principal component analysis on Facebook interests and on WVS questions, using the common sample of countries covered by both data sources. The goal is to reduce the dimensionality of interests and questions, and assess whether a limited number of principal components are able to explain a large share of the variance in Facebook interests across countries and a large share of the variance of WVS questions across countries. When conducting principal component analysis, we use the covariance matrix, since our variables are population shares, measured on the same scale, from 0 to 1. In *Appendix B* we show that the results are similar when using the correlation matrix instead.

Dendrograms

To generate the dendrograms, we use an implementation of hierarchical agglomerative clustering (HAC) that is provided by the python package *scipy.cluster.hierarchy*. This implementation of HAC allows for a choice of a distance method and a linkage criterion. We use cosine distance for consistency across other analyses and employ the Ward variance minimization linkage criterion. The input is the interest share vector of each country.

Hierarchical clustering algorithms are a general family of clustering algorithms that build nested clusters by merging or splitting the clusters successively in iterations. The nested structure can be represented as a dendrogram, in which the root of the tree is the highest level cluster that includes all data, while the leaf of the tree represents the lowest level cluster that includes a single data point.

Hierarchical clustering can be performed through a bottom-up (agglomerative) strategy or a top-down (divisive) strategy. In the agglomerative strategy, hierarchical clustering is performed via a bottom-up process in which every data point starts as its own cluster. The algorithm builds nested clusters by merging or splitting the clusters successively in multiple iterations and represents the nested structure as a dendrogram.

Adopting notations from Arbelaitz et al.[13], a data set DS is defined as a set of N data points, each of which

is a vector of F dimensions. A clustering of DS is a set of disjoint clusters that partitions DS into K groups

$$C = \{c_1, c_2, \ldots, c_K\} \text{ where } \forall K > 1 \cup_{c_k \in C} c_k = DS \text{ and } \cap_{c_k \in C} c_k = \varnothing$$

In the first iteration, $\binom{K}{2}$ distance values between base K clusters are computed using a distance metric. After the distance values have been evaluated, a linkage criteria is used to determine two clusters c_i and c_j to combine to form a new cluster c_k. The Ward variance minimization linkage criterion is given by the following formula:

$$d(c_k, c_l) = \sqrt{\frac{|c_l| + |c_i|}{T} d(c_l, c_i)^2 + \frac{|c_l| + |c_j|}{T} d\left(c_l, c_j\right)^2 - \frac{|c_l|}{T} d\left(c_i, c_j\right)^2}$$

where c_k is a newly formed cluster consisting of clusters c_i and c_j, c_l is an unused cluster, $T = |c_l| + |c_i| + |c_j|$ and $|\cdot|$ is the cardinality of its argument.

Dendrogram of full set of countries

Figure 3 depicts a dendrogram based on the set of 72 countries common to both the WVS and Facebook sample. Appendix Figure B16 shows a dendrogram for the full set of 225 countries for which we have Facebook interest shares.

Regional, gender, and age divisiveness

Our measure of regional divisiveness is as follows. Consider a country c with a number of regions indexed by k or l. Denote the Facebook cosine distance between any two regions k and l by cos dist (k, l), and denote the population shares of k and l by s_k and s_l. We then measure the regional divisiveness of country c as the population-weighted Facebook cosine distance between its different subnational regions:

$$\frac{\sum_{k \neq l} \sum_l \cos \text{dist}\,(k, l)\, s_k s_l}{\sum_{k \neq l} \sum_l s_k s_l} \tag{3}$$

One way of interpreting (3) is as the expected Facebook distance between two randomly drawn individuals from different regions. If there are only two regions in a country, then (3) would simplify to the bilateral distance between them, i.e., $\cos \text{dist}(k, l)$. More simply, (3) is an average interregional distance measure. Gender and age divisiveness are measured similarly.

Country networks and maps

From the 18 countries for which we downloaded regional data, we chose Germany and India to show network and map, given their rank according to divisiveness in Figure 6A. For each country, we first constructed a network in which each node represents a region, and we used links to connect every pair of regions. Links are weighted by the cosine similarity between regions (1 - cosine distance). Link weights are rescaled using standardization and then shifted by a constant value as to have only non-negative weights (a condition on the input for the community detection method). Then, to detect communities, we used an *igraph* (from *R*) implementation of *multi-level modularity optimization* algorithm (also known as *Louvain* method)[14], which is based on the modularity measure and a hierarchical approach. Modularity is a measure of a quality of a partition (codomain: [-1, 1]), which measures the density of links inside communities as compared to links between communities[15,16]. Then, we colored nodes according to community affiliation, and links according to

their adjacent nodes. To plot country maps, we used *R* package *rnaturalearth* to access geometry of regions' layout from *Natural Earth* map data[17]. We colored map regions according to the communities detected from the corresponding network.

B Supplementary analysis

B.1 Partial correlations of distance measures

To further explore whether Facebook distances capture cultural distances, we look at various partial correlations. That is, we analyze the correlation between one distance measure (e.g., WVS distances) and the Facebook distance measure, controlling for all other distance measures (e.g., genetic, geographic, linguistic and religious distances). The goal is to discover which type of distance measure correlates most strongly with FB distances.

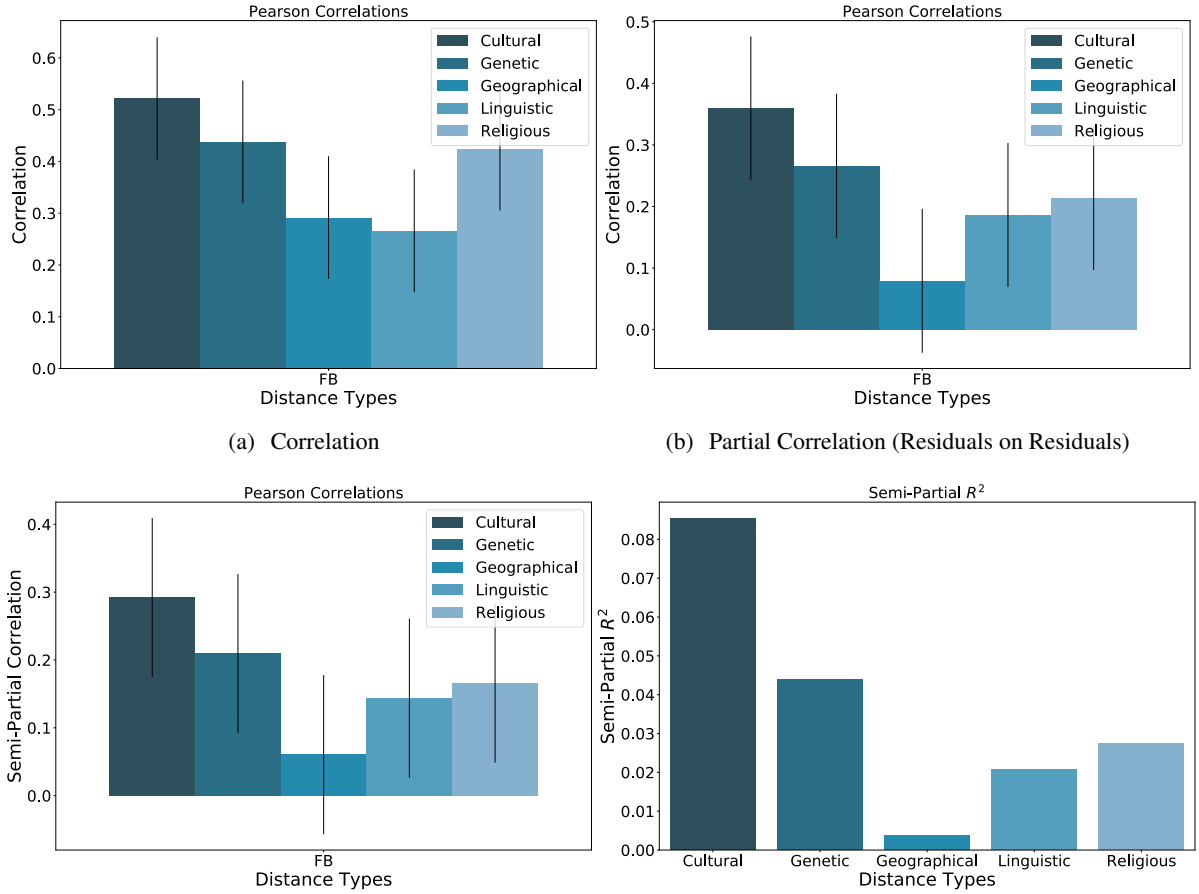
We start by focusing on one measure for each type of distance proxy. That is, rather than using different ways of measuring each distance proxy, we choose one measure for each one of the five distance proxies (values, genetic, geographic, linguistic, and religious).[1] Before looking at partial correlations, Figure B1(a) plots the correlations between Facebook distances and each one of the distance measures, not controlling for any other distance. The correlations of Facebook distances with genetic, geographic, linguistic and religious distances are all positive and statistically significant at the 95% confidence level, but the strongest correlation continues to be with the most direct survey-based measure of cultural distances.[2] This confirms that our bottom-up Facebook measure of cultural distance corresponds well to the standard top-down measure of cultural distance.

Figure B1(b) plots the partial correlations between Facebook distances and each one of the distance measures. To give a specific example, consider the partial correlation between Facebook distances and value-based cultural distances (slightly above 0.35). This number represents the correlation between the residuals of a regression of Facebook distances on all other distances (genetic, geographic, linguistic and religious) and the residuals of a regression of value-based cultural distances on all other distances (genetic, geographic, linguistic and religious). It hence tells us how correlated Facebook and value-based cultural distances are, after controlling for all other distances. The same partial correlations with other distances are all lower: for example, the partial correlation between Facebook and geographic distances is below 0.1, and not statistically significant at the 95% confidence level. Hence, when controlling for all other distances, the strongest partial correlation is between Facebook distances and value-based cultural distances. This shows that Facebook distances are not just picking up geographic, genetic, linguistic or religious distances.

The other two panels of Figure B1 confirm this finding. Panel (c) shows the semi-partial correlations between Facebook distances and each one of the distance measures. This represents the correlations between Facebook distances and the residuals of a regression of one of the distance measures on all other distance measures. Once again, we find that the strongest semi-partial correlation is with survey-based cultural distances.

[1] In particular, value distances are based on 98 questions from the World Values Survey, spanning the period 1981- 2000, as in [7], with the only difference that we use cosine distance; genetic distances come from [8] and measure population-weighted F_{ST} genetic distances between countries using genetic data by [9]; geographic distances are between country capitals; linguistic distances are based on the Ethnologue database and use the formula by [11] to measure the linguistic distance between two randomly drawn individuals of two different countries; and religious distances are based on the population-weighted distance using a religion tree from the World Christian Database [7].

[2] As is usual with correlations between distance matrices, all confidence intervals are based on Mantel tests[5].



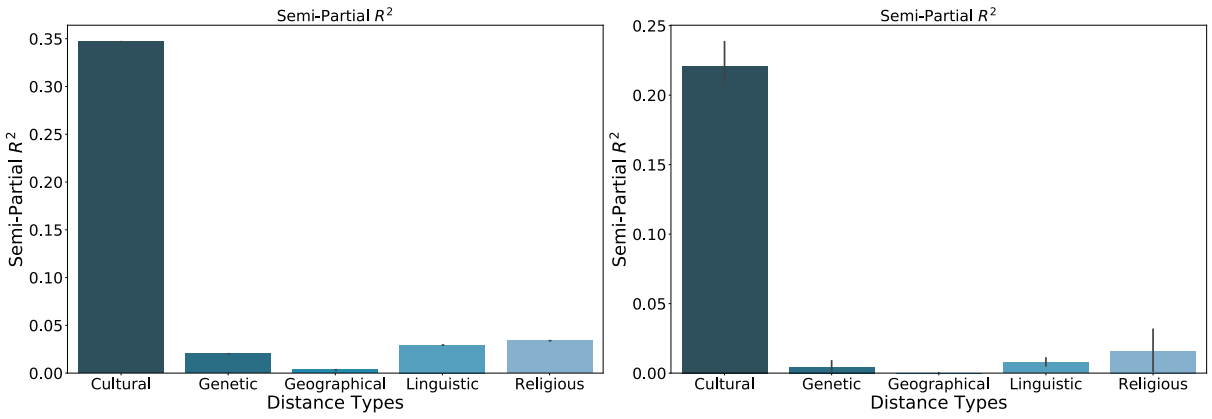
(a) Correlation

(b) Partial Correlation (Residuals on Residuals)

(c) Semi-Partial Correlation (Actual on Residuals)

(d) Semi-Partial R^2

Figure B1: ***Correlations Between Facebook and Selected Distance Measures.*** *Panel (a) plots the correlations between Facebook distances and each one of the distance measures, not controlling for any other distance. Panel (b) plots the partial correlations between Facebook distances and each one of the distance measures. For example, the partial correlation between Facebook distances and value-based distances corresponds to the correlation between the residuals of a regression of Facebook distances all other distances (genetic, geographic, linguistic and religious) and the residuals of a regression of value-based cultural distances on all other distances (genetic, geographic, linguistic and religious). Panel (c) plots the semi-partial correlations between Facebook distances and each one of the distance measures. For example, the semi-partial correlation between Facebook distances and value-based distances corresponds to the correlation between Facebook distances and the residuals of a regression of value-based cultural distances on all other distances (genetic, geographic, linguistic and religious). Panel (d) plots the semi-partial R^2 of Facebook distances on each one of the distance measures. For example, the semi-partial R^2 between Facebook and value-based cultural distances corresponds to the difference in R^2 of a regression of Facebook distances on all other distance measures (including value-based cultural distances) and a regression of Facebook distances on all other distance measures (excluding value-based cultural distances). The distance measures for each proxy (value-based, genetic, geographic, linguistic and religious) are the ones given in footnote 1.*



(a) Semi-Partial R^2 (Joint)

(b) Semi-Partial R^2 (Individual)

Figure B2: ***Semi-Partial R² of Facebook Distance on All Distance Measures.*** *Panel (a) plots the semi-partial R^2 of Facebook distances on all alternative measures of each one of the distance proxies. For example, the semi-partial R^2 between Facebook and value-based cultural distances corresponds to the difference in R^2 of a regression of Facebook distances on all alternative distance measures of the five proxies (including all alternative measures of value-based cultural distances) and a regression of Facebook distances on all alternative measures of the distance proxies (excluding all alternative measures of value-based cultural distances). Panel (b) plots the average semi-partial R^2 of Facebook distances on alternative measures of each one of the distance proxies. For example, the average semi-partial R^2 between Facebook and value-based cultural distances corresponds to the average difference in R^2 of a regression of Facebook distances on all alternative distance measures of the proxies (including one of the measures of value-based cultural distances) and a regression of Facebook distances on all alternative measures of the distance proxies (excluding value-based cultural distances). The alternative distance measures for each proxy (value-based, genetic, geographic, linguistic and religious) are the ones given in Figure 2 of the main paper, and further described in the Methods section.*

Panel (d) reports the semi-partial R^2 of Facebook distances on each one of the distance measures, after controlling for all others. For example, the semi-partial R^2 of 0.085 between Facebook and value-based cultural distances means that a regression of Facebook distances on all other distance measures (including value-based cultural distances) explains 8.5% more of the variation in Facebook distances than a regression that excludes value-based cultural distances.

We now turn to using more than one measure for each type of distance proxy. As in the main paper, we use all alternative measures of the five distance proxies (values, genetic, geographic, linguistic, and religious). Figure B2 reports the results. To explain the difference between the graphs, we focus on the first bar of each graph. The first bar in Panel (a) shows the semi-partial R^2 of a regression of Facebook distances on all value-based cultural distances, after controlling for all other measures of distance (genetic, geographic, linguistic and religious). The first bar in Panel (b) shows the average semi-partial R^2 of a regression of Facebook distances on each value-based cultural distance separately, after controlling for all other measures of distance (genetic, geographic, linguistic and religious). All other bars in the two panels show the same information, but for genetic, geographic, linguistic and religious distances. These graphs confirm our main finding: although all distance measures partly explain Facebook distances, survey-based cultural distances have the strongest explanatory power.

