


RESEARCH ARTICLE

History and cultural evolution: measuring the relationship through the Wikipedia network

Matthew J. Histén 

Department of Economics, California State University, Northridge, Northridge, CA, USA
Email: matthew.histen@csun.edu

(Received 3 August 2022; revised 28 January 2024; accepted 7 February 2024)

Abstract

Even at long time horizons, modern outcomes are in some sense bounded by history. Culture shapes how people interact and as it propagates across generations, groups with more common ancestors face less frictions to cooperation. This, in turn, affects institutional and technological diffusion, implying a society's history plays a crucial role in the causes of sustained long-run economic growth. To test this, we follow other studies by proxying for historical effects with genetic relatedness, which yields a temporal proportionality of shared common ancestry. Measuring cultural traits are more challenging. We develop a new systematic measure through network analysis of Wikipedia. Connectivity statistics over the encyclopaedia's hyperlink-directed network captures unique features of cultural relatedness. Further, as we index pages, we can coarsen the network into specific topics. The results show how history correlates broadly over a range of cultural factors. Differences across the coarsened networks demonstrate not simply that history matters, but where it matters less.

Keywords: cultural transmission; economic history; informal institutions; network analysis; technological diffusion

JEL classification: F14; O33; O47; O57; Z1

Introduction

History casts a long shadow. The fact of the matter is that there is enormous persistence in economic prosperity even at long time horizons. Modern outcomes in this sense are bounded by their histories, which lets us look at an outcome today and relate it to an item in its past. A key finding is that the history of populations matters much more than the history of locations (Putterman and Weil, 2010). These populations hold within them certain types of institutions, human capital, ideologies, norms – more broadly, culture. Culture shapes how cooperation takes place along a variety of activities. As it propagates across generations, groups with more common ancestors face less frictions to cooperating in these categories. Populations can more easily adopt new developments from societies similar to themselves, while those historically and culturally farther face higher barriers on the flow of technologies, goods, and people (Spolaore and Wacziarg, 2013). Thus, a society's history plays a crucial role in the diffusion of productivity enhancing innovations fundamental to the causes of sustained long-run economic growth. If economic development is a product of the past, how Laplacian are modern outcomes generally? Many studies have given evidence showing history matters (e.g. Acemoglu *et al.*, 2001; Ashraf and Galor, 2013; Nunn, 2020a; Spolaore and Wacziarg, 2016). We complement these results with novel data to demonstrate not simply that history matters, but where it matters less.

The mechanism, that ancestry affects culture and more similar cultures can more easily interact, can be empirically tested through measures of genealogical and cultural relatedness. Regarding the former, we have an idea of the composition of modern populations, which then approximates the cultural and

human capital carried along with them through their history. Genetic information from today's populations can describe the relations between populations in the past and serves as a summary statistic for a wide array of cultural traits and has been shown to correlate with a variety of proxies (e.g. Spolaore and Wacziarg, 2016). However, cultural traits are challenging to characterize and measure. Most empirical approaches use survey answers or experiments focusing on general attitudes related to trust, individualism, and perceptions of work and poverty (Alesina and Giuliano, 2015). But by pinning culture to explicitly stated preferences, these studies rely on idiosyncratic sources with results that are hard to generalize.

We contribute to this literature by developing a new systematic measure of culture through network analysis over Wikipedia. We derive cultural measures through statistics about article connectivity, an untapped data source in cultural research. Articles cross reference each other to create a directed network between topics. When aggregating network statistics across thousands of hyperlinks, we obtain a systematic measure of cultural relatedness between countries. Furthermore, because each page in Wikipedia is indexed by categories, we can disentangle similarity measures along different dimensions including religion, language, cuisine, and more. These measures are consistent with previous results across different source data including World Value Surveys, linguistic bifurcation graphs, religious genealogies, and others. Thus, Wikipedia data can be used to generate, separate, and compare cultural similarity across a range of dimensions, and statistical variations across them can remark on the relative reach of history.

The regression results are consistent with an argument that genetic distance serves as a summary statistic for cultural traits. Since the primary focus of this literature has been documenting historical persistence, little attention has been paid to its dampening (Nunn, 2020a). We further the argument with a process of how history replaces itself over time through impersonalizing institutions (Henrich, 2020, 48; North *et al.*, 2009, 2). Here, we describe how a modularizing of social roles makes a person (e.g. son of Uther) interchangeable between a persona (e.g. citizen of England). As a social persona becomes more anonymized from personal identity – as institutions become more impersonalized – interactions spread over wider areas of social behaviour without needing to be cognizant of the individual characteristics of a partner. Participants trust the persona instead of the person, increasing mobility between parts and attenuating regional or personal lock-in effects. The results suggest that a modularizing of social relationships offers some counterbalance historical determinism.

Overall, these results support the value of Wikipedia as data. The network statistics are consistent with previous questions and shed light on new ones. Because the measures are systematic and can be parsed in a variety of ways, Wikipedia as data is primed for further research proxying cultural salience. The paper proceeds as follows. The second section describes the related literature. The third section describes the data, including our novel measure of cultural institutions through a network analysis over Wikipedia and its indexing system. The penultimate section provides the empirical results and discussion, and we offer concluding remarks in the last section.

Related literature

Institutions

The argument – that history affects cultural relatedness which in turn influences how well groups cooperate – connects to a rich literature in economic history and evolutionary cultural transmission (e.g. Acemoglu *et al.*, 2001; Engerman and Sokoloff, 1997; Nunn, 2020a, 2020b; Rubin, 2014). Researchers coordinate on a definite conclusion: history persists in the long term. They also share two challenges: obtaining appropriate evidence and teasing out causal mechanisms. Regarding the former, the scholarship depends on the art of compiling novel data that convincingly trace some past event and its developments across time. Many pieces of a larger historical puzzle have been identified, but many remain buried. Nor have many studies comprehensively approached how those pieces fit together (Nunn, 2020a). Although the extent of these channels and their complementarities are yet to be understood, significant progress has been made in empirically testing the lasting impacts of historical events (Nunn, 2020a). In particular, the literature moved from examining proximate

determinants of growth to analysing more fundamental factors (Spolaore and Wacziarg, 2013). Studies on technology, productive capacities, organization of scale economies, etc. may reveal differences in outcomes. But they are not causes of growth – they *are* growth (North and Thomas, 1973, 2). This interpretation emphasizes the institutional differences across societies.