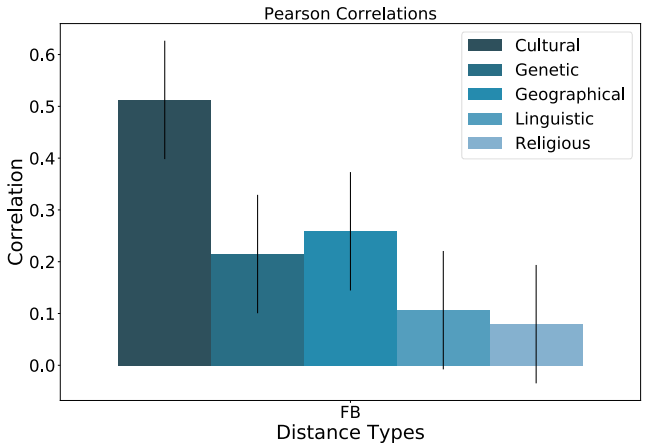


Figure B3: ***Correlations Between Facebook and Selected Distance Measures, Using Full Sample.*** *This figure plots the correlation between Facebook distances and select measures of each of the distance proxies (value-based, genetic, geographic, linguistic and religious). It uses the full sample of countries, rather than the sample that is common to all measures. The select measures of the distance proxies are the ones given in footnote 1.*

B.2 Robustness to alternative samples

In this section we explore the robustness of our analysis to various samples. First, Figure B3 plots the correlation between Facebook distances and select measures of each of the distance proxies, using the full sample of countries rather than the common sample. The results confirm that the strongest correlation with Facebook distances are value-based distances from the WVS. Second, Figure B4 shows the results of replicating the main analysis when we constrain the sample to countries with more than 300,000 people and a Facebook penetration above 5%. This decreases the number of countries for which we have Facebook distances from 225 to 161. The results are quantitatively and qualitatively very similar. This is not surprising: most of the countries that drop out were not in the common sample. Third, to further explore representativity based on Facebook penetration, we create two groups of countries: a more representative group with Facebook penetration above the median, and a less representative group with Facebook penetration below the median. Figures B5 and B6 show that our results do not qualitatively change.

Fourth, we explore the representativity of Facebook users in terms of gender. For this analysis, we start by comparing, for each country, its Facebook user composition in terms of gender to the composition of its actual gender composition using data from the World Development Indicators in 2017. Starting from the common sample, we create two groups of countries: a first group of countries that are above the median difference in gender composition when comparing Facebook users and the actual population, and a second group of countries that are below the median difference in age composition. In the first group Facebook is less representative of the population than in the second group. Figures B7 and B8 show the results of splitting the sample based on gender composition. Reassuringly the results are similar to the main analysis.

Fifth, we explore the representativity in terms of age composition. Focusing on the population aged 15-64, we compare the share of FB users aged 15-29 to the share of the actual population aged 15-29. We select this age split because on average about half of the Facebook users aged 15-64 are in the group 15-29. As before, we create two groups of countries: a less representative group of countries that are above the median difference in age composition when comparing Facebook users and the actual population, and a more representative group

of countries that are below the median difference in age composition. Figures B9 and B10 show that similar results are obtained if we split the sample based on age composition.

In addition to exploring robustness to different samples of countries, we also analyze robustness to different samples of interest categories. Figure B11 plots the correlation between Facebook distances and each one of the distance measures for each one of the 14 macro-categories of interests: people; lifestyle and culture; travel, places and events; empty; hobbies and activities; news and entertainment; shopping and fashion; business and industry; food and drink; sports and outdoors; education; technology; fitness and wellness; and family and relationships. Quite a few interests are marked by Facebook as a local business. To ensure differences between countries are not driven by such local businesses, we add a robustness check that focuses exclusively on interests that are not marked as local businesses. As can be seen in Figure B11, the results do not differ substantially across these different sub-samples of interests. Similarly, Figure B12 displays the division of Germany, based exclusively on interests that are not marked as local businesses.

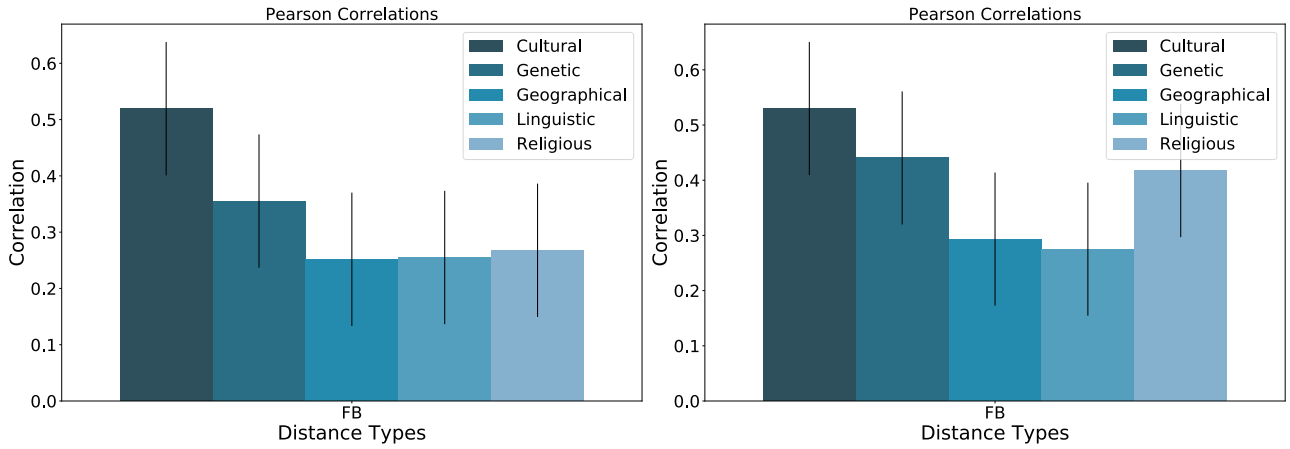
B.3 Robustness to other distance measures

Figure B13 reports our main results based on normalized Euclidean distances, rather than cosine distance. Since the former is a simple transformation of the latter, our findings are unchanged.

B.4 Alternative methods for principal component analysis

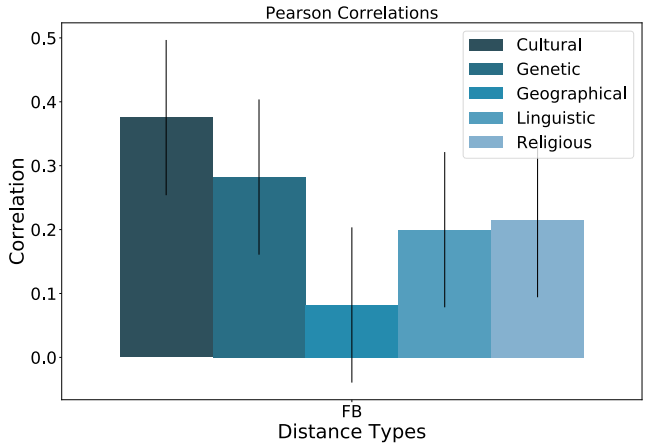
As mentioned in the *Data appendix*, principal component analysis can be done on the covariance matrix or on the correlation matrix. When using the correlation matrix, we are standardizing the population share for each one of the interests or questions, whereas when using the covariance matrix, we are not. In the context of our data it is not obvious whether one should standardize or not. Standardization is often done to make the variables scale-independent. Since our variables are population shares, they are already measured on the same scale, from 0 to 1. Of course standardization still matters, since there are interests or questions with very low average shares, and others with very high average shares. Standardizing puts equal weight on all interests, whereas not standardizing puts greater weight on interests with larger average shares.

Figure 2C in the main text is based on the covariance matrix, whereas Figure B14 shows results for both the correlation and the covariance matrix. In particular, Figure B14 plots the share of the overall variance in questions and interests that is explained by principal components as a function of their number: the left panel is based on the correlation matrix (i.e., standardized population shares), whereas the right panel is based on the covariance matrix (i.e., non-standardized population shares). Focusing on the standardized shares, the first ten principal components of FB explain slightly more than 40% of the overall variance in FB interests, whereas the first ten principal components of WVS explain slightly less than 70% of the variance in WVS questions. When using non-standardized shares, the share of the variance that is explained by the first ten principal components increases by about ten percentage points, to slightly less than 60% in the case of FB and to around 80% in the case of WVS. From this we conclude that FB captures more dimensions of culture than the WVS.

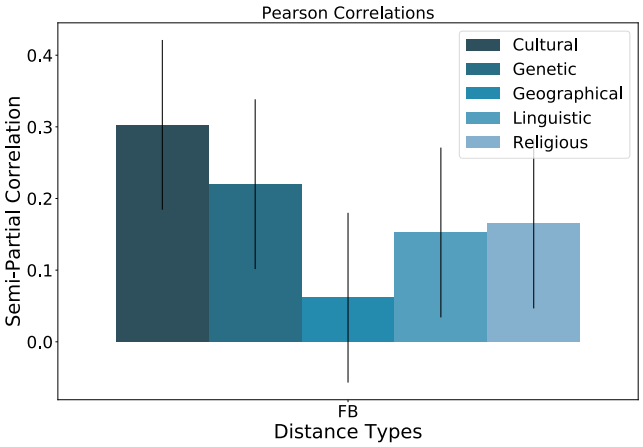


(a) Full Sample

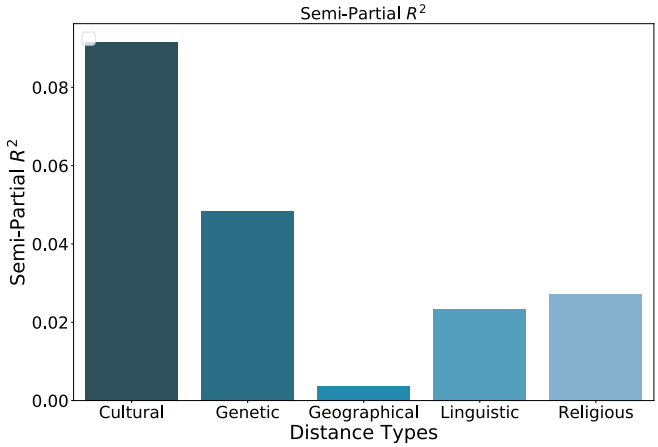
(b) Common Sample



(c) Partial Correlation (Residuals on Residuals)

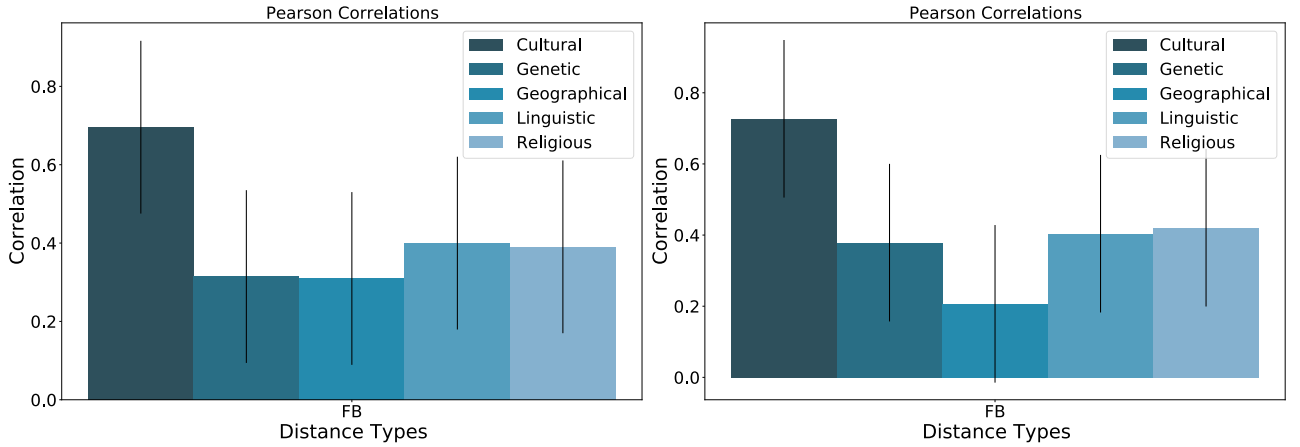


(d) Semi-Partial Correlation (Actual on Residuals)



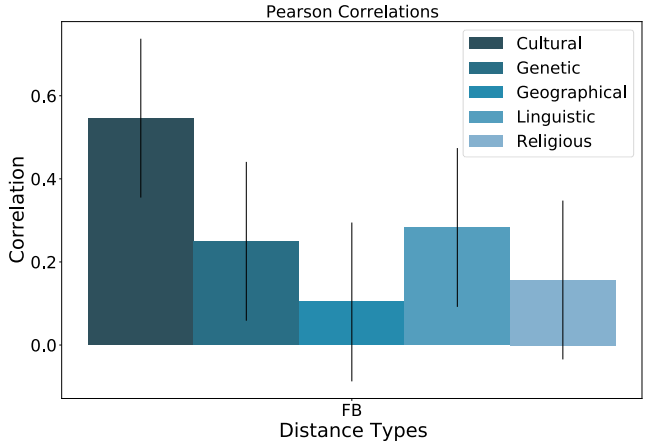
(e) Semi-Partial R^2

Figure B4: ***Correlations Between Facebook and Selected Distance Measures, Robustness to Population Size and Facebook Penetration.*** *This figure shows the same information as Figure B1 and B2, with one difference: it does not include the countries with a population of less than 300,000 and a Facebook penetration of less than 5%. For that different sample, Panel (a) corresponds to Panel (a) of Figure B2, and Panels (b) through (e) correspond to Panels (a) through (d) of Figure B1.*

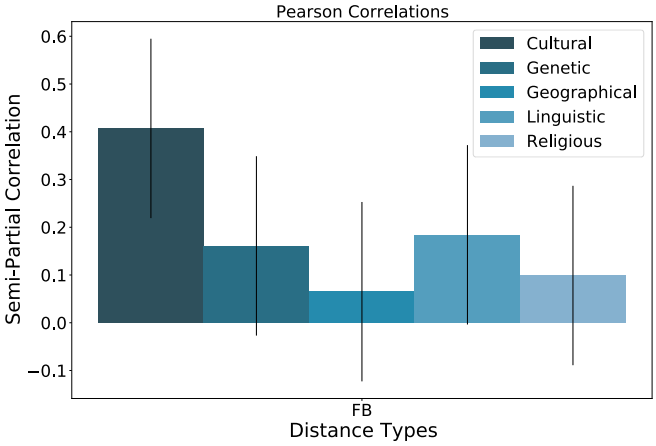


(a) Full Sample

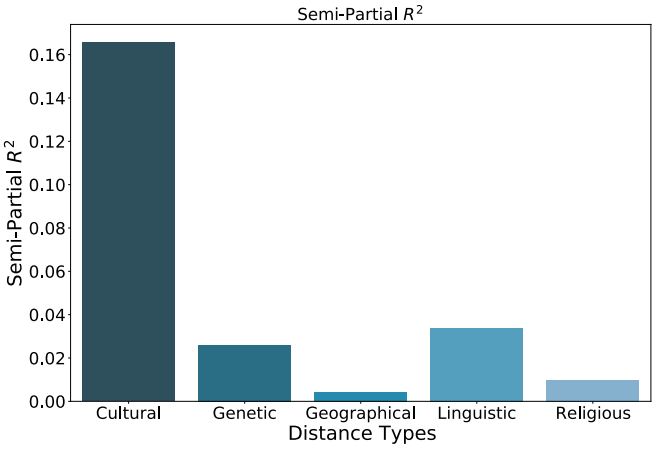
(b) Common Sample



(c) Partial Correlation (Residuals on Residuals)

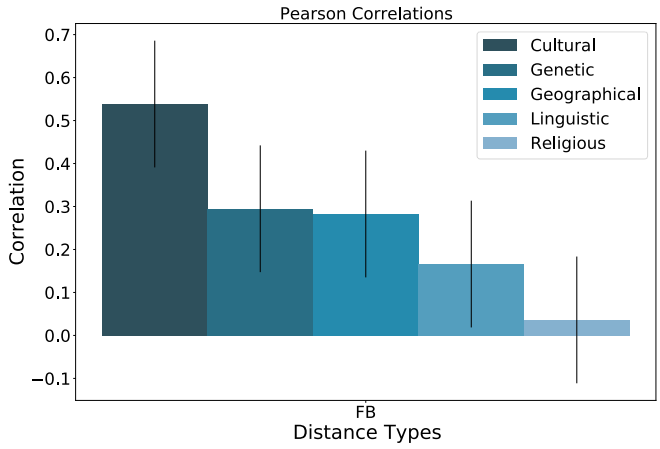


(d) Semi-Partial Correlation (Actual on Residuals)

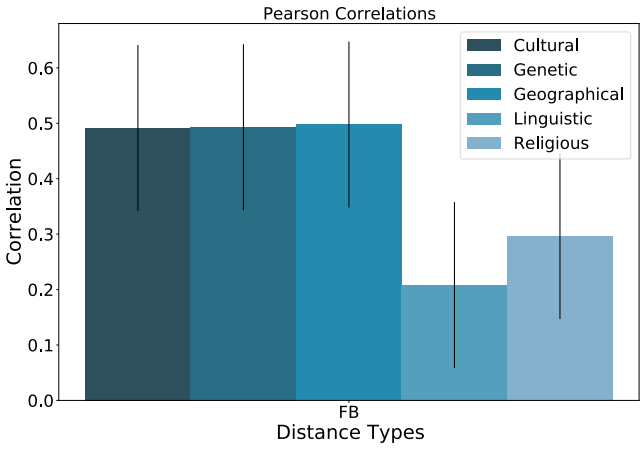


(e) Semi-Partial R^2

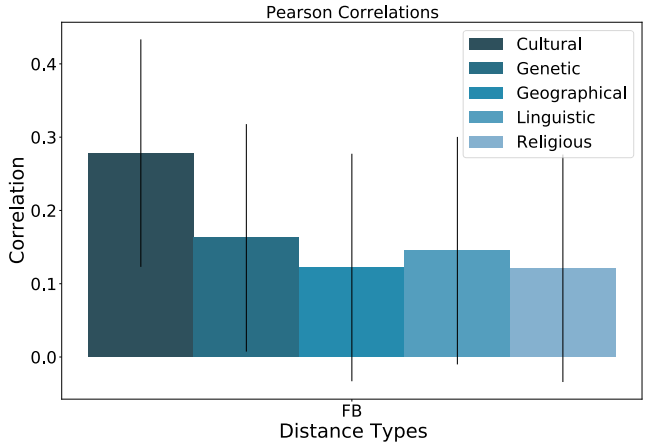
Figure B5: ***Correlations Between Facebook and Selected Distance Measures, Robustness to Representativity (Facebook penetration above Median).*** *This figure shows the same information as Figure B1 and B2, with one difference: using the common sample, it only retains the countries where the penetration of Facebook users is above the median of all countries. Hence, it focuses on the subset of countries where Facebook users are more representative of the actual population in terms of penetration. For that sample, Panel (a) corresponds to Panel (a) of Figure B2, and Panels (b) through (e) correspond to Panels (a) through (d) of Figure B1.*



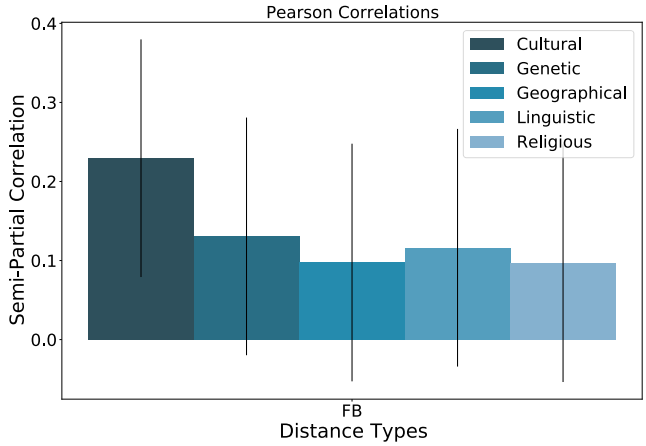
(a) Full Sample



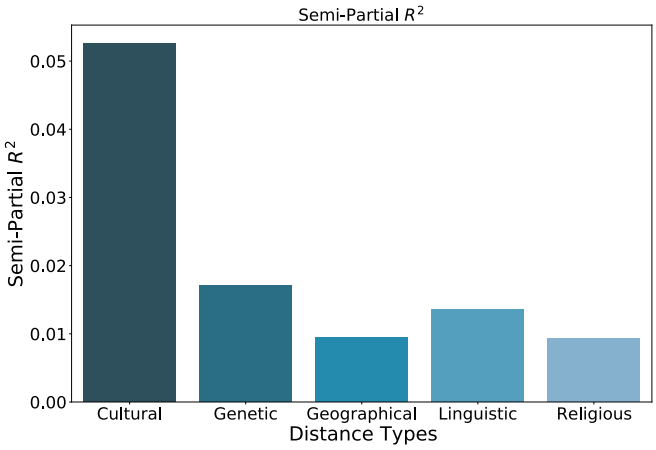
(b) Common Sample



(c) Partial Correlation (Residuals on Residuals)

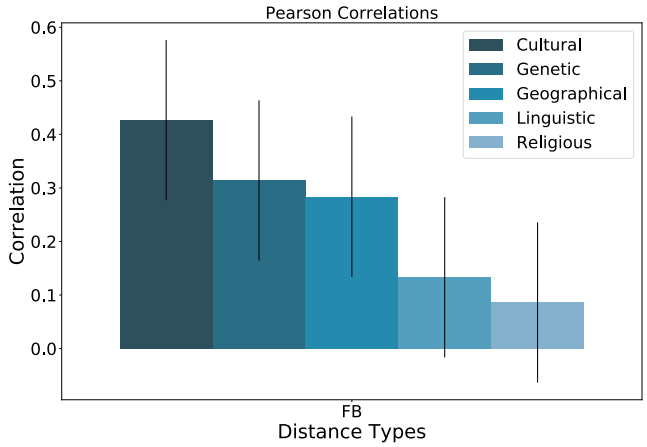


(d) Semi-Partial Correlation (Actual on Residuals)

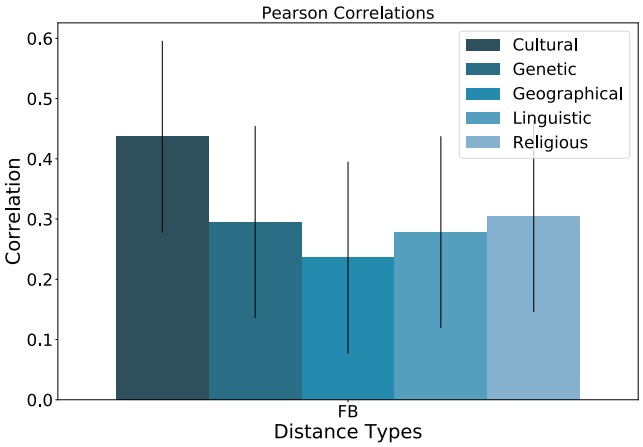


(e) Semi-Partial R^2

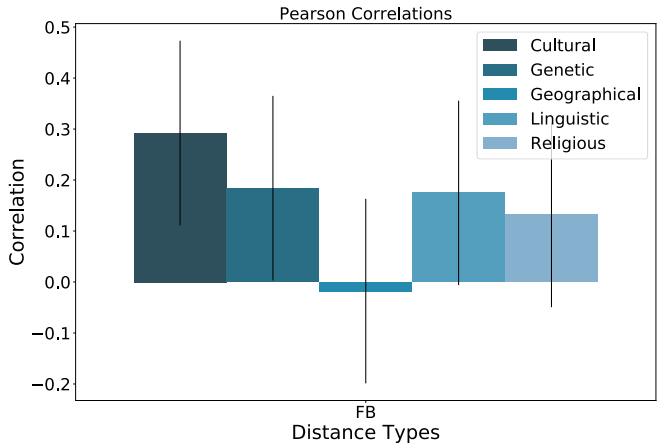
Figure B6: ***Correlations Between Facebook and Selected Distance Measures, Robustness to Representativity (Facebook penetration below Median).*** *This figure shows the same information as Figure B1 and B2, with one difference: using the common sample, it only retains the countries where the penetration of Facebook users is below the median of all countries. Hence, it focuses on the subset of countries where Facebook users are least representative of the actual population in terms of penetration. For that sample, Panel (a) corresponds to Panel (a) of Figure B2, and Panels (b) through (e) correspond to Panels (a) through (d) of Figure B1.*



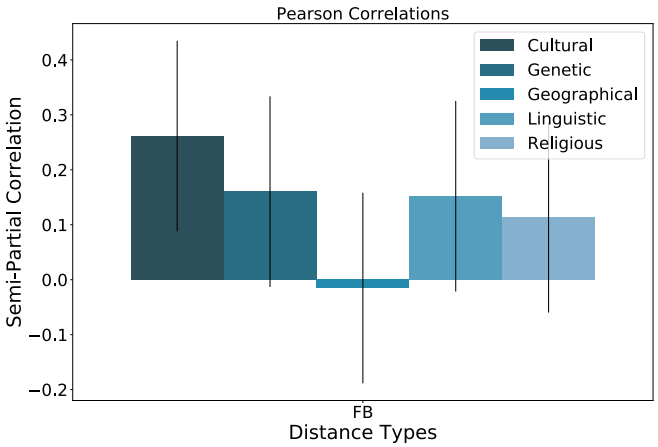
(a) Full Sample



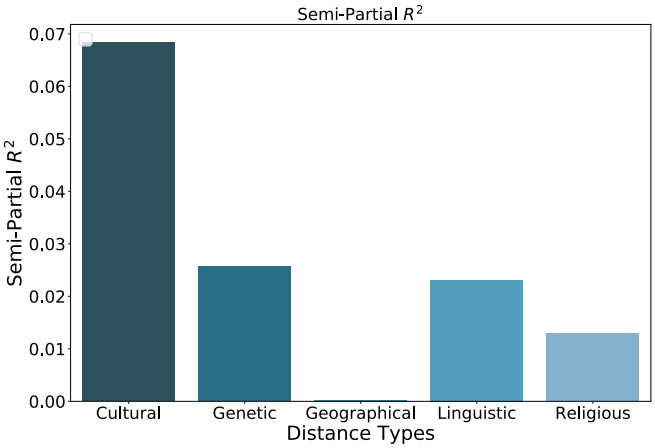
(b) Common Sample



(c) Partial Correlation (Residuals on Residuals)

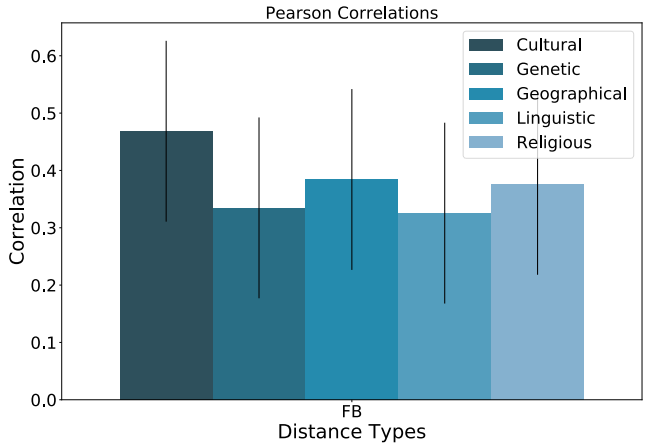


(d) Semi-Partial Correlation (Actual on Residuals)

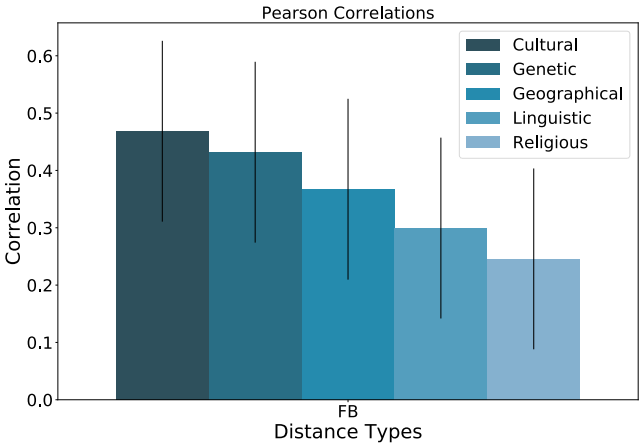


(e) Semi-Partial R^2

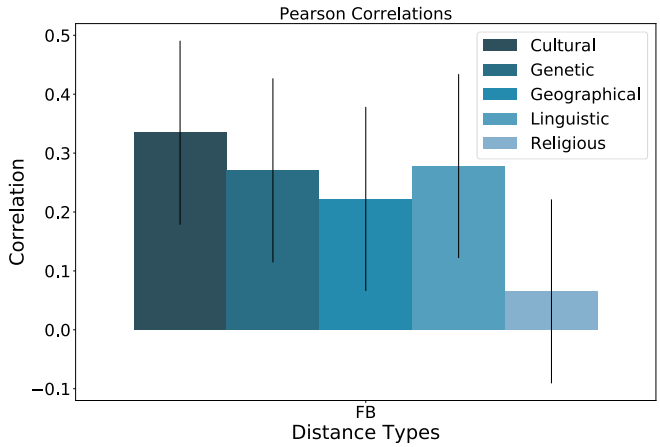
Figure B7: ***Correlations Between Facebook and Selected Distance Measures, Robustness to Representativity (Gender Difference above Median).*** *This figure shows the same information as Figure B1 and B2, with one difference: using the common sample, it only retains the countries where the difference in gender composition of Facebook users and the actual population is above the median. Hence, it focuses on the subset of countries where Facebook users are least representative of the actual population in terms of gender. For that sample, Panel (a) corresponds to Panel (a) of Figure B2, and Panels (b) through (e) correspond to Panels (a) through (d) of Figure B1.*



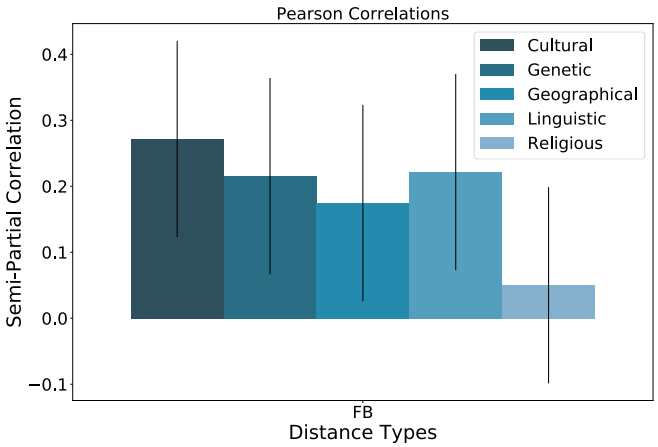
(a) Full Sample



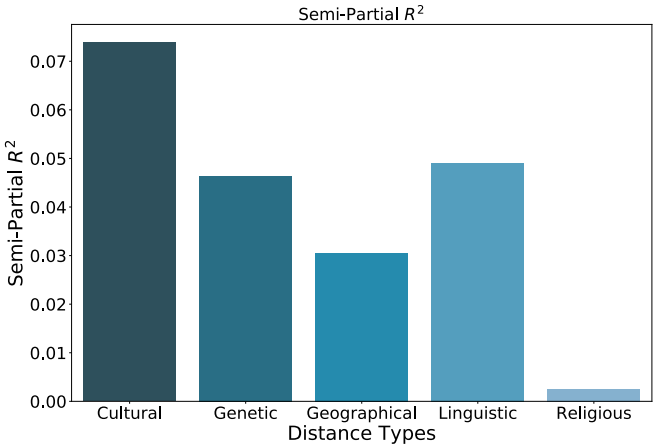
(b) Common Sample



(c) Partial Correlation (Residuals on Residuals)

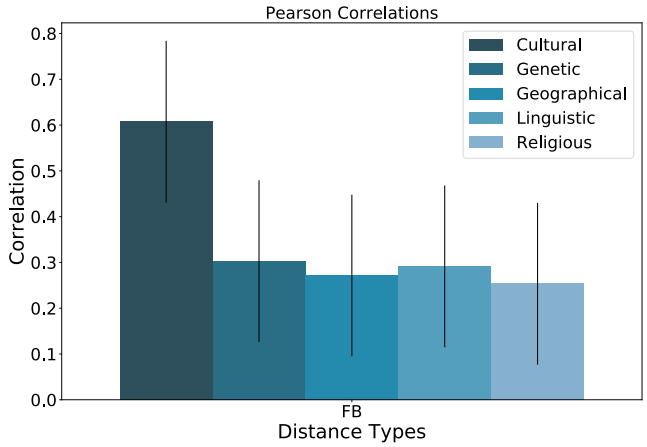


(d) Semi-Partial Correlation (Actual on Residuals)