Institutions are the ‘rules of the game’ that standardize the patterns of interaction, encompassing formal rules and social conventions, written laws, informal norms, and shared beliefs (North, 1990, 3). Standardization, by definition, implies interactions conform to an expectation. Institutions define the form of such association (Hodgson, 2006). These rules provide in advance knowledge about how an interaction will proceed, providing a means for planning, predicting, and cooperating (Hayek, 2011, 268). Many researchers have asked which institutions survive given their context and behaviour. The process of establishing the right institutions is endogenous (Chang, 2010; Rubin, 2017, 251). For example, the consequences of European colonialization stem from settlers introducing ‘good’ institutions that led to rich outcomes today (Spolaore and Wacziarg, 2013). Generally, these were institutions supporting trade and innovation. But these settlers also brought themselves – that is, their human capital (Glaeser *et al.*, 2004). They carried with them scientific and technical knowledge, access to international markets, and human capital creating institutions, *and* they brought ideologies, values, social norms, etc. (Easterly and Levine, 2003). This theoretical ambiguity is equally consistent with the empirical evidence.

Institutions, culture, and human capital lie at the heart of the debate regarding patterns of comparative development over the past few centuries (Ashraf and Galor, 2013). Disentangling societal characteristics – be they institutions, norms, culture, etc. – is incredibly challenging because these variables are interdependent and complementary (Chang, 2010; Spolaore and Wacziarg, 2013). The bottom line is that comparative development patterns are rooted in the composition of populations (Spolaore and Wacziarg, 2016). These specific histories play an undeniable role in modern outcomes. Success rests on the fact that an individual can benefit from more knowledge than she is aware of through products of cumulative growth in cultural adaptations (Hayek, 2011, 90). The notion of culture to economists remains rather vague (Nunn, 2020a), and definitions of ‘institutions’ often overlap with definitions of ‘culture’ (Alesina and Giuliano, 2015; Spranz *et al.*, 2012). Still, empirical papers attempt to distinguish formal institutions (formal legal or regulatory systems) from informal ones (norms, beliefs, and informal rules), but these tend to be performed in isolation of each other (Alesina and Giuliano, 2015). Moreover, merely bucketing what’s left after subtracting a legal apparatus as ‘informal institutions’ or ‘culture’ is both theoretically and empirically insufficient.

Culture

To avoid relegating either to secondary importance, we take a holistic view of culture as encompassing a society’s formal and informal institutions affecting cooperation. That is, institutions are the rules of the game *and* how we play it. Cultural similarity aims at how easily populations can share knowledge, communicate, and interact. It is often unclear how to distinguish between formal and informal institutions as conventions and norms tend to become codified overtime, which feedback to affect the evolution of these conventions and norms (Nunn, 2020a). Moreover, informal institutions are often prerequisites to formal ones (e.g. democratic values are necessary for legalized democratic institutions to operate [Besley and Persson, 2019]). Ultimately, similar cultural institutions (including formal ones) influence the diffusion of complex technological innovations (Spolaore and Wacziarg, 2016) and the facilitation of trade flows (Guiso *et al.*, 2006).

However, culture is hard to pin down. Methodologically, a focus on culture struggles to bring to bear tools of quantitative methods to parse the characteristics of specific cultural traits. The quest for a summary measure of cultural differences is challenging since many cultural practices – food, dress, accent, etc. – are not collected in any systematic manner (Abramitzky, 2015). Most empirical papers represent culture as the broad attitudes and values related to generalized trust, individualism, and perceptions of work and poverty (Alesina and Giuliano, 2015). Typically, these values are

measured through survey questions aggregated at the country level (e.g. the World Values Survey) and then correlated with economic outcomes. These studies calculate a culture trait through questionnaires and correlate it to an aggregate measure (e.g. more individualistic cultures are less wedded to tradition and less likely to show nepotism, which affects the organization of companies and politics and has consequences for long run growth [Acemoglu and Robinson, 2012; Gorodnichenko and Roland, 2011]). Another line of research involves economic experiments. For example, participants from different cultures are briefed on a coordination game (e.g. the dictator game Duhaime, 2015) or the ultimatum game [Henrich, 2000]) or a task with priming (e.g. unscramble sentences with words influenced by a studied topic such as religion [Norenzayan and Shariff, 2008]). Their results are again interpreted as an index of attitudes and values (e.g. individualism) which can be aggregated and regressed against different outcome variables. Studies show relationships across a range of phenomena ranging from cooperation with strangers, voluntary blood donations, or parking tickets for diplomats (Everett *et al.*, 2016; Henrich, 2020, 212).

Clearly, these approaches detect intriguing relationships. But imposing a top-down structure of culture framed by predefined traits or attitudes constrains broader connotations of culture. These studies also tend to collapse culture into a single proxy such as trust, individualism, or whether two countries have been to war with each other (e.g. Guiso *et al.*, 2006). This often lacks generalizability since these proxies are unique and from non-comparable data sources, yet measures are aggregated and conclusions drawn across studies (Alesina and Giuliano, 2015). Furthermore, they miss crucial elements that are harder to represent as a regression variable, or aspects that haven't been thought of altogether. Elements such as music, cuisine, fashion, etc. matter significantly in how populations interact but aren't easily quantified through surveys or experiments.

Quantitative methods over big data sources are beginning to offer more generalizability. Parsing large corpuses of text to investigate cultural ideas through dictionaries or topic models has been used to generate cultural proxies in regression models (Blaydes *et al.*, 2018; Grajzl and Murrell, 2019). For example, Michalopoulos and Xue (2021) analyse cultural attitudes by analysing topics from country specific folklore. We contribute to the study of culture by extending these big data analyses to networks. We argue that the network structure of Wikipedia provides a novel dimensional space to measure cultural relatedness. A country's Wikipedia page contains thousands of hyperlinks to other pages within the encyclopaedia. These links reveal aspects of their cultural attention, and clusters around salient items are quantifiable.¹ When aggregated across thousands of links for each country, we obtain a broad, systematic measure of cultural similarity between countries. Similarity is decentralized and emergent compared to textual analyses which involve more researcher judgement over which corpuses are included. Moreover, Wikipedia categorizes each page into topics. Separating by category tags, then, allows us to disentangle relationships along the lines of religion, language, cuisine, and more.

Overall, the body of research studying comparative development increasingly emphasizes the role of ethnic, linguistic, religious, and other cultural effects (Michalopoulos and Papaioannou, 2016; Michalopoulos and Xue, 2021; Nunn, 2020b). We hope to open the door to additional channels of cultural impact and unearth even deeper roots. In investigating these relationships, our paper is most similar to Spolaore and Wacziarg (2016), Ashraf and Galor (2013), Desmet *et al.* (2012), and Desmet *et al.* (2011). Each study corresponds a proxy of history from genetic measures onto contemporary outcome variables with positive correlations.² In particular, the outcome variables in Spolaore

¹For example, consider cultural activities through sports. Even if India's Wikipedia page does not contain a direct link to 'baseball', there's a short path to that topic through other pages in the network (e.g. 'cricket'). Properties about this distance provide a measurable closeness around different cultural items. In fact, to celebrate 60 years of diplomatic relations, in 2013 joint efforts by the Japanese and Indian governments created an anime series called 'Suraj: The Rising Star'. A reboot of a Japanese anime about baseball, its re-release in India centred around cricket, and required few other changes. The leaders of the project argued that it synthesized Indo-Japanese values and strengthened economic, diplomatic, and cultural traditions across the two countries.

²Technically, Ashraf and Galor (2013) trace a hump-shaped effect whereby a moderate level of genetic diversity outperforms too much or too little.

and Wacziarg (2016) consider three measures of culture. First, they construct an attitudinal space from World Values Survey answers (similar to Desmet *et al.*, 2011 and Spolaore and Wacziarg, 2009) across 100 questions. Second, they measure linguistic differences through lexicostatistics of language evolution (similar to Desmet *et al.*, 2012). Third, they develop an index of religious difference through bifurcation diagrams tracking major separations between religious traditions. Under each index, constructed from different data sources and methods, their results find positive correlations between genetic and cultural distance after controlling for confounding effects. We generate results consistent with equivalent categorical dimensions from a systematic method over a single data source and extend the analysis onto categories not yet considered. Thus, Wikipedia as data offers several advantages compared to previous research. Its decentralized, emergent structure allows for an extemporaneous capture of broad cultural traits *versus* predefined buckets. Its massive size stretches across many topics to allow for a broad and systematic characterization of culture, and its organizational characteristics allow for slicing the data along many dimensions. Wikipedia as data provides a novel measure of culture and connecting it to genealogy reinforces old answers and contributes new insights regarding critical phenomena in comparative economic development.

Data and methodology

We proxy for historical effects through genetic relatedness and cultural measures constructed by the network connectivity of Wikipedia pages. The genetic data have been developed by population geneticists in other studies. The Wikipedia data are novel. While it has been used by some researchers to estimate election (Smith and Gutafson, 2017) or health outcomes (Smith, 2020) from article page views, the use of Wikipedia as data for other relationships is quite limited. In this section, we describe structural features of these data with an emphasis on the encyclopaedia project, its hyper-link structure, and its indexing system. Although there are reasons to be cautious, we clarify the encyclopaedia's reliability in its usage as an organizational structure *versus* content verification. We webscraped links across pages to generate adjacency matrices for each cultural category in the network analysis.³

Genealogy

A genetic measure explains how distant human societies are from each other. The argument is not that such a measure captures anything meaningful about genetic traits; it simply serves as a molecular clock characterizing relatedness between populations in terms of the number of generations separating them from a common ancestor population. Genetic information from today's populations can describe the relations between populations in the past and allows for a reconstruction of a history of humankind. Groups with more common ancestors face less frictions to cooperating and can more easily adopt new developments from societies similar to themselves, while those historically and culturally farther face higher barriers on the flow of technologies, goods, and people. Genetic similarity contains a temporal proportionality because it contains markers of shared common ancestry. Although all humans share the same set of alleles, they appear in different frequencies in different populations. These frequencies vary as an effect of genetic drift over time. A random change in allele frequency between generations affects the frequency distribution of the next. Given the combinatoric space of possibilities, it is much more likely that populations with similar distributions share common ancestors rather than those distributions arising out of chance.⁴

³Unique Python packages used include Scikit-learn, BeautifulSoup4, and NetworkX. Documentation for these packages is open source and scripts are available upon request.

⁴The English having similar allele frequencies to the Danish implies they share recent ancestry. Those frequencies are less similar between the English and the Turks, and even less so between the English and the Malagasies, implying their common ancestry is generationally farther away (Spolaore and Wacziarg, 2016). In fact, the smallest genetic distribution observed in the data is between the English and the Danish. The largest distance is between Mbuti Pygmies and Papua New Guineans.

By considering many such markers, population geneticists have been able to measure global differences across populations (Cavalli-Sforza *et al.*, 1994). This has only improved with advances in DNA sequencing that consider allele frequency more precisely. The commonly used fixation index (F_{ST}) is estimated from genetic polymorphism data from samples of individuals in thousands of locations across the world. The index takes the ratio of allele variance in different subpopulations (weighted by the size of the subpopulation) relative to the variance of the allelic state in the total population to track separations over time. The F_{ST} provides a type of clock since two populations shared their last common ancestors. Although these genetic distance data were originally collected at the population level (Cavalli-Sforza *et al.*, 1994), economists used ethnic composition data by country (Alesina *et al.*, 2003) to match F_{ST} genetic distance between countries (Spolaore and Wacziarg, 2009). The results provide bilateral distances between country pairs with available genetic data. The dataset constructed by Spolaore and Wacziarg (2009) attempted to match 207 countries and dependencies. However, matches could not occur where ethnic group share data were limited (as was the case for many island countries), leaving approximately 180 country pairs for 16,000 observations. Because many countries contain sets of subpopulations that are genetically distant (like the United States), they also constructed another weighted measure of expected distances. The measure yields the expected genetic distance between two randomly selected individuals from each country pair. These measures are highly correlated, and though the results below show the weighted measure, the other yields very similar results.