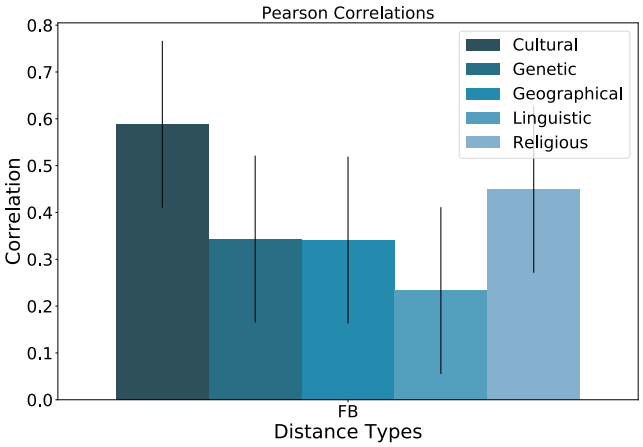


(e) Semi-Partial R^2

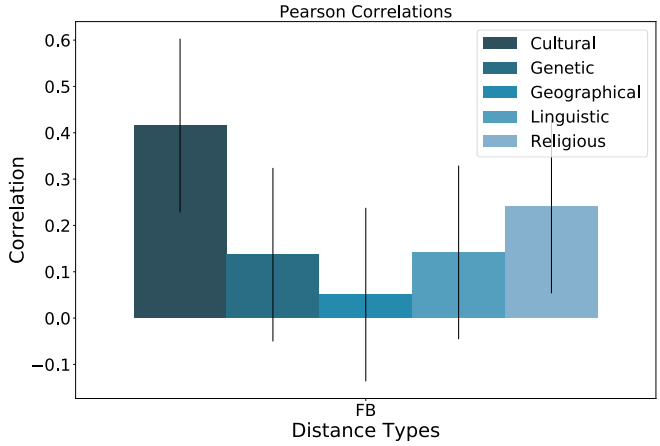
Figure B8: ***Correlations Between Facebook and Selected Distance Measures, Robustness to Representativity (Gender Difference below Median).*** *This figure shows the same information as Figure B1 and B2, with one difference: using the common sample, it only retains the countries where the difference in gender composition of Facebook users and the actual population is below the median. Hence, it focuses on the subset of countries where Facebook users are most representative of the actual population in terms of gender. For that sample, Panel (a) corresponds to Panel (a) of Figure B2, and Panels (b) through (e) correspond to Panels (a) through (d) of Figure B1.*



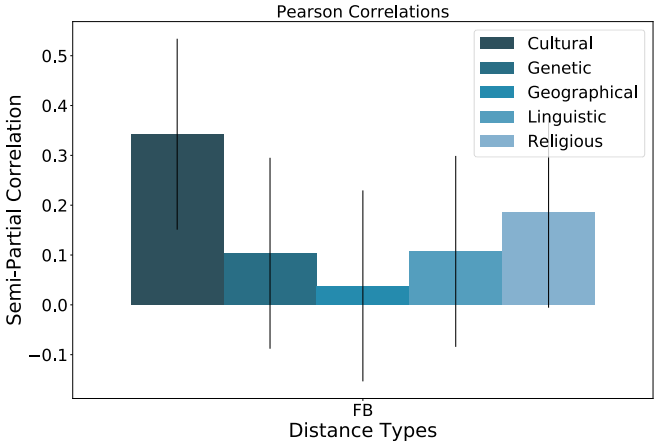
(a) Full Sample



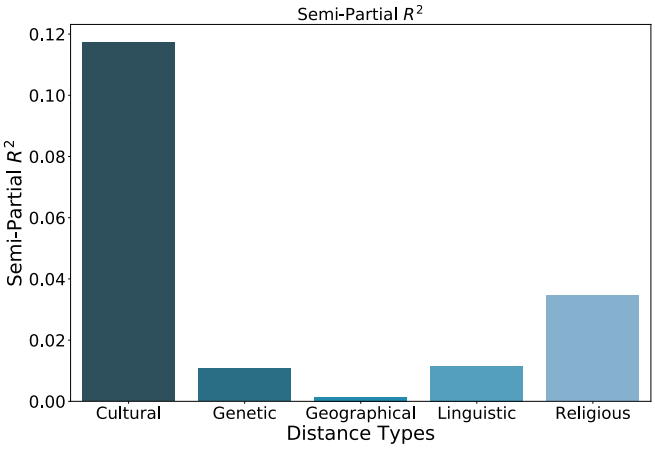
(b) Common Sample



(c) Partial Correlation (Residuals on Residuals)

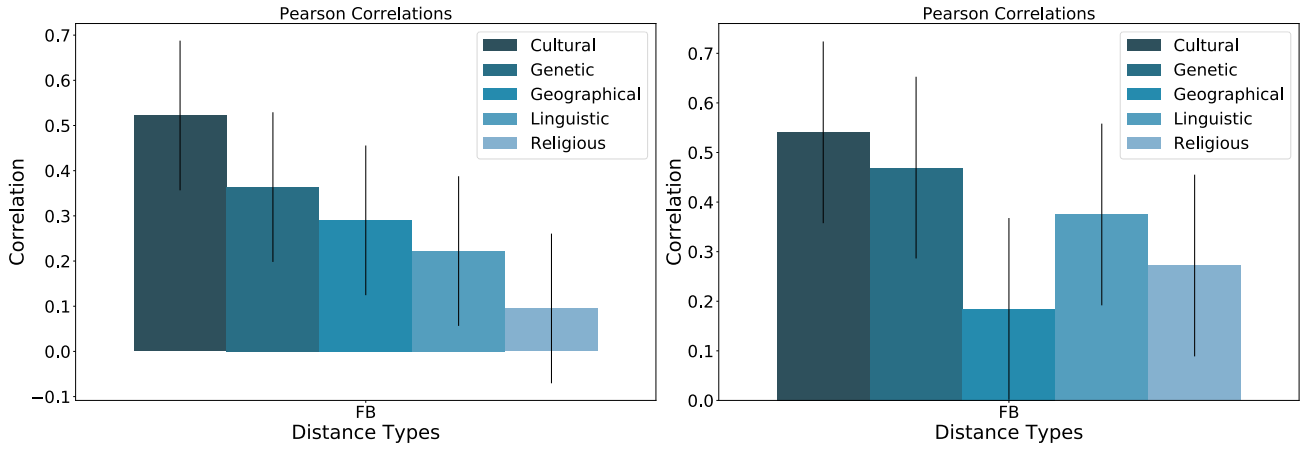


(d) Semi-Partial Correlation (Actual on Residuals)



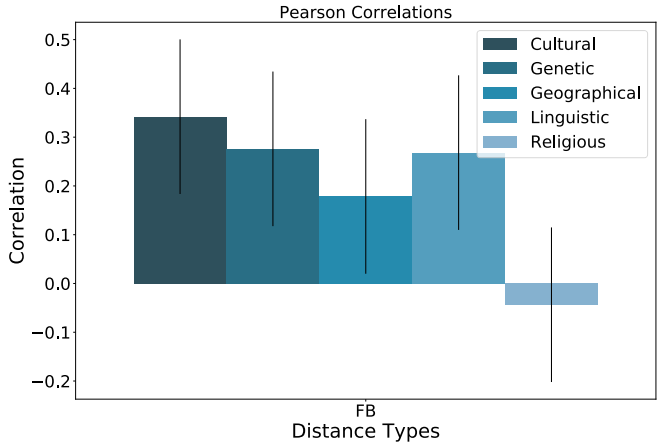
(e) Semi-Partial R^2

Figure B9: ***Correlations Between Facebook and Selected Distance Measures, Robustness to Representativity (Age Difference Above Median).*** *This figure shows the same information as Figure B1 and B2, with one difference: using the common sample, it only retains the countries where the difference in age composition of Facebook users and the actual population is above the median. Hence, it focuses on the subset of countries where Facebook users are least representative of the actual population in terms of age. For that sample, Panel (a) corresponds to Panel (a) of Figure B2, and Panels (b) through (e) correspond to Panels (a) through (d) of Figure B1.*

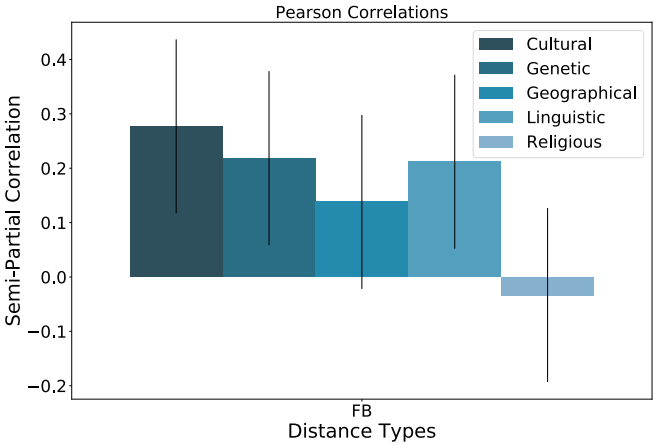


(a) Full Sample

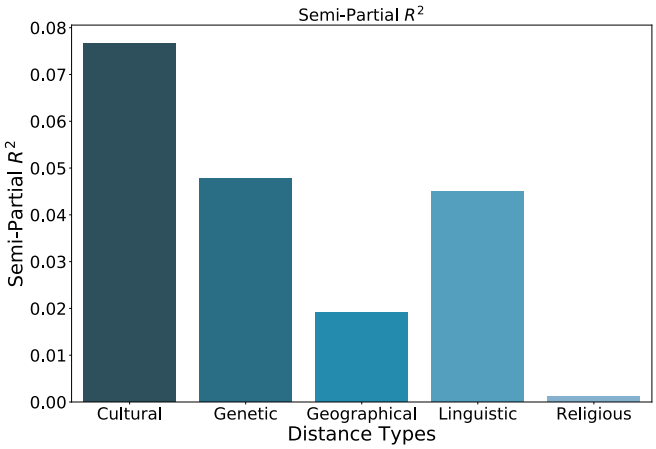
(b) Common Sample



(c) Partial Correlation (Residuals on Residuals)

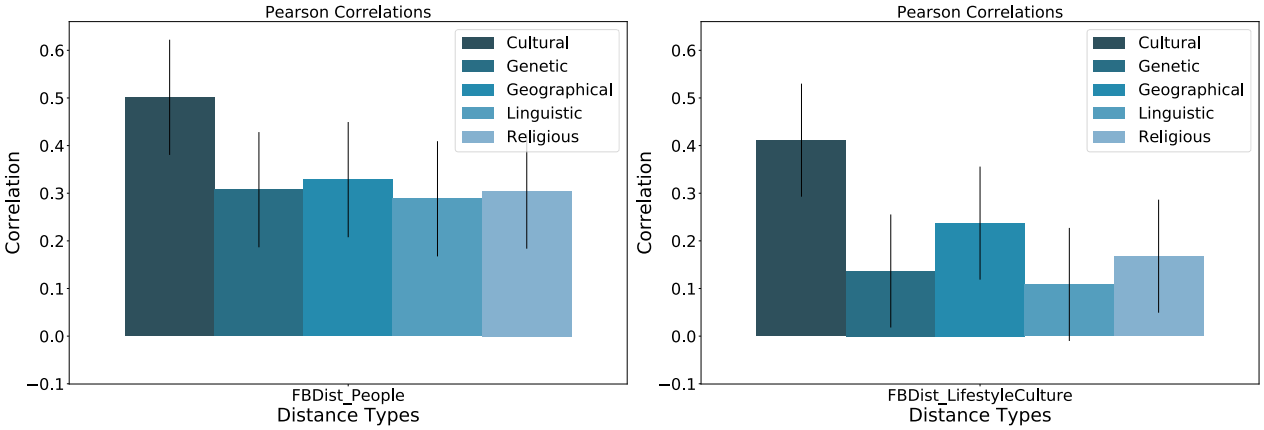


(d) Semi-Partial Correlation (Actual on Residuals)



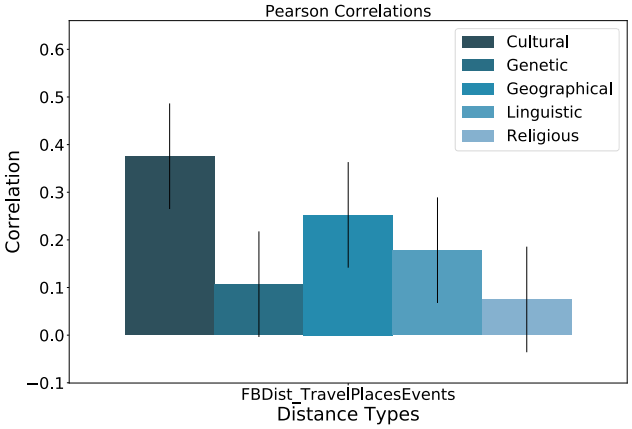
(e) Semi-Partial R^2

Figure B10: ***Correlations Between Facebook and Selected Distance Measures, Robustness to Representativity (Age Difference Below Median).*** *This figure shows the same information as Figure B1 and B2, with one difference: using the common sample, it only retains the countries where the difference in age composition of Facebook users and the actual population is below the median. Hence, it focuses on the subset of countries where Facebook users are most representative of the actual population in terms of age. For that sample, Panel (a) corresponds to Panel (a) of Figure B2, and Panels (b) through (e) correspond to Panels (a) through (d) of Figure B1.*

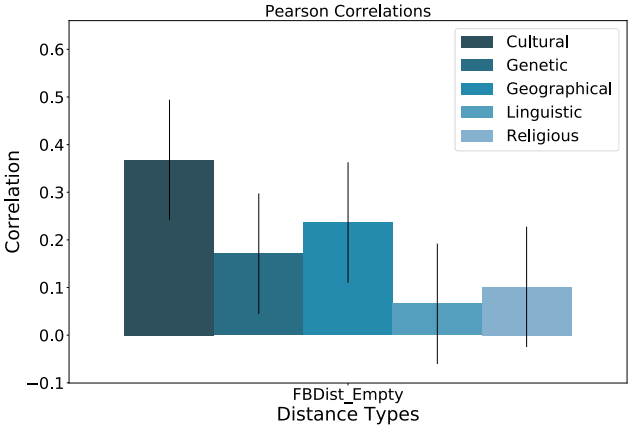


(a) People

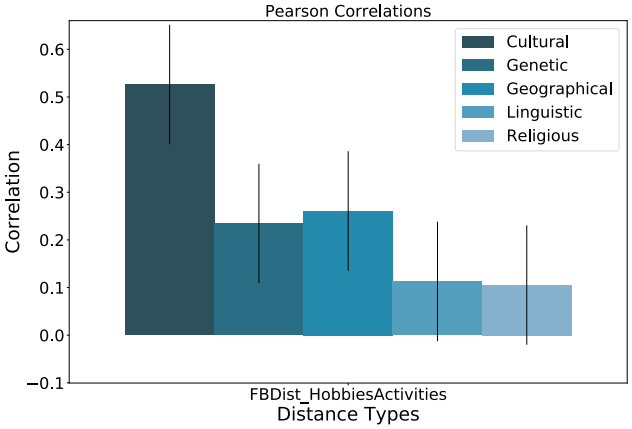
(b) Lifestyle & Culture



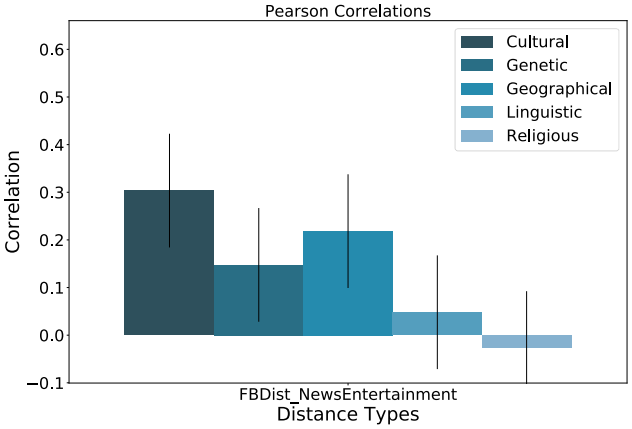
(c) Travel, Places & Events



(d) Empty

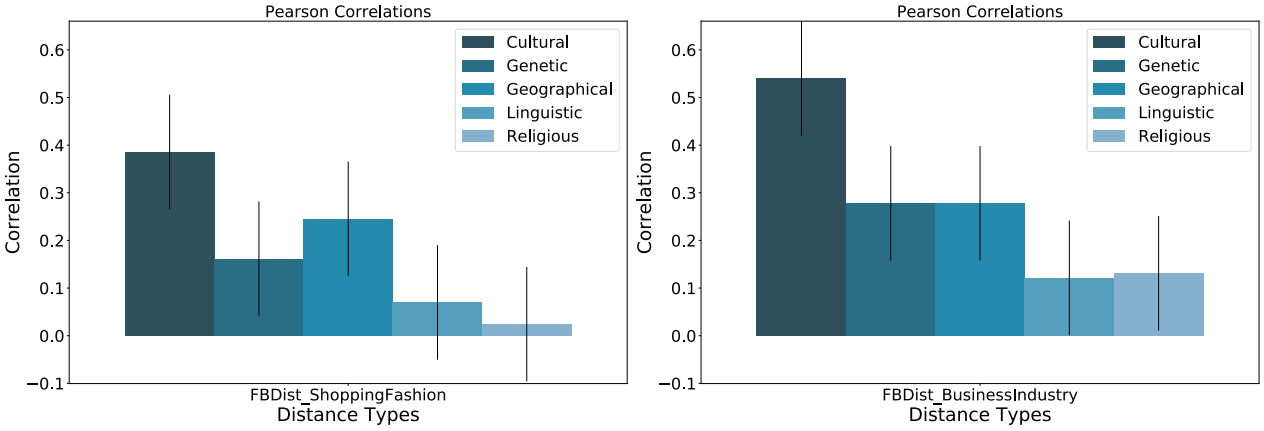


(e) Hobbies & Activities



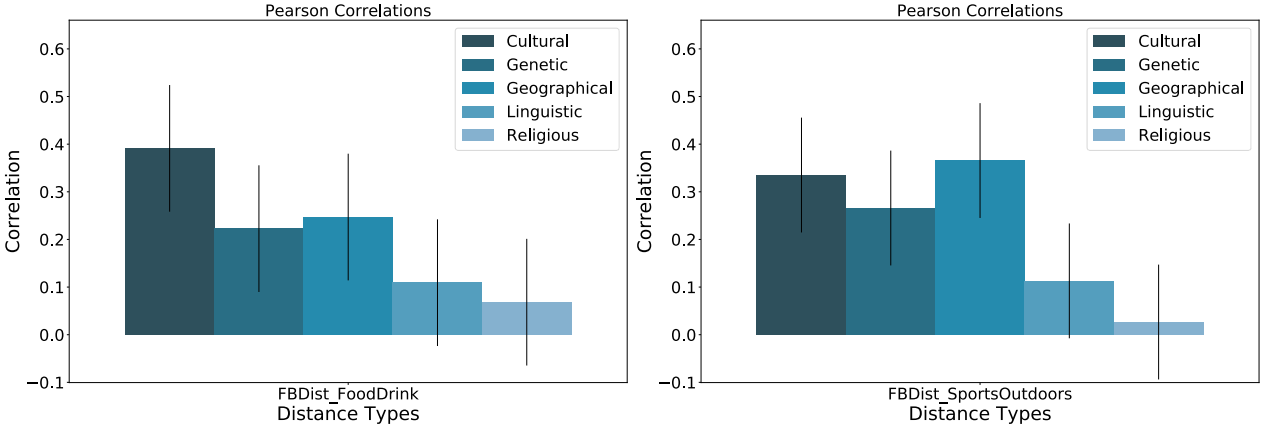
(f) News & Entertainment

Figure B11: ***Correlations Between Facebook and Selected Distance Measures, Robustness to Facebook Categories.*** *Each panel plots the correlations between Facebook distances and each one of the distance measures, not controlling for any other distance. Panels (a) through (n) depict these correlations for each one of the 14 macro-categories of interests: people; lifestyle and culture; travel, places and events; empty; hobbies and activities; news and entertainment; shopping and fashion; business and industry; food and drink; sports and outdoors; education; technology; fitness and wellness; and family and relationships. Panel (o) depicts these correlations for all interests that are not marked as local businesses.*



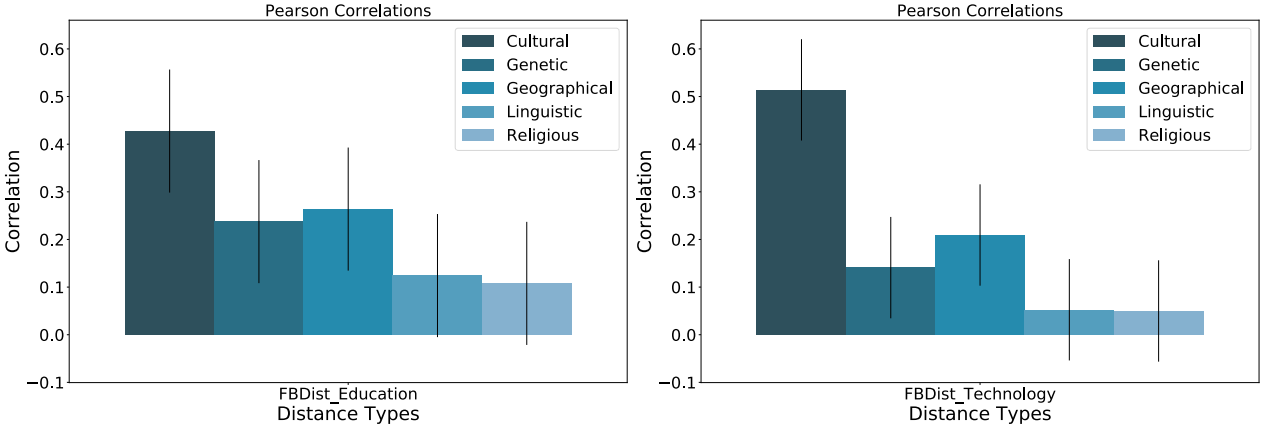
(g) Shopping & Fashion

(h) Business & Industry



(i) Food & Drink

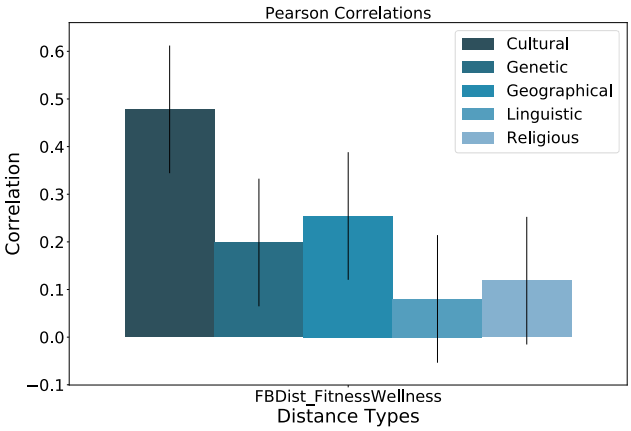
(j) Sports & Outdoors



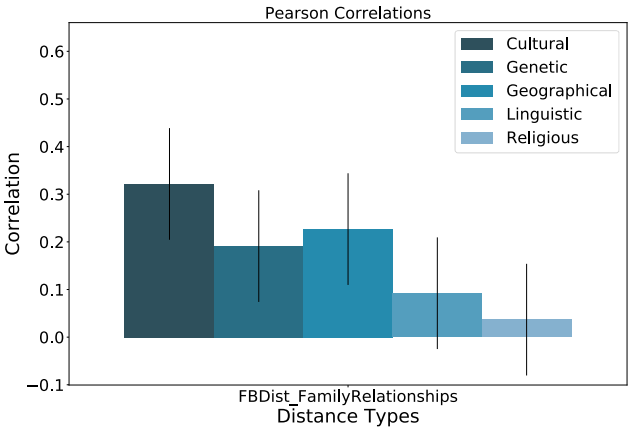
(k) Education

(l) Technology

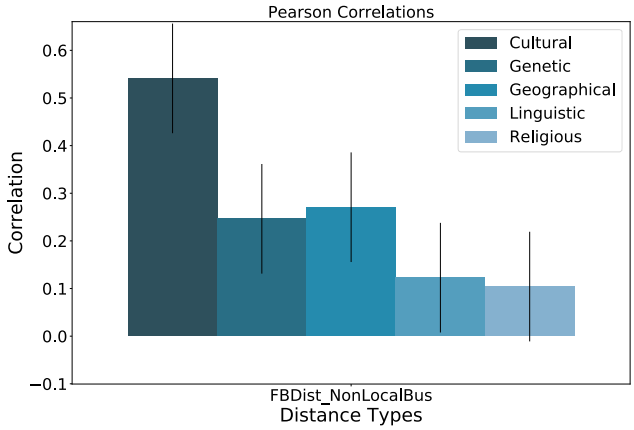
Figure B11: ***Correlations Between Facebook and Selected Distance Measures, Robustness to Facebook Categories (cont.).*** *Each panel plots the correlations between Facebook distances and each one of the distance measures, not controlling for any other distance. Panels (a) through (n) depict these correlations for each one of the 14 macro-categories of interests: people; lifestyle and culture; travel, places and events; empty; hobbies and activities; news and entertainment; shopping and fashion; business and industry; food and drink; sports and outdoors; education; technology; fitness and wellness; and family and relationships. Panel (o) depicts these correlations for all interests that are not marked as local businesses.*



(m) Fitness & Wellness



(n) Family & Relationships



(o) Non Local Business

Figure B11: ***Correlations Between Facebook and Selected Distance Measures, Robustness to Facebook Categories (cont.).*** *Each panel plots the correlations between Facebook distances and each one of the distance measures, not controlling for any other distance. Panels (a) through (n) depict these correlations for each one of the 14 macro-categories of interests: people; lifestyle and culture; travel, places and events; empty; hobbies and activities; news and entertainment; shopping and fashion; business and industry; food and drink; sports and outdoors; education; technology; fitness and wellness; and family and relationships. Panel (o) depicts these correlations for all interests that are not marked as local businesses.*

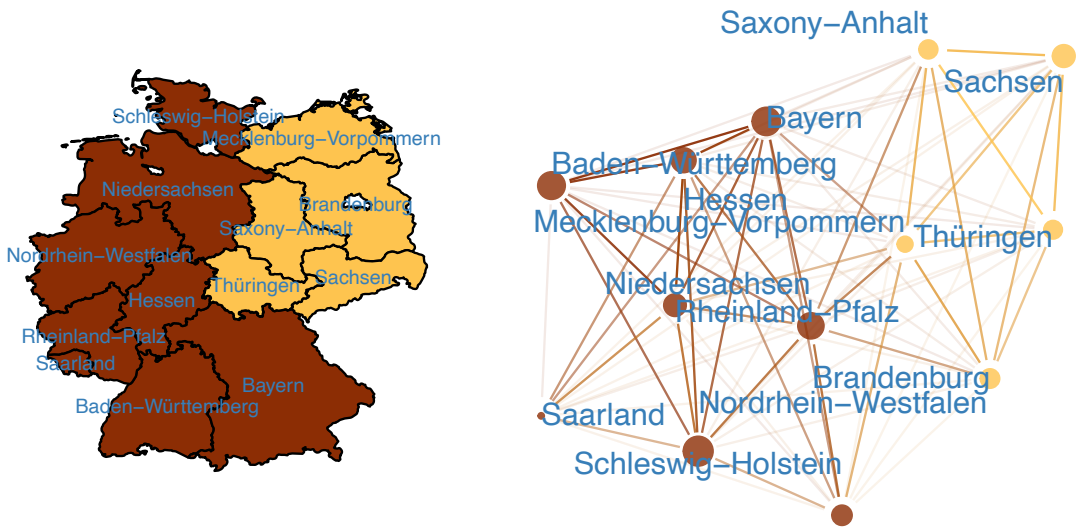
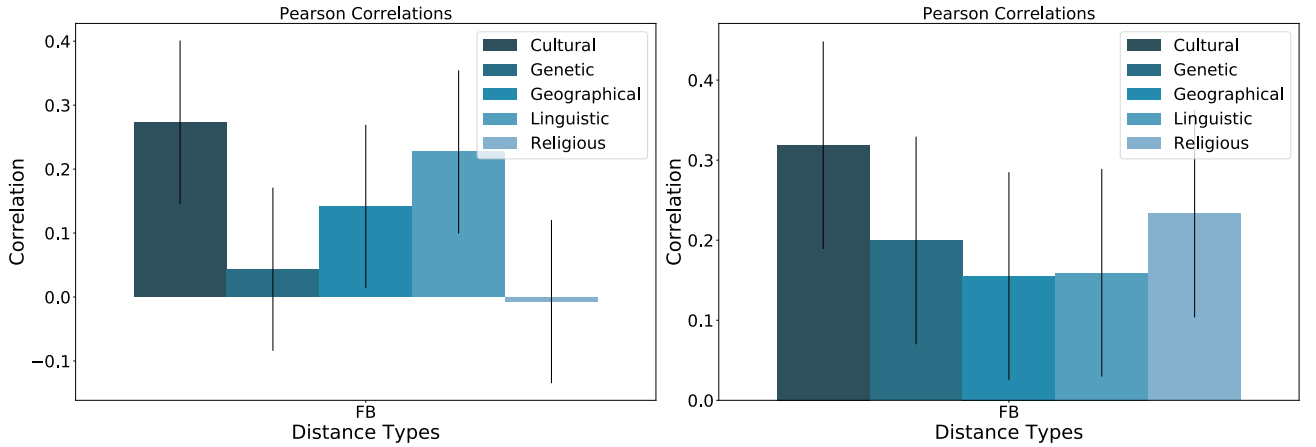
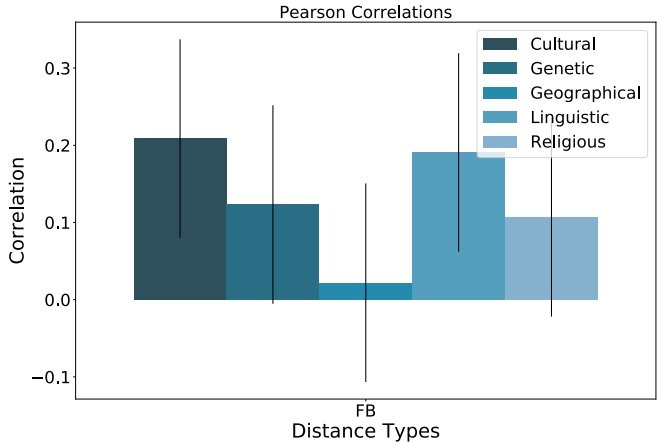


Figure B12: ***Geographical map and network of regions in Germany based on interests not marked as local businesses.*** *Networks are constructed from nodes as regions, and links are weighted by standardized cosine similarity between regions. Communities are detected using a multi-level modularity optimization algorithm (Louvain method). Nodes are resized proportionally to region population, and are colored according to community affiliation. Links are colored according to adjacent nodes, with lower transparency applied to higher weight links. Map regions are colored according to communities calculated from the corresponding network.*

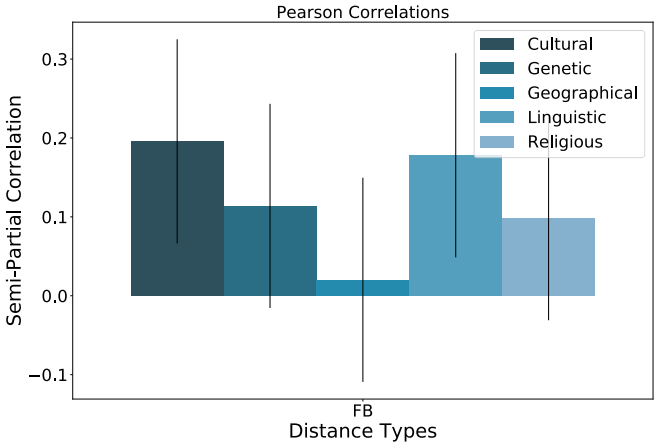


(a) Full Sample

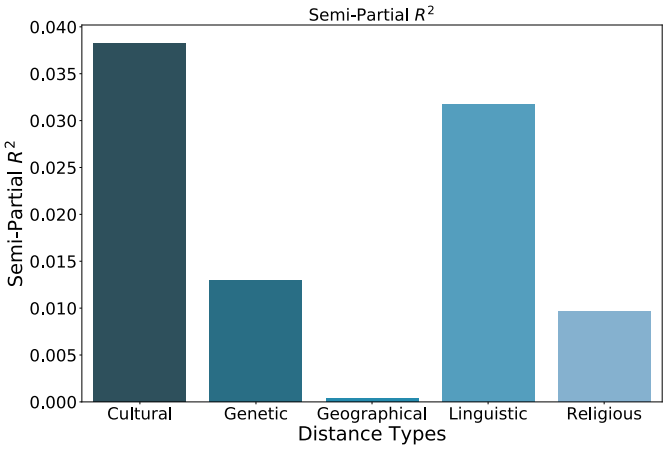
(b) Common Sample



(c) Partial Correlation (Residuals on Residuals)