Wikipedia

Open-source collaboration involves a unique production model where users trade effort outside of traditional organizational forms such as firms or markets (Langlois and Garzarelli, 2008). Provided the architecture for interaction is well-established, users select into activities suiting their specialization. Wikipedia is an open-source encyclopaedia project based on a model of freely editable content. By taking advantage of dispersed and idiosyncratic knowledge, 125,000 active monthly contributors and 40 million total accounts (Wikistats, 2021) exchange their specialized knowledge currently. This has led Wikipedia to becoming far more comprehensive than any other encyclopaedia. At the time of analysis, it contained more than 57 million articles across 300 languages, with over 6 million articles written in English alone (Wikistats, 2021).

Because it's written collaboratively by anonymous volunteers, anyone can edit its media, leaving articles prone to misinformation, errors, or vandalism. These concerns vary substantially across language versions and articles.⁵ Reliability is by far most consistent in the English version (Steinsson, 2023), which populates the data analysed here. In the English Wikipedia, several mechanisms are in place to maximize contributions while preserving quality. Articles are standardized through an established manual of style, both for ordering of sections and formatting. The site tries to operate under strict policies of viewpoint neutrality and information verifiability. Aiming to be only a tertiary source, Wikipedia is explicit about not performing original research and instead acting entirely on previously published information, requiring material to be verified through qualifying sources with direct inline citation. Sources must also be publicly available in some form. When claims are exceptional, multiple high-quality sources are required.

Wikipedia operates by open and transparent consensus for editorial disputes, too, with well-defined processes for resolution. Editorial control falls on a network of self-organizing editors subject to peer review. These administrative roles are only achieved by nomination, which helps enforce communal standards of conduct backed by experience and familiarity with its content policies. Thousands of active editors will be using, monitoring, or editing the articles and assisted by automated programs to help watch for problematic edits (Wikistats, 2021). Popular or controversial articles are reviewed

⁵Note that language versions of the encyclopaedia are not translations of each other. Rather, contributors of each language generate their own articles. This means articles can vary substantially between versions, if they exist across them at all.

more closely to make it harder for vandalism. An arbitration committee sits on top of all editorial disputes, composed of elected members in regularly rotated tranches. This committee has the authority to impose binding solutions that the community struggles to resolve on its own.

Because Wikipedia has so many users, articles are scrutinized by a massive number of readers and editors, with every edit logged and every debate about edits logged. The fact that anyone can edit articles increases the chances that errors will be corrected rather than propagated. Reliability of content has been analysed many times. Although results vary, project outcomes are generally positive (Mesgari *et al.*, 2015; Petiška and Moldan, 2019; Steinsson, 2023). Since 2005, Wikipedia has shown an error rate similar to Britannica (Giles, 2005) with most vandalism resolved within 5 minutes (Viégas *et al.*, 2005). Although most contributors are deeply committed to curating accurate information according to their dispersed knowledge, criticisms are justified over content errors and editorial processes. There have been many examples of false or misleading information persisting for months, a systematic political bias (Greenstein and Feng, 2012), and demographic concerns related to the composition of editors who skew heavily male from Western countries (Meyer, 2013). Given that there is no systematic process to generate ‘obviously important’ articles, they are populated by the relative attention of its users’ linguistic and cultural interests.

Can the English Wikipedia be trusted, then? The short answer is, on average, yes, and accuracy improves substantially for popular articles. Popular articles, by the nature of network analysis, populate most of the data in this study. Even if we discount each article’s efficacy, the unit of analysis is not the content itself but the network of hyperlinks cross-referencing articles. In fact, this structure is even more emergent than the generation of articles and harder to intentionally manipulate. Ultimately, Wikipedia is a widely contributed platform aggregating knowledge from many different viewpoints, offering a unique insight into different cultures’ attention and priorities.

Data collection

Wikipedia as data involves representing the robust network of hyperlinks between pages as a mathematical object. All articles are linked or cross-referenced throughout the encyclopaedia to guide users to related pages with additional information. This network, then, can be transformed into a directed graph. As part of its manual of style, contributors are encouraged to link to other articles with relevant connections regarding background information, events, proper nouns, technical information, and more (Wikistats, 2021). Downward pressure exists to avoid overlinking to pages considered commonly understood. These pathways form a network structure that connects topics organically. To derive this network, we began with 207 countries of the 20th century (following Spolaore and Wacziarg, 2009) and scraped all internal hyperlinks that appear in paragraph html. After removing redundant, administrative, and miscellaneous pages, each country had on average 523 content links to other material articles. Italy had the most with 1,700 while Democratic Republic of Congo had the fewest with 100 links. Since the analysis is over the English Wikipedia, Western countries receive more attention by contributors. Further research can clarify these differences across language versions of the encyclopaedia. Collecting these country page links summed to about 70,000 unique pages, which were also scraped for all their internal hyperlinks appearing in paragraph html. This generated an adjacency matrix of about 1.5 million unique articles.

Simultaneously, all category tags from each page were scraped. Wikipedia is so vast that several features have been introduced to facilitate navigation, including categorizing every page by topics. Although categorization efforts are continuous, the bottom of every page contains relevant tags that classify pages into larger groupings. Each category can be browsed for related pages clustered under the same tag to create an interconnected hierarchy. Similar to the content itself, categorization develops collaboratively by end users applying public tags. Wikipedia’s general taxonomy contains about 40 irreducible topics.⁶ After collecting tags for all articles in the matrix, we can coarsen the network into

⁶This list contains {‘Academic disciplines’, ‘Business’, ‘Concepts’, ‘Culture’, ‘Economy’, ‘Education’, ‘Energy’, ‘Engineering’, ‘Entertainment’, ‘Entities’, ‘Ethics’, ‘Events’, ‘Food and drink’, ‘Geography’, ‘Government’, ‘Health’, ‘History’, ‘Human

specific categories (e.g. drop all pages not related to a topic). We took two approaches to this categorization: ‘All tags’ and ‘First Tag Only’. The former considers all relevant pages and therefore generates a larger network, while the latter considers category tags without duplication (i.e. each network is mutually exclusive regarding pages contained). Their development is explained in detail in [Appendix A](#), and results from both are provided in the analysis below. [Table 1](#) presents more detailed information on the coarsened categories, which Wikipedia main topics fall under their umbrellas, and additional information related to the types of pages.

Empirical analysis

Dependent variable

Using tools from network analysis, we move from qualitative concepts of cultural similarity to quantitative measures through clusters in the network. A network simplifies a system to an abstract structure preserving only its connections (Newman, 2010, 105). These connections describe the paths along which something flows. Centrality statistics are real-valued measures of importance along this interpretation, where values provide rankings of critical nodes (Newman, 2010, 160). Applications of centrality include super-spreaders of disease or key infrastructure in transportation networks. One of the most common is degree centrality. The degree of a node refers to the number of connections it has to other nodes, which can be interpreted as a measure of likelihood for a node to catch flows through the network. Degree centrality, then, considers the number of links incident upon a node to rank its relative importance by weighting the number of walk lengths across the network (Newman, 2010, 160). For example, consider the network shown in [Figure 1](#). The node with the highest degree centrality would be ‘1’ which is a connection point for most information flowing through the network, while node ‘7’ would have lowest degree centrality. For Wikipedia, the network forms from cross-referenced hyperlinks between articles and flow captures how information propagates through nodes. [Table 2](#) shows articles with the highest degree centrality of the different subnetworks.⁷

Our dependent variable comes from a network statistic that effectively inverts centrality measures: connectivity. Connectivity measures the minimum number of paths that need to be removed before two nodes are no longer connected (Newman, 2010, 262). This can be visualized in terms of bottlenecks. In the example from [Figure 1](#), the number of required separations for ‘1’ and ‘7’ (1 cut) differs from ‘1’ and ‘4’ (2 cuts). A larger cut set implies the network has to be dismantled more substantially to separate pairs, implying more similarity between nodes. In the Wikipedia network, we consider linkages between country pairs, which can take hundreds of cuts to separate. The statistic interprets similarity according to the commonality of topics and articles between them. The approach is slightly more complicated because hyperlinks run in a specific direction; a path along the network from one page to another is not necessarily reversible. Therefore, we measure how strongly nodes are independently connected through a connectivity min-cut statistic normalized by the average number of cuts needed to disconnect two nodes in the network.

To ensure connectivity accurately characterizes the network structures, we run robustness tests with two additional network statistics. The first calculates the cosine similarity between a country pair’s links, which essentially measures how many connections throughout Wikipedia a country pair has in common normalized by the first term’s total connections. Second, we consider a shortest path measure, which calculates the fewest number of nodes between country pairs from every hyperlink in one country page to the other normalized by the average distance between any two paths. Results are consistent across all network statistics. We also test if results are sensitive to the topic boundaries themselves (e.g. the ‘business’ or ‘technology’ tag rolling up different topic combinations). Results are again consistent, so we report outcomes from the largest network. Finally, results are also

behavior’, ‘Humanities’, ‘Industry (economics)’, ‘Information’, ‘Knowledge’, ‘Language’, ‘Law’, ‘Life’, ‘Mass media’, ‘Mathematics’, ‘Military’, ‘Music’, ‘Nature’, ‘People’, ‘Philosophy’, ‘Politics’, ‘Religion’, ‘Science’, ‘Society’, ‘Sports’, ‘Technology’, ‘Universe’}.

Table 1. Coarsened network criteria by Wikipedia main topic classifications

Regression variable	Wikipedia main topic classifications	Sample pages	All tags network size	First tag only network size
Business	{'Business', 'Economy', 'Industry'}	{'World Trade Organization', 'Bloomberg L.P.', 'Lome Convention', 'Kmart', 'Jollibee', 'Outsourcing', 'Tender board', 'Bretton Woods Conference', 'Regional Comprehensive Economic Partnership', 'Market capitalisation'}	68,565	3,740
Education	{'Education'}	{'College-preparatory school', 'Universitas 21', 'Bocconi University', 'Art schools', 'Licentiate (degree)', 'Science, technology, engineering, and mathematics', 'Direct Subsidy Scheme', 'The three Rs', 'Catholic school'}	29,195	9,840
Government	{'Government', 'Politics'}	{'National personification', 'Commonwealth republic', 'Royalism', 'Dwifungsi', 'Ministry of Interior', 'Carousel voting', 'Arengo', 'Supermajority', 'Wealth tax', 'Dutch Sandwich'}	382,030	24,132
Language	{'Language'}	{'Ghanaian Pidgin English', 'Ie (digraph)', 'Deutsches Wörterbuch', 'Adyghe language', 'Jopara', 'Ishkashimi language', 'Ems Ukaz', 'Qatari Sign Language', 'Tajik alphabet', 'Cursive script (East Asia)'}	92,919	12,802
Military	{'Military'}	{'Brigade (military)', 'Battle of Salamis', 'Balkan Wars', 'Conscription', 'Nuclear proliferation', 'Fieldcraft', 'Lajes Air Base', 'V-2 rocket', 'Biological Weapons Convention', 'Conscientious objection'}	88,618	28,565
Religion	{'Religion'}	{'Saltigue', 'Stations of the Cross', 'Iviron monastery', 'Millenarianism', 'Shwedagon Pagoda', 'Neoplatonists', 'Hindu calendar', 'Temple of the Tooth', 'Ise Grand Shrine', 'Islamic cosmology'}	86,458	11,702
Social	{'Culture', 'Entertainment', 'Food and Drink', 'Music', 'Sports'}	{'Tinkling', 'Steinstossen', 'Iranian art', 'Beef Stroganoff', 'Asphyx', 'Hua-Yi distinction', 'Batwing sleeves', 'Pakol', 'Walkabout long Chinatown', 'The Adventures of Pinocchio'}	297,572	67,472
Technology	{'Engineering', 'Mathematics', 'Technology'}	{'Stabilizer (aircraft)', 'Satellite dishes', 'Transistor', 'Welding', 'Nuclear power plants', 'Materials science', 'Semiconductor fabrication plant', 'Vernacular architecture', 'Stem cell', 'Wireless internet'}	62,056	6,900