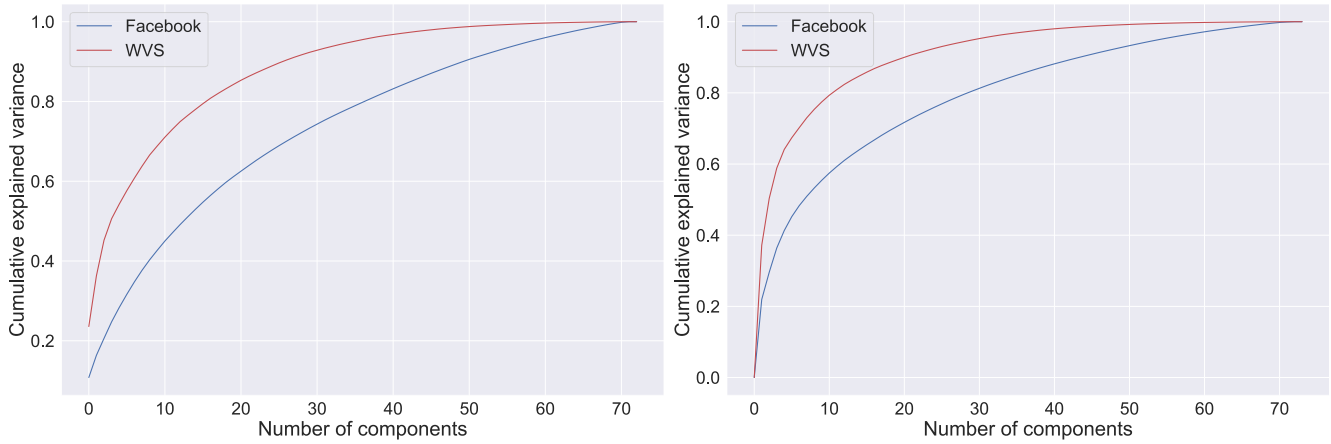


(d) Semi-Partial Correlation (Actual on Residuals)



(e) Semi-Partial R^2

Figure B13: ***Correlations Between Facebook and Selected Distance Measures, Euclidean Distances.*** *This figure shows the same information as Figure B1 and B2, with one difference: it uses Euclidean, rather than cosine distances. Using Euclidean distances, Panel (a) corresponds to Panel (a) of Figure B2, and Panels (b) through (e) correspond to Panels (a) through (d) of Figure B1.*



(a) Variance explained by PC (correlation)

(b) Variance explained by PC (covariance)

Figure B14: ***Variance explained by principal components.*** *This figure shows the number of independent underlying factors behind the variation in our Facebook measure and compares it to the traditional values-based measure. Specifically, it shows the cumulative share of total variation in our Facebook measure and the values-based measure explained by principal components. Panel (a) plots explained variance after standardizing, i.e. using the correlation matrix. Panel (b) plots explained variance without standardizing, i.e. using the covariance matrix.*

B.5 Cultural center of the world

In this subsection we explore which country is the cultural center of the world. Suppose the entire world population needs to meet in a particular country. We will refer to the country that minimizes the average Facebook distance traveled by the world population as the cultural center of the world. To determine this country, compute for each country k, the average distance traveled by the world population to meet in country k:

$$\cos\text{dist}(k, W) = \sum_{l} s_l \cos\text{dist}(l, k) \tag{4}$$

When using all 225 countries, Table B1 shows that the cultural center of the world is South Africa, probably the only country in the world with a large African, European and Asian population. Other countries in the top ten include some of the Gulf countries with large immigrant populations (44% in Oman, 37% in Saudi Arabia). If we limit ourselves to the sample of 161 countries with more than 300,000 people and a Facebook penetration above 5%, the cultural center of the world is India, although South Africa continues to be in the top-3. If instead of taking the actual population shares as the shares in (4) we consider the Facebook user shares, the cultural center of the world for the same sample of 161 countries becomes Switzerland. The cultural center based on Facebook users benefits either countries with large Facebook populations (such as the US) or countries that are culturally close to the bulk of world Facebook users (such as Switzerland or Belgium).

One advantage of using Facebook data to determine the cultural center of the world is its global coverage. For example, standard values surveys often undersample certain regions of the world. Needless to say, this introduces important distortions.

1. Top-10, weighted by actual populations, sample of 225 countries: South Africa, India, China, Oman, Russia, Benin, France, Tajikistan, Central African Republic, Saudi Arabia
2. Top-10, weighted by actual populations, sample of 161 countries: India, Oman, South Africa, Saudi Arabia, Germany, Qatar, France, Kyrgyzstan, Bahrain, Switzerland
3. Top-10, weighted by Facebook users, sample of 225 countries: South Africa, France, Switzerland, Canada, US, Belgium, Germany, Great Britain, Guyana, Panama
4. Top-10, weighted by Facebook users, sample of 161 countries: Switzerland, France, Germany, Belgium, Canada, South Africa, Great Britain, Panama, US, Guyana

Table B1: ***Cultural center of the world.*** *This table reports the top-10 countries that minimize the average distance the world population (or the world Facebook users) would need to travel if they were to meet in the same location. Calculations are based on (4), and differ depending on samples and weights. The sample can either be all 225 countries for which Facebook data are available (cases 1 and 3) or the subsample of 161 countries with a population of more than 300,000 and a Facebook penetration of more than 5% (cases 2 and 4). The distance can either be weighted by actual populations (case 1 and 2) or by Facebook users (cases 3 and 4).*

B.6 Cultural *sister* regions

In this subsection we identify for each subnational region of ten countries the "sister region", defined as the closest foreign region according to Facebook distances. Table B2 shows the results for the subnational regions of ten countries: United States (US), Spain (ES), France (FR), Germany (DE), Italy (IT), Portugal (PT), Great Britain (GB), Belgium (BE), Netherlands (NL) and Ireland (IE). For each subnational region, the cultural sister has to be in one of the other nine countries.

Several results stand out. First, the sister regions tend to be located in the countries that are closest in the dendrogram in Figure 3. For example, the sister regions of U.S. states are in Great Britain, the sister regions of France are in Belgium, and the sister regions of the German Länder are in the Netherlands. Second, subnational regions sometimes have sister regions in different countries, depending on linguistic or geographic proximity. For example, most regions in Spain have sister regions in Italy. However, Galicia, a region in northwest Spain that speaks a language closely related to Portuguese, has as sister region Lisbon. Likewise, the Dutch-speaking region of Belgium has as sister region Limburg in the southern Netherlands, whereas the French-speaking region of Belgium has as sister region Lorraine in France. As another example, England has as sister region the state of New York, whereas Scotland has as sister region County Louth on the eastern coast of Ireland. Italy also shows some interesting patterns: while most of its regions are paired to Corsica, a French island with an Italian dialect, the region of Trentino-Alto Adige with a large German-speaking population is paired to Bavaria. Third, the urban nature of regions also sometimes matters. For example, the Paris region of Ile-de-France has as sister region Brussels, the capital of Belgium. The same is true for the French region of Rhône-Alpes, home to the second-largest metropolitan area Lyon.

Region (US)	Sister	Region (ES)	Sister	Region (DE)	Sister	Region (PT)	Sister
US/Alabama	GB/England	ES/Balearic Islands	IT/Lombardy	DE/Baden-Wurttemberg	NL/North Holland	PT/Aveiro	IT/Friuli-Venezia
US/Alaska	GB/England	ES/La Rioja	IT/Lombardy	DE/Bayern	NL/North Holland	PT/Beja	IT/Friuli-Venezia
US/Arizona	GB/England	ES/Madrid	IT/Lombardy	DE/Hessen	NL/North Holland	PT/Braga	IT/Friuli-Venezia
US/Arkansas	GB/England	ES/Murcia	IT/Lombardy	DE/Niedersachsen	NL/Overijssel	PT/Braganca	IT/Friuli-Venezia
US/California	GB/England	ES/Navarre	IT/Lombardy	DE/Nordrhein-Westfalen	NL/Overijssel	PT/Castelo Branco	IT/Friuli-Venezia
US/Colorado	GB/England	ES/Asturias	PT/Lisbon D	DE/Rheinland-Pfalz	NL/Overijssel	PT/Coimbra	IT/Friuli-Venezia
US/Connecticut	GB/England	ES/Cantabria	IT/Lombardy	DE/Saarland	NL/Overijssel	PT/Evora	IT/Friuli-Venezia
US/Delaware	GB/England	ES/Andalusia	IT/Lombardy	DE/Schleswig-Holstein	NL/Overijssel	PT/Faro	IT/Friuli-Venezia
US/Florida	GB/England	ES/Aragon	IT/Lombardy	DE/Brandenburg	NL/Overijssel	PT/Madeira	IT/Friuli-Venezia
US/Georgia	GB/England	ES/Canary Islands	PT/Lisbon D	DE/Mecklenburg-Vorpom.	NL/Overijssel	PT/Guarda	IT/Friuli-Venezia
US/Hawaii	GB/England	ES/Castilla-La Mancha	IT/Lombardy	DE/Sachsen	NL/Overijssel	PT/Leiria	IT/Friuli-Venezia
US/Idaho	GB/England	ES/Castile and Leon	IT/Lombardy	DE/Saxony-Anhalt	NL/Overijssel	PT/Lisbon	IT/Friuli-Venezia
US/Illinois	GB/England	ES/Catalonia	IT/Lombardy	DE/Thuringen	NL/Overijssel	PT/Portalegre	IT/Friuli-Venezia
US/Indiana	GB/England	ES/Extremadura	IT/Lombardy	**Region (IT)**	**Sister**	PT/Porto	IT/Friuli-Venezia
US/Iowa	GB/England	ES/Galicia	PT/Lisbon D	IT/Abruzzo	FR/Corse	PT/Santarem	IT/Friuli-Venezia
US/Kansas	GB/England	ES/Basque Country	IT/Lombardy	IT/Basilicata	FR/Corse	PT/Setubal	IT/Friuli-Venezia
US/Kentucky	GB/England	ES/Valencia	PTLombardy	IT/Calabria	FR/Corse	PT/Viana do Castelo	IT/Friuli-Venezia
US/Louisiana	GB/England	**Region (FR)**	**Sister**	IT/Campania	FR/Corse	PT/Vila Real	IT/Friuli-Venezia
US/Maine	GB/England	FR/Aquitaine	BE/Wallonia	IT/Emilia-Romagna	FR/Corse	PT/Viseu	IT/Friuli-Venezia
US/Maryland	GB/England	FR/Auvergne	BE/Wallonia	IT/Friuli-Venezia Giulia	FR/Corse	PT/Azores	IT/Friuli-Venezia
US/Massachusetts	GB/England	FR/Basse-Normandie	BE/Wallonia	IT/Lazio	FR/Corse	**Region (NL)**	**Sister**
US/Michigan	GB/England	FR/Bourgogne	BE/Wallonia	IT/Liguria	FR/Corse	NL/Drenthe	BE/Flemish Region
US/Minnesota	GB/England	FR/Bretagne	BE/Wallonia	IT/Lombardy	FR/Corse	NL/Friesland	BE/Flemish Region
US/Mississippi	GB/England	FR/Centre	BE/Wallonia	IT/Marche	FR/Corse	NL/Gelderland	BE/Flemish Region
US/Missouri	GB/England	FR/Champagne-Ardenne	BE/Wallonia	IT/Molise	FR/Corse	NL/Groningen	BE/Flemish Region
US/Montana	GB/England	FR/Corse	BE/Wallonia	IT/Piedmont	FR/Corse	NL/Limburg	BE/Flemish Region
US/Nebraska	GB/England	FR/Franche-Comte	BE/Wallonia	IT/Puglia	FR/Corse	NL/North Brabant	BE/Flemish Region
US/Nevada	GB/England	FR/Haute-Normandie	BE/Wallonia	IT/Sardinia	PT/Lisbon D	NL/North Holland	BE/Flemish Region
US/New Hampshire	GB/England	FR/Ile-deFR	BE/Brussels	IT/Sicilia	FR/Corse	NL/Utrecht	BE/Flemish Region
US/New Jersey	GB/England	FR/Languedoc-Roussillon	BE/Wallonia	IT/Tuscany	FR/Corse	NL/Zeeland	BE/Flemish Region
US/New Mexico	GB/England	FR/Limousin	BE/Wallonia	IT/Trentino-Alto Adige	DE/Bayern	NL/Zuid-Holland	BE/Flemish Region
US/New York	GB/England	FR/Lorraine	BE/Wallonia	IT/Umbria	FR/Corse	NL/Overijssel	BE/Flemish Region
US/North Carolina	GB/England	FR/Midi-Pyrenees	BE/Wallonia	IT/Aosta Valley	FR/Corse	NL/Flevoland	BE/Flemish Region
US/North Dakota	GB/England	FR/Nord-Pas-de-Calais	BE/Wallonia	IT/Veneto	FR/Corse	**Region (BE)**	**Sister**
US/Ohio	GB/England	FR/Pays de la Loire	BE/Wallonia	**Region (GB)**	**Sister**	BE/Brussels	FR/Ile-deFR
US/Oklahoma	GB/England	FR/Picardie	BE/Wallonia	GB/England	US/New York	BE/Flemish Region	NL/Limburg
US/Oregon	GB/England	FR/Poitou-Charentes	BE/Wallonia	GB/Wales	IE/County Meath	BE/Wallonia	FR/Lorraine
US/Pennsylvania	GB/England	FR/Provence-Alpes-C. Azur	BE/Wallonia	GB/Scotland	IE/County Louth		
US/Rhode Island	GB/England	FR/Rhone-Alpes	BE/Brussels	GB/Northern Ireland	IE/County Louth		
US/South Carolina	GB/England						
US/South Dakota	GB/England						
US/Tennessee	GB/England						
US/Texas	GB/England						
US/Utah	GB/England						
US/Vermont	GB/England						
US/Virginia	GB/England						
US/Washington	GB/England						
US/West Virginia	GB/England						
US/Wisconsin	GB/England						
US/Wyoming	GB/England						

Table B2: ***"Sister region" for each region in a sample of ten countries****: the United States (US), Spain (ES), France (FR), Germany (DE), Italy (IT), Portugal (PT), Great Britain (GB), Belgium (BE), Netherlands (NL) and Ireland (IE). A sister region is defined as the foreign region in one of the other nine countries that is closest according to Facebook distances.*

B.7 Regional divisiveness and gender divisiveness

Given their importance, it is interesting to compare regional differences and gender differences. Broadly speaking, our findings suggest that in most developed countries the gender divide is larger than the regional divide, whereas in many developing countries the regional divide continues to be important (Figure 7).

Rather than comparing the *average* interregional distance with the gender distance, we could also compare *all* bilateral distances between regions in a given country to the bilateral distance between genders in that same country. Figure B15 displays for each country a kernel density plot of the bilateral distances between regions as well as the bilateral distance between genders. Two observations stand out. First, interregional distances vary widely, both in their variance across countries and in some cases in their variance within countries. Second, in comparison, gender differences are relatively similar across countries.

In some countries, such as France and Germany, the distance between genders is greater than the bilateral distance between *any* two regions. For many other countries, the picture is more complex. Take, for instance, the United States. There, the gender difference is greater than many, but not all, interstate differences. As a comparison: while the average distance between Texas and California residents is larger than the average distance between men and women in the U.S., the opposite is true for the average distance between Massachusetts and Connecticut residents or between New Mexico and Colorado residents. As another example, consider Kenya. There, the average gender divide is similar to the average regional divide. However, the regional divide between the North Eastern Province, inhabited by Somalis, and the rest of the country is huge. As a last example, take India. In that country, the regional differences are on average much larger than the gender difference, but there are exceptions, such as the distance between Uttar Pradesh and Madhya Pradesh, two neighboring northern states.