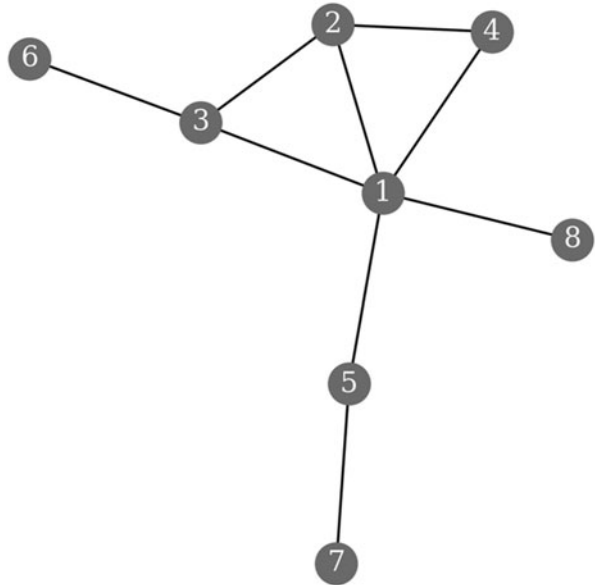


Figure 1. Force-directed graph drawing of a randomly generated network.

consistent with alternative genetic measures (e.g. plurality instead of weighted), and the Wikipedia measures correlate with statistical significance to cultural proxies in previous research.⁸

Regression results

We test the hypothesis that longer ancestral separation correlates with less similar culture, with the novel culture variable derived by comparing the connectivity of topics and articles between countries' Wikipedia pages. Descriptive statistics for the ordinary least squares regimes are listed in Table 3. All results consider treatments unique to network analysis for unobserved heterogeneity across neighbouring observational units by using arbitrary clustering structures (Colella *et al.*, 2019). The regression takes the following form:

$$N_{ij} = \beta_0 + \beta_1 F_{ij} + \beta_k X_{ij} + \varepsilon$$

where subscripts i, j refer to country-direction pair, N refers to the connectivity statistic between pages, F refers to the F_{ST} distance between them, and X is a vector of k controls between countries. Because ancestry similarity owes much to geographic proximity, and to mitigate concern that the relationships merely go through geographic distance, we include controls for contiguity, geography, water, and historical relationships, elaborated in Appendix B.

First, we consider the results of genetic relatedness on cultural similarity broadly. Table 4 shows the regression results when considering country pairs across the complete sample of Wikipedia pages. We see a robust statistical relationship with the signs in the expected direction for columns 1 and 2, which include data from all countries in the sample. Note that the smaller the F_{ST} distance, the closer the ancestry. Since connectivity measures how many paths must be broken to disconnect two nodes, a larger number indicates more links in common for the country pair. Table 4 columns 2 and 4 therefore show that seven fewer F_{ST} units (more similar genetics) imply one additional normalized connection (more similar cultures).⁹ Moreover, given the effect of population movements after the discovery

⁸The measures will not correlate perfectly because network statistics capture connections that tree-branching systems of languages (e.g. Desmet *et al.*, 2012) or religions (e.g. Spolaore and Wacziarg, 2016) do not (e.g. religious bifurcations won't measure concept similarities; lexicostatistics won't capture dialect or slang).

⁹In log terms, a 0.1% decrease in genetic dissimilarity leads to about a 1% increase in network connectivity.

Table 2. Descending order of articles by highest degree centrality for the first 10 nodes

Regression variable	All tags degree centrality	First tag only degree centrality
Business	{'World Bank', 'London', 'Euro', 'Multinational corporation', 'Obama Administration', 'World Trade Organization', 'Currency', 'Kyoto Protocol', 'OECD', 'Socialism'}	{'World Trade Organization', 'Fiscal year', 'IMF', 'Eurasian Union', 'Inditex', 'Trump administration', 'Cairns Group', 'Foreign exchange market', 'Outsourcing', 'Forbes'}
Education	{'UNESCO', 'Public University', 'University', 'London', 'Columbia University', 'Bachelor of Science', 'Higher education', 'QS World University Rankings', 'University of Oxford', 'Age of Enlightenment'}	{'University', 'Bachelor of Science', 'Higher education', 'QS World University Rankings', 'Medical school', 'Master's degree', 'Kindergarten', 'Times Higher Education World University Rankings', 'Secondary education', 'Academic Ranking of World Universities'}
Government	{'Ottoman Empire', 'Second World War', 'London', 'Paris', 'World Bank', 'UNESCO', 'Joseph Stalin', 'Socialism', 'New York City', 'Democracy'}	{'Democracy', 'Socialism', 'World Bank', 'Political party', 'UNESCO', 'Proportional representation', 'European Parliament', 'United Nations Security Council', 'Nationalism', 'Republic'}
Language	{'English language', 'Latin', 'French language', 'Arabic language', 'Greek language', 'Spanish language', 'German language', 'Portuguese language', 'Russian language', 'Italian language'}	{'English language', 'French language', 'Spanish language', 'Latin', 'Arabic language', 'Lingua franca', 'Portuguese language', 'Dialect', 'Italian language', 'Russian language'}
Military	{'Ottoman Empire', 'Second World War', 'Royal Navy', 'Paris', 'Vietnam War', 'British Army', 'Royal Air Force', 'Winston Churchill', 'First World War', 'Joseph Stalin'}	{'Royal Navy', 'British Army', 'International military intervention against ISIL', 'Iraq War', 'First World War', 'Conscription', 'Military service', 'Battalion', 'Nuclear weapon', 'Militia'}
Religion	{'Islam', 'Christianity', 'Greek language', 'Buddhism', 'Roman Catholic Church', 'Jerusalem', 'Muhammad', 'Rome', 'Hinduism', 'Jesus'}	{'Islam', 'Christianity', 'Buddhism', 'Hinduism', 'Muhammad', 'Judaism', 'Jesus', 'Roman Catholic Church', 'Hindu', 'Muslim world'}
Social	{'New York City', 'Paris', 'Greek language', 'London', 'Islam', 'Christianity', 'The New York Times', 'German language', 'Russian language', 'North America'}	{'FIFA World Cup', 'Basketball', 'Jazz', 'Rock music', 'Folk music', 'UEFA', 'The New York Times', 'Cricket', 'Opera', 'Pop music'}
Technology	{'London', 'The New York Times', 'NASA', 'International airport', 'Internet', 'Petroleum', 'California', 'South America', 'Carbon dioxide', 'Vienna'}	{'International airport', 'NASA', 'Runway', 'Indian Space Research Organisation', 'Satellite', 'Standard gauge', 'Nuclear power', 'Airport', 'California', 'Uranium'}

of the New World, columns 3 and 4 exclude pairs between the Americas or Oceania to only examine Old World countries. Results are consistent, demonstrating the strong effect genetic history holds on cultural similarity even when we recognize significant population movements.

The more interesting feature of these data stems from separating culture into different categories. In [Tables 5 and 6](#), the network was coarsened to only include pages from a relevant category tag along the two measures discussed in the previous section. We consider eight dimensions from [Table 1](#): business, education, government, language, military, religion, social, and technology. Aspects of how societies interact along these dimensions are influenced by intergenerational characteristics. Of course, these effects differ across categories and across societies, and other forces might superimpose an order without a corresponding genetic effect. Note that because 'All tags' categories can overlap, the increased commonality of articles washes out some network uniqueness and curtails the effects.

Across all regressions, we see a robust statistical relationship with expected signs. Given the construction of the connectivity variable, in absolute value terms, a smaller coefficient implies less impact

Table 3. Descriptive statistics for the explanatory/dependent variable regimes

Variable	Mean	Std. dev.	Min	Max
Genetic distance				
F_{ST} , weighted	3.698	1.852	0.000	9.496
F_{ST} , plurality	3.768	2.245	0.000	10.64
Connectivity distance				
All	125.7	74.19	32.637	930.8
Business	14.59	9.689	1.447	147.0
	4.095	1.832	0.840	26.31
Education	8.009	5.187	1.120	76.81
	4.370	2.154	1.042	33.60
Government	35.59	23.00	3.225	377.1
	13.08	7.468	1.271	131.6
Language	15.02	9.616	1.610	102.1
	7.235	3.692	1.597	39.02
Military	14.88	10.44	1.819	164.9
	5.960	3.385	1.461	53.81
Religion	12.78	9.190	1.823	121.1
	6.137	3.442	1.412	46.85
Social	30.87	21.96	1.279	256.1
	12.37	8.138	1.035	112.2
Technology	12.04	8.493	1.252	118.3
	5.370	3.206	0.983	49.90

For the eight categorical variables, the first row shows statistics for the ‘All tags’ network, and the second row shows the ‘First Tag Only’ network.

Table 4. Regression results for the full network with connectivity as the dependent variable

	(1) All	(2) All	(3) Old World	(4) Old World
Weighed F_{ST} distance	-10.10***	-7.21***	-11.00***	-7.43***
	(0.233)	(0.258)	(0.270)	(0.384)
Contiguity controls	n	y	n	y
Geography controls	n	y	n	y
Water controls	n	y	n	y
History controls	n	y	n	y
Constant	1.35***	0.92***	1.37***	0.94***
	(0.011)	(0.034)	(0.013)	(0.039)
Observations	15,576	15,576	10,011	10,011
R^2	0.105	0.31	0.127	0.323

Coefficients for controls are provided in [Appendix B](#) and refer to [Table 1](#) for details regarding the number of items in the network statistic calculation.

Robust standard errors in parentheses.

*** $P < 0.01$, ** $P < 0.05$, * $P < 0.1$.

Table 5. Regression results for the ‘All tags’ submatrices with connectivity as the dependent variable

	(1) Business	(2) Education	(3) Government	(4) Language	(5) Military	(6) Religion	(7) Social	(8) Technology
Weighed F_{ST} distance	-1.25*** (0.047)	-0.60*** (0.026)	-3.08*** (0.104)	-1.32*** (0.046)	-1.20*** (0.046)	-1.15*** (0.040)	-2.80*** (0.103)	-1.05*** (0.042)
Contiguity controls	y	y	y	y	y	y	y	y
Geography controls	y	y	y	y	y	y	y	y
Water controls	y	y	y	y	y	y	y	y
History controls	y	y	y	y	y	y	y	y
Constant	0.16*** (0.006)	0.08*** (0.003)	0.38*** (0.015)	0.16*** (0.007)	0.16*** (0.007)	0.12*** (0.006)	0.34*** (0.015)	0.13*** (0.006)
Observations	15,576	15,576	15,576	15,576	15,576	15,576	15,576	15,576
Centred R^2	0.287	0.254	0.315	0.333	0.302	0.346	0.324	0.289

Coefficients for controls are provided in [Appendix B](#) and refer to [Table 1](#) for details regarding the number of items in the network statistic calculation.

Robust standard errors in parentheses.

*** $P < 0.01$, ** $P < 0.05$, * $P < 0.1$.

Table 6. Regression results for the first tag only submatrices with connectivity as the dependent variable

	(1) Business	(2) Education	(3) Government	(4) Language	(5) Military	(6) Religion	(7) Social	(8) Technology
Weighed F_{ST} distance	-1.45*** (0.095)	-2.32*** (0.113)	-9.34*** (0.350)	-4.68*** (0.178)	-3.53*** (0.159)	-3.19*** (0.155)	-8.90*** (0.401)	-3.59*** (0.162)
Contiguity controls	y	y	y	y	y	y	y	y
Geography controls	y	y	y	y	y	y	y	y
Water controls	y	y	y	y	y	y	y	y
History controls	y	y	y	y	y	y	y	y
Constant	0.43*** (0.013)	0.46*** (0.014)	1.34*** (0.049)	0.69*** (0.025)	0.64*** (0.020)	0.61*** (0.020)	1.47*** (0.057)	0.57*** (0.022)
Observations	15,576	15,576	15,576	15,576	15,576	15,576	15,576	15,576
Centred R^2	0.168	0.205	0.268	0.251	0.263	0.277	0.291	0.219

Coefficients for controls are provided in [Appendix B](#) and refer to [Table 1](#) for details regarding the number of items in the network statistic calculation.

Robust standard errors in parentheses.

*** $P < 0.01$, ** $P < 0.05$, * $P < 0.1$.

of ancestry on culture – that is, less sensitivity to historical effects. Although differences in network sizes allow some hesitation in directly comparing coefficients, the R^2 statistics offer additional interpretation. Here, differences in how much the model explains across categories captures an additional sensitivity to historical effects. These statistics reveal a pattern: those categories consistently least affected by ancestry include business, education, and technology. Those most affected are government and social. Some of these measures (language or religion) have been tested by previous researchers and are consistent with their findings (e.g. Desmet *et al.*, 2011, 2012; Spolaore and Wacziarg, 2016). Others, such as technology or government, are much harder to obtain value measures through previous methods, and the results here provide new insights on these dynamics. The consistency with previous research supports analysis using Wikipedia as data and corroborates new findings in traditionally hard to characterize domains.

Discussion

The above results provide additional evidence on the relationship between history and modern outcomes. One interpretation of variations across categories can be understood as variations in sensitivity to historical effects. We can map these variations according to specific features of the categories. In particular, we emphasize their divergent role in impersonalizing institutions. That is, historical sensitivity hinges in part on whether cultural relationships are structured via personal ties or accessed more openly. Historically, even in powerful states, most people were constrained by relationship-specific obligations and privileges, often derived from kin-based institutions (Henrich, 2020, 119). Interpersonal relationships formed the basis for political organization and granted access to valuable resources and activities. Why would the beneficiaries of such privileges ever give them up by allowing wider participation? North *et al.* (2009, 25) argued they would be enticed to do so if it made their rents even greater. These advances can be brought about when elite privileges are transformed into impersonal rights, motivated largely through voluntary organizations (Henrich, 2020, 191) and open access institutions (North *et al.*, 2009, 2). Impersonality breaks the constraints of personal ties by transforming the web of interactions into a more modular architecture. Just as commodification standardizes a bundle of attributes associated with a category of good (Baldwin, 2008), ‘personafication’ standardizes a bundle of duties and obligations associated with a category of personhood. A codified social persona (e.g. citizen of England, member of the bookbinder guild) makes any member of that category interchangeable and recombinable. An established social persona grants relational freedom, reducing the complexity of new interactions instead of working strictly through vetted members of a personalized social network. Rather than learning the specifics of a stranger, you need to only learn the specifics of a persona and ascertain if the stranger is a bona fide member. This can cheapen cooperation between strangers in mutually beneficial transactions, vastly increasing the space of possible combinations and externalizing innovations to all participants (Langlois, 2002).

Similar to any modular system, creating a social persona can enable large-scale cooperation but must be worth the cost. Modularity requires agreed-upon visible design rules (Baldwin, 2008), established standards that hold fixed how members interact (a parallel to a transaction in markets). Arriving at a common description of obligations and duties is expensive. Moreover, standards vary across types of interactions (Baldwin, 2008). Tolerance in machining refers to the permissible margin of error for parts to plug in with other components (Winchester, 2018, 16). Similarly, discrepancies between expectations and performance of a persona might be intolerable without great expense. While impersonal relationships can increase at a macro level, individuals still preserve interpersonal relationships and comparatively assess value. Interpersonal relationships allow for a broader range of requests beyond a persona’s standardized description of duties and obligations (Hodgson, 2006) and contain vetted norms over search, bargain, and enforcement (Coase, 1937). In fact, barter historically only occurred between strangers while exchanges with people you knew relied on kinship customs (Graeber, 2011, 31). In short, occasionally it’s easier to ask your neighbour to watch your cat in lieu of hiring a stranger.

A tilt towards more voluntary organizations and open access institutions depends on net returns from the creation of a social persona compensating elites above the counterfactual. This unwinds

personal identity and cultural context in the structure of relationships, and in turn, lessens genealogical effects embedded in history. Of course, outcomes are not binary; they occur along a spectrum with systems being more loosely or tightly coupled (Simon, 1962) along personal relationships. Returning to the data, we can map these categories onto such a spectrum of whether a social persona is more likely to occur or elites are more likely to hold on to a rent position. Here, we can understand variations across the coarsened networks as corresponding to variations across the benefits of impersonalizing institutions. In particular, networks that scale well from creating personas generate significant returns that will make the costs of setting up a modular system worth it and pay off incumbent elites.

North *et al.* (2009, 23) identified a main feature of open access societies the impersonalizing of markets and exchange. In this vein, commerce provides the canonical example of a comedy of the commons where value increases with participation (Rose, 1986). Smith *et al.* recognized the market as a civilizing process (Muller, 2003, 56) wherein participants become ‘impartial spectators’ (Smith, 1976, 783) – that is, adopt norms promoting impersonal fairness. Further increased through voluntary organizations such as guilds, impersonality greatly expanded opportunities for cooperation between strangers and associated gains from trade. In Europe, an emerging package of norms gave rise to *lex mercatoria*, a set of guiding principles that stripped personal relationships out of exchange and encouraged individuals to engage in commerce completely separated from relational and emotional ties (Henrich, 2020, 318).

Technology, too, scales with network size. It can also face more resistance. As the Luddites realized, the benefits of new technologies often accrue to a few and disrupt the status quo (Schumpeter, 1980, 85), leading to frictions in their implementation (Kuhn, 1962, 115). For example, the printing press, electricity, and mechanical refrigeration faced heavy barriers in their implementation (Juma, 2016, 43). Scale returns feature in education as well (e.g. Romer, 1986). Its category formulation above includes universities, art institutes, and pedagogical practices. The university was a critical voluntary organization in breaking down kinship networks (Henrich, 2020, 442), and the persona of scholar was historically welcomed into royal courts. In each of these categories, competency disarmed the gatekeepers. Elites learned long ago that they must promote competent non-elites to different domains of society, incentivizing a process of persona creation since rents increased drastically in consequence (Henrich, 2020, 117).

In contrast, the social and government categories consistently showed high coefficient magnitudes and R^2 values, implying deeper roots in history. The military, language, and religion categories also appear in the upper end. First, the social category rolls up entertainment, sports, music, food and drink, and social status. Unlike commerce, it’s not obvious how some of these activities (e.g. cuisine) benefit from scaling. Compared to the value of relying on interpersonal relationships, undertaking the setup costs of a persona is discouraged. In fact, these features may even serve as deliberate constraints to network effects. For example, food taboos were probably more about ingroup social bonding than food sanitation reasons, where preventing someone from breaking bread with a stranger helped cut down on defection (Meyer-Rochow, 2009).

Other activities were likely too valuable for elites to risk giving up, again existing to deliberately preserve distinctions that prevent outsiders from joining a network. Regimes that maintain a hierarchy of personal relationships typically specialize in a range of military, political, and religious activities (North *et al.*, 2009, 18). Government’s closer tie to history might seem counterintuitive since ‘citizen’ was their canonical example of an impersonal category (North *et al.*, 2009, 