B.8 Specific cultural traits

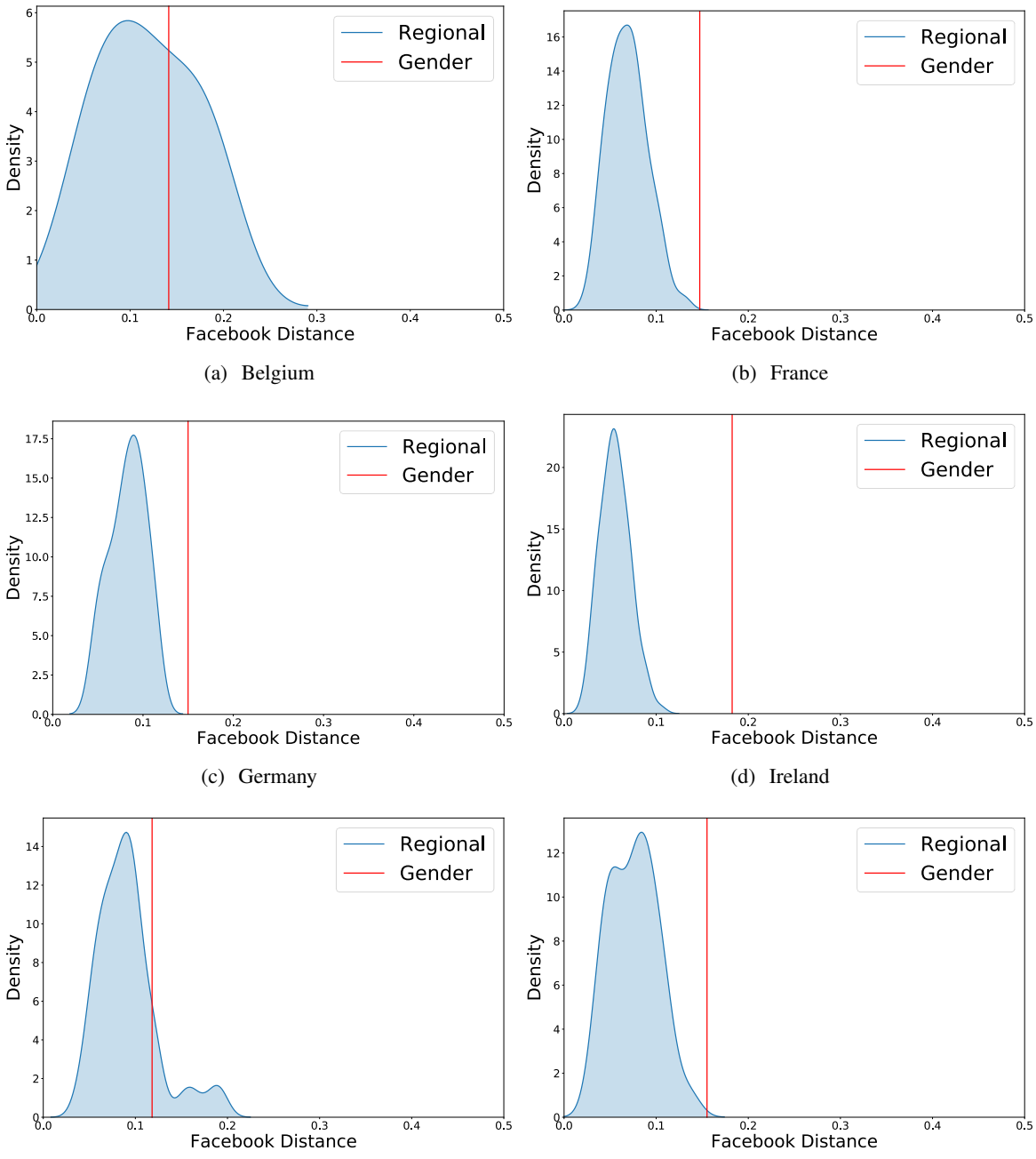
While our results suggest that the broad set of Facebook interests are able to provide an overall measure of culture, this section explores whether they can also capture *specific* cultural traits of interest to social scientists. To that end, we use a supervised machine learning algorithm that uses Facebook interests to predict close to 50 specific cultural traits or attributes, ranging from generosity to kinship tightness, from uncertainty avoidance to son bias, and from beef consumption to contraceptive use. The second column of Table B3 then reports the correlations between the predicted cultural attributes and the observed cultural attributes. We find an average correlation of 0.59, suggesting that the broad set of Facebook interests is able to capture specific, more traditional cultural traits, providing further validation of our measure.

To describe our approach in more detail, suppose we want to predict a country's degree of generosity. As inputs, we use the matrix of all Facebook interests and a vector of the degree of generosity from another data source (in this case, from the World Happiness Report). Using a training sample that consists of 90% of the countries, the Mathematica command "predict" then chooses among a set of standard supervised machine learning algorithms and predicts the degree of generosity in the test sample that consists of the 10% remaining countries. By running this algorithm 100 times on varying training samples,we obtain values for the average predicted generosity by country. We then correlate the generosity as predicted by Facebook to the generosity as observed in the data.

Alternatively, instead of using all Facebook interests to predict cultural traits, we can also use the principal components of the Facebook interests. The third column in Table B3 reports the correlations between predicted and observed cultural traits, based on the first 20 principal components. The correlations are slightly higher, at an average of 0.64.

B.9 Dendrogram with full sample of countries

Figure B16 depicts the dendrogram from the main text, employing the full sample of Facebook countries.

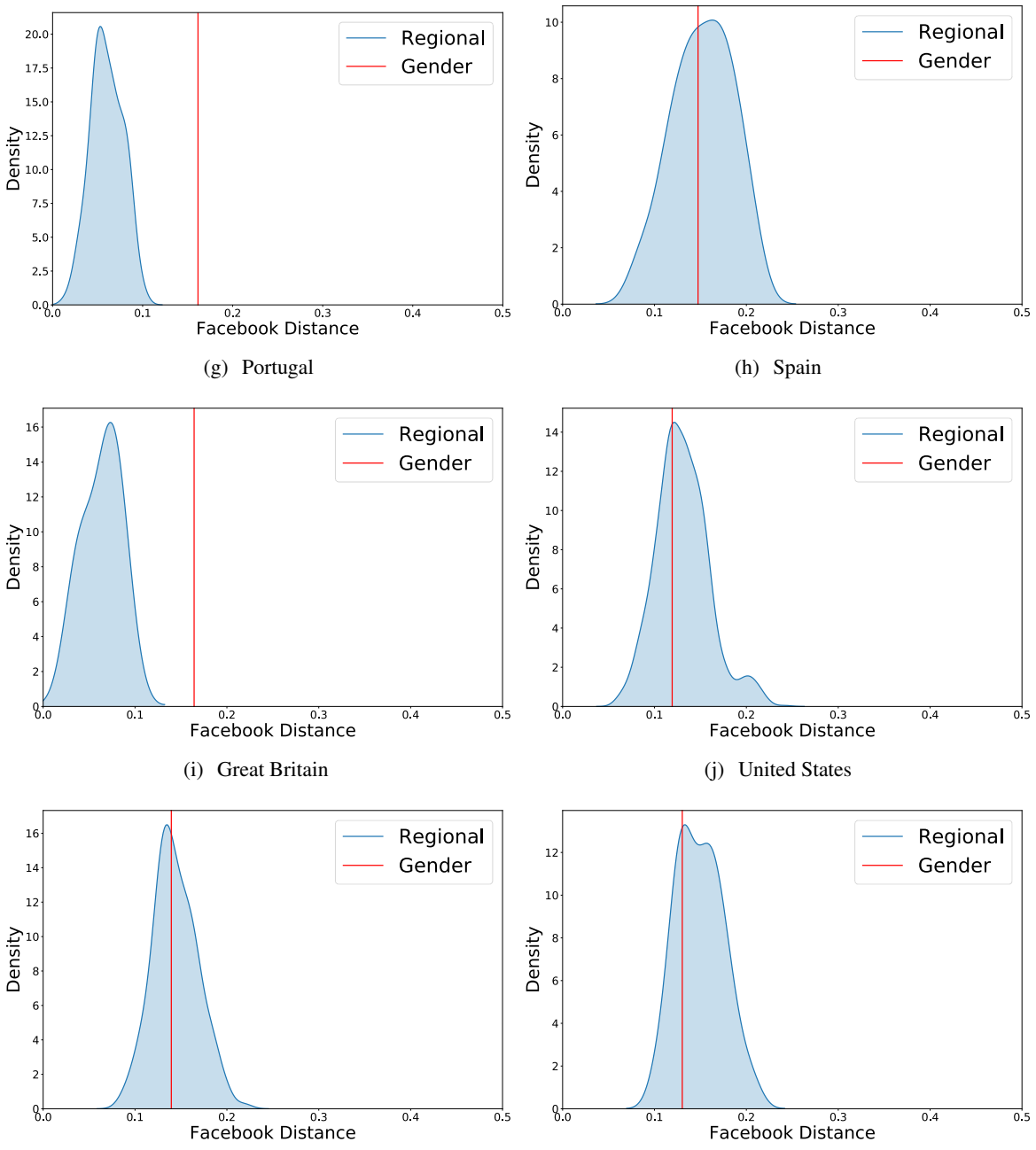


(a) Belgium (b) France

(c) Germany (d) Ireland

(e) Italy (f) Netherlands

Figure B15: ***Kernel density plots of interregional bilateral distances (blue) and gender distances (red).*** *This figure displays histograms of bilateral distances between regions and bilateral distances between genders for 18 countries (BE: Belgium, BR: Brazil, DE: Germany, ES: Spain, FR: France, GB: Great Britain, IE: IN: India, Ireland, IT: Italy, KE: Kenya, MU: Mauritius, MX: Mexico, NG: Nigeria, NL: Netherlands, PT: Portugal, SV: El Salvador, TH: Thailand, US: United States). All figures have the same horizontal scale.*

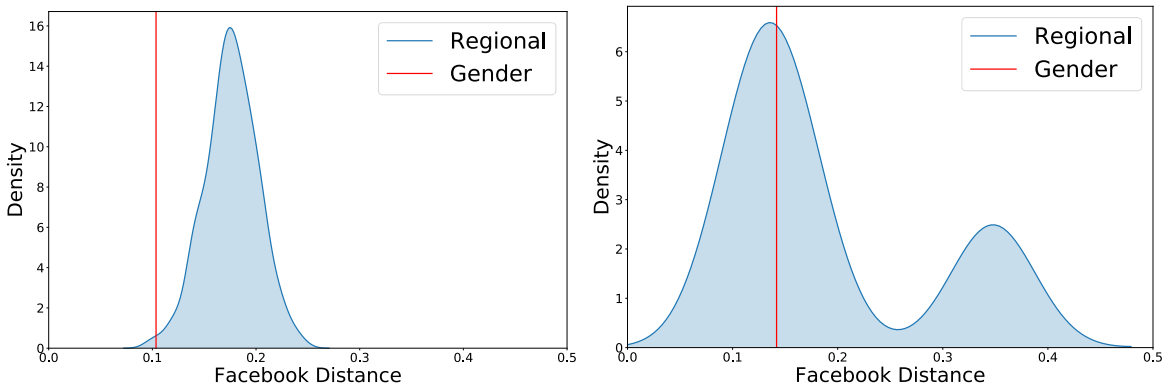


(g) Portugal (h) Spain

(i) Great Britain (j) United States

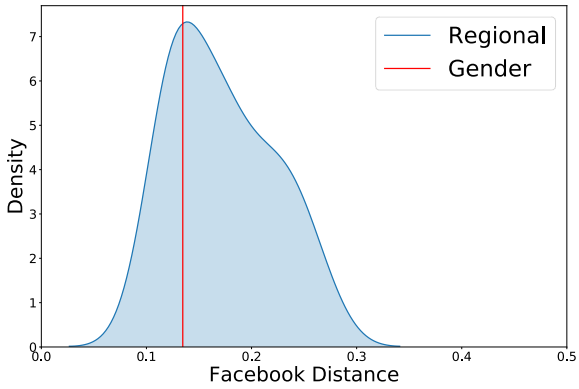
(k) Brazil (l) El Salvador

Figure B15: ***Kernel density plots of interregional bilateral distances (blue) and gender distances (red) (continued).*** *This figure displays histograms of bilateral distances between regions and bilateral distances between genders for 18 countries (BE: Belgium, BR: Brazil, DE: Germany, ES: Spain, FR: France, GB: Great Britain, IE: IN: India, Ireland, IT: Italy, KE: Kenya, MU: Mauritius, MX: Mexico, NG: Nigeria, NL: Netherlands, PT: Portugal, SV: El Salvador, TH: Thailand, US: United States). All figures have the same horizontal scale.*

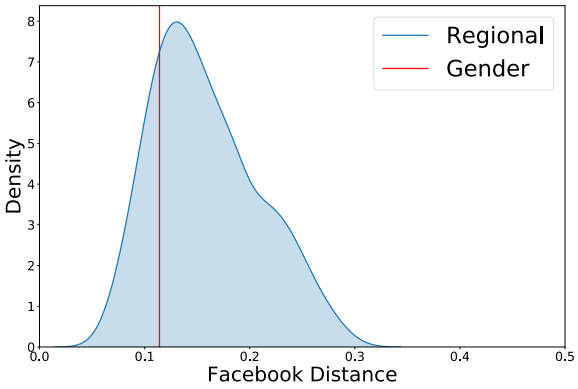


(m) Mexico

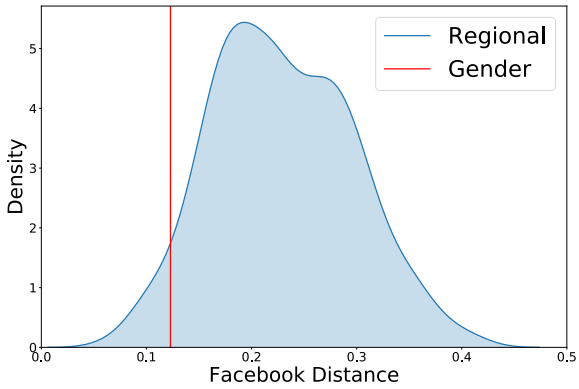
(n) Kenya



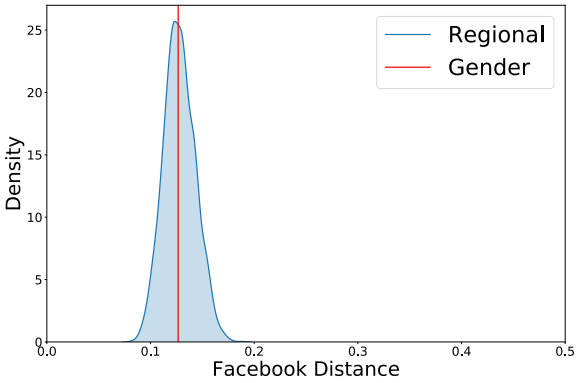
(o) Mauritius



(p) Nigeria



(q) India



(r) Thailand

Figure B15: ***Kernel density plots of interregional bilateral distances (blue) and gender distances (red) (continued).*** *This figure displays histograms of bilateral distances between regions and bilateral distances between genders for 18 countries (BE: Belgium, BR: Brazil, DE: Germany, ES: Spain, FR: France, GB: Great Britain, IE: IN: India, Ireland, IT: Italy, KE: Kenya, MU: Mauritius, MX: Mexico, NG: Nigeria, NL: Netherlands, PT: Portugal, SV: El Salvador, TH: Thailand, US: United States). All figures have the same horizontal scale.*

Cultural Dimension / Sub-Dimension	FB Interests	20 PC	Obs	Source
Demographic				
Fertility rate	0.85	0.88	156	Fertility rate 2017, WDI, World Bank[18]
Age first marriage, men	0.66	0.72	131	Age first marriage men 2005-2014 (United Nations)[19]
Age first marriage, women	0.73	0.79	147	Age first marriage women 2005-2014 (United Nations)[19]
Food and Drinks				
Alcohol, both sexes	0.77	0.83	150	Alcohol per capita, both sexes 2010-2016 (WHO)[20]
Alcohol, women	0.79	0.85	150	Alcohol per capita, female 2010-2016 (WHO)[20]
Alcohol, men	0.78	0.83	150	Alcohol per capita, male 2010-2016 (WHO)[20]
Beef consumption	0.53	0.63	142	Beef and buffalo consumption kg per capita 2013 (FAO)[21]
Pork consumption	0.84	0.88	138	Pig consumption kg per capita 2013 (FAO)[21]
Poultry consumption	0.62	0.72	142	Poultry consumption kg per capita 2013 (FAO)[21]
Mutton and goat consumption	0.52	0.55	142	Mutton and goat consumption kg per capita 2013 (FAO)[21]
Other meat consumption	0.38	0.49	142	Other meat consumption kg per capita 2013 (FAO)[21]
Gender issues				
Contraceptive use, women	0.78	0.78	118	Contraceptive use, women ages 15-49, 2010-2019, UNICEF[18]
Discriminatory work norms against women	0.78	0.81	148	Males 15+ who agree it is unacceptable for women to work, 2016, Georgetown GIWPS[22]
Gender inequality index, UN	0.80	0.93	141	Gender inequality index UN 2018[23]
Female employment	0.67	0.62	149	Employment women ages 25+, 2018, Georgetown GIWPS[22]
Gender gap	0.72	0.73	134	World Economic Forum Global Gender Gap, 2017-2020[24]
Community safety, women	0.57	0.60	148	Perception community safety among women ages 15+, 2010-16, Georgetown GIWPS[22]
Intimate partner violence	0.66	0.66	147	Intimate partner violence experienced by women, 2000-2017, Georgetown GIWPS[22]
Son bias	0.58	0.61	149	Male to female ratio at birth, 2015-20, Georgetown GIWPS[22]
Women in parliament	0.45	0.48	148	Parliamentary seats held by women, 2019, Georgetown GIWPS[22]
Wellbeing				
Subjective well-being	0.76	0.84	126	Life ladder 2018 (World Happiness Report)[25]
Healthy life expectancy	0.83	0.90	123	Healthy life expectancy at birth 2018 (World Happiness Report)[25]
Insufficient physical exercise	0.44	0.52	128	Insufficient physical activity adults (%), 2016 (WHO)[20]
Freedom to make life choices	0.44	0.55	126	Freedom to make life choices 2018 (World Happiness Report)[25]
Generosity	0.42	0.40	117	Generosity 2018 (World Happiness Report)[25]
Social support	0.75	0.78	126	Social support 2018 (World Happiness Report)[25]
Suicide rate	0.56	0.61	148	Crude suicide rates (per 100 000 population) 2016, WHO[26]
Society				
Social capital	0.75	0.77	129	Social capital 2018-2019 (World Economic Forum)[27]
Relational mobility	0.70	0.65	39	Relational mobility Thomson et al PNAS[28]
Nepotism in business	0.68	0.85	106	Nepotism in business (Enke, based on Van de Vliert)[29]
Kinship tightness	0.72	0.73	149	Kinship tightness score (Enke)[29]
Nuclear family	0.56	0.53	147	Nuclear family (Enke, based on Ethnographic Atlas and Giuliano)[29]
Perception of corruption	0.62	0.72	121	Perception of corruption 2018 (World Happiness Report)[25]
Ethnolinguistic fractionalization	0.48	0.44	153	ELF6 (Desmet, Ortuño and Wacziarg)[11]
Global Preferences Survey				
Altruism	-0.15	0.12	72	Altruism (Global Preferences Survey)[30,31]
Trust	0.41	0.43	72	Trust (Global Preferences Survey)[30,31]
Negative reciprocity	0.04	0.06	72	Negative reciprocity Global Preferences Survey)[30,31]
Positive reciprocity	0.08	0.09	72	Positive reciprocity (Global Preferences Survey)[30,31]
Risk taking	0.33	0.37	72	Willingness to take risks (Global Preferences Survey)[30,31]
Patience	0.63	0.71	72	Patience (Global Preferences Survey)[30,31]
Hofstede values				
Indulgence vs restraint	0.68	0.66	88	Indulgence vs restraint (Hofstede)[32]
Uncertainty avoidance	0.56	0.65	69	Uncertainty avoidance index (Hofstede)[32]
Long-term orientation	0.77	0.82	89	Long-term orientation vs short-term orientation (Hofstede)[32]
Masculinity vs femininity	0.44	0.50	69	Masculinity vs femininity (Hofstede)[32]
Individualism vs collectivism	0.82	0.81	69	Individualism vs collectivism (Hofstede)[32]
Power distance	0.48	0.66	69	Power distance index (Hofstede)[32]

Table B3: ***Correlations between observed cultural traits and those same cultural traits as predicted by FB interests using a machine learning algorithm***. *Column 2 uses all FB interests in predicting, Column 3 uses the first 20 principal components of FB interests in predicting, and Column 4 gives the number of observations.*

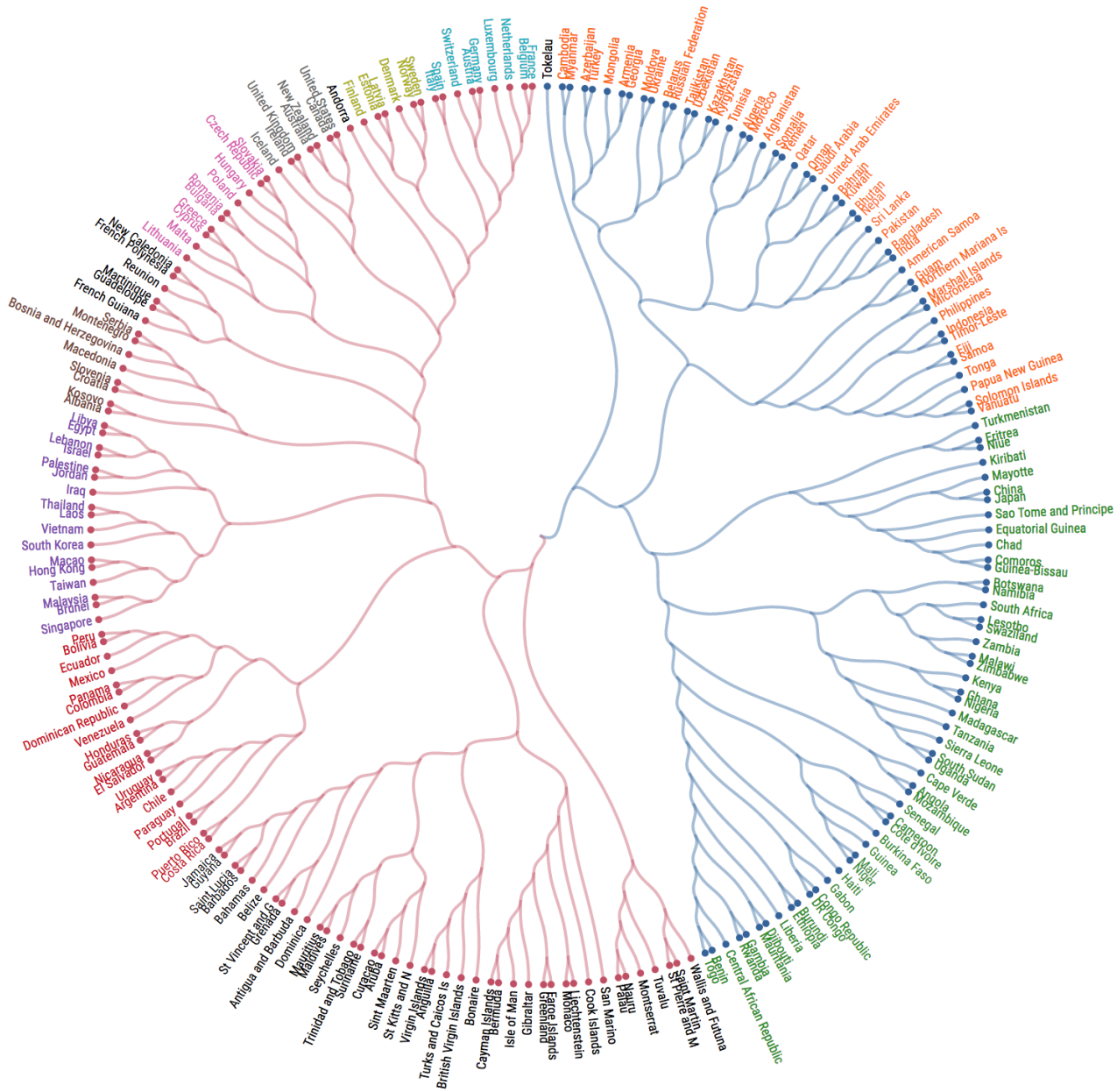


Figure B16: ***Hierarchical clustering of all in-sample countries based on Facebook distances.*** *Dendrogram is generated using the cosine distance and Ward linkage method. All countries in the Facebook data are included. The color of a country's link represents its membership to a main cluster, while the color of its name represents its membership to a sub-cluster. Two countries of the same name color (respectively link color) are closer to each other than to a country of a different name (respectively link) color.*

B.10 Examples of FB interests associated with different traditional and non-traditional cultural traits

Cultural trait	Category	Examples of FB Interests
Arts	**Traditional**	Abstract Art, African Art, Art Deco, Art museum, Art history, Art rock, Body art, Byzantine Art, Ceramic Art, Conceptual art, Contemporary Art Gallery, Cooking art, Cover art, Digital art, Fine-art photography, Folk art, Glass art, Gothic art, History of art, Interactive art, Japanese art, Louvre Art Museum, Make Up Art, Medieval art, Metropolitan Museum of Art, Mexican art, Modern art, Museum of Contemporary Art, Museum of Modern Art, Nail Art, National Gallery of Art, New media art, Performance art, Pixel art, Pop art, Public art, Red Ted Art, Sound art, Tattoos and Tattoo art, Thai Temple art and architecture, Van Gogh Museum Art Museum. Video art, Visionary art, Wearable art, Wood art, Latte art**.**
Formalities	**Traditional**	Arranged Marriage, Civil Marriage, Marriage (Catholic Church), Marriage in Islam, Marriage license, Marriage vows, Civil procedure, Criminal procedure, Ritual, Ritual purification, Rite, Rite Aid, Rite of passage, Prenuptial Agreement, Vehicle Registration Plate.
Religion	**Traditional**	Religion, Afro-american Religion, Confession (religion), Glory (religion), Transcendence (religion), Yoruba religion, Islam, Abraham in Islam, David in Islam, Five Pillars of Islam, God in Islam, Jesus in Islam, Intimate parts (Islam), Islam is Beautiful, Islam in the United States, Islam in India, Islam in Malaysia, Marriage in Islam, Shia Islam, Shirk (Islam), Studying Islam, Sunni Islam, Women in Islam, Born again (Christianity), Christianity, Disciple (Christianity), Early Christianity, Eastern Christianity, Eternal life (Christianity), God in Christianity, Grace (Christianity), Heaven (Christianity), Holy Spirit (Christianity), Minister (Christianity), Mission (Christianity), Passion (Christianity), Western Christianity, Judaism, Jewish culture, Jewish prayer, Conservative Judaism, Messianic Judaism, Orthodox Judaism, Reform Judaism, Christian prayer, Mass (liturgy).
Politics	**Traditional**	Politics, Centre-right politics, Far-left politics, Far-right politics, Gun politics in the United States, Green politics, Law and order (politics), Left-wing politics, Opposition (politics), Politics and Social Issues, Right-wing politics, Speaker (politics), Whip (politics), Brazilian Republican Party, Idaho Republican Party, Republican National Convention, Republican National Committee, Republican Left of Catalonia, Republican Party of Florida, Republican Party of Texas, Republican People's Party (Turkey), Republican Party (United States), Democrat Party (Thailand), Brazilian Democratic Movement Party, Christian Democratic Party (Chile), Christian Democratic People's Party of Switzerland, Christian Democratic Union (Germany), Idaho Democratic Party, Liberal Democratic Party (Japan), Liberal Democracy, Democratic Party (United States), Democratic Socialism, Socialism, Communism, Brazilian Socialist Party, Italian Socialist Party, Spanish Socialist Worker's Party, Barack Obama, Donald Trump, Bill Clinton, Hillary Clinton.
Social Structures	**Traditional**	Family, Marriage, Town hall, Municipality, State, Government, Executive (government), Church service, Social Security, Police, Army, Non-Governmental Organisation.
Angry Birds	**Non-Traditional**	Angry Birds, Angry Birds (video game), Angry Birds 2 Angry Birds Epic, Angry Birds POP! -Bubble Shooter, angry birds friends
Soccer fans	**Non-Traditional**	Real Madrid C.F, History of Real Madrid C.F., Real Madrid Fans, Real Madrid Fans Club, FC Barcelona, FC Barcelona Fans Club, F.C. United of Manchester, Manchester City F.C, Manchester City F.C. Supporters.
Loungewear	**Non-Traditional**	Tracksuit, Shorts, Bermuda shorts, Boxer shorts, T-shirt, Sports bra.
Running	**Non-Traditional**	Running, Barefoot running, Cross country running, Long-distance running, Middle-distance running, Mizuno running North America, Road running, Sprint (running), Stadion (running race), Trail running, Women's Running Magazine, Adidas running, Brooks running, Ironman Triathlon, Half marathon, Marathon

Table B4: ***Table mapping FB interests to cultural traits both used in traditional measures of culture and non-traditional ones***. *This table illustrates actual FB interests included in the interest set used in the paper that can be mapped into specific cultural traits. We have selected nine cultural traits in this table divided into two groups. First, cultural traits that are usually employed in traditional measures of culture (arts, formalities, religion, politics and social structures). Second, traits that our measure of culture also captures but are not used in traditional measures of culture (angry birds, football fans, loungewear and running).*

Appendix References

1. Englehardt, S. & Narayanan, A. Online Tracking. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS'16* (2016) doi:10.1145/2976749.2978313.

2. Cabañas, J. G., Cuevas, Á. & Cuevas, R. Unveiling and Quantifying Facebook Exploitation of Sensitive Personal Data for Advertising Purposes. (2018).

3. `https://transparency.facebook.com/community-standards-enforcement#fake-accounts`, consulted May 5, 2021.

4. Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A. & Rieke, A. Discrimination through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. *Proc. ACM Hum.-Comput. Interact.* **3**, CSCW, Article 199 (2019).

5. Legendre, P. & Legendre, L. *Numerical ecology*. vol. 24 (Elsevier, 2012).

6. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825-2830 (2011).

7. Spolaore, E. & Wacziarg, R. Ancestry, language and culture. in *The Palgrave handbook of economics and language* 174-211 (Springer, 2016).

8. Spolaore, E. & Wacziarg, R. Ancestry and development: New evidence. *J. Appl. Econometrics* **33**, 748-762 (2018).

9. Creanza, N. *et al.* A comparison of worldwide phonemic and genetic variation in human populations. *Proceedings of the National Academy of Sciences* **112**, 1265-1272 (2015).

10. Mayer, T. & Zignago, S. Notes on CEPII's distances measures: The GeoDist database. (2011).

11. Desmet, K., Ortuño-Ortín, I. & Wacziarg, R. The political economy of linguistic cleavages. *J. Dev. Ec.* **97**, 322-338 (2012).

12. Mecham, R. Q., Fearon, J. & Laitin, D. Religious classification and data on shares of major world religions. *unpublished, Stanford University* **1**, 18 (2006).

13. Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M. & Perona, I. An extensive comparative study of cluster validity indices. *Pattern Recognit.* **46**, 243-256 (2013).

14. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* vol. 2008 P10008 (2008).

15. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* vol. 99 7821-7826 (2002).

16. Newman, M. E. J. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* vol. 103 8577-8582 (2006).

17. Natural Earth. http://www.naturalearthdata.com.

18. World Bank. *World Development Indicators*.

19. United Nations. Department of Economic and Social Affairs, Population Division. *World Marriage Data 2015*, POP/DB/Marr/Rev2015, (2015).

20. World Health Organization. *Global Health Observatory data repository*.

21. United Nations Food and Agricultural Organization (FAO). *FAOSTAT.*

22. Georgetown Institute for Women, Peace and Security Contact (GIWPS). *WPSIndex*.

23. United Nations Development Programme (UNDP). Human Development Report 2019.

24. World Economic Forum. *Global Gender Gap Report 2020*.

25. Helliwell, J., Layard, R., & Sachs, J. *World Happiness Report 2019*. (New York: Sustainable Development Solutions Network, 2019).

26. World Health Organization. *Global Health Observatory (GHO) data.*

27. World Economic Forum. *Global Competitiveness Index.*

28. Thomson, R. *et al.* Relational mobility predicts social behaviors in 39 countries and is tied to historical farming and threat. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 7521-7526 (2018).

29. Enke, B. Kinship, cooperation, and the evolution of moral systems. *Quarterly Journal of Economics* **134**, 953-1019 (2019).

30. Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. Global evidence on economic preferences. *Quarterly Journal of Economics* **133**, 1645-1692 (2018).

31. Falk, A., Becker, A., Dohmen, T., Huffman, D., & Sunde, U. The preference survey module: A validated instrument for measuring risk, time, and social preferences. IZA Discussion Paper No. 9674 (2016).

32. Hofstede, G., Hofstede, G. J. & Minkov, M. *Cultures and organizations: software of the mind.* (New York: McGraw-Hill USA, 2010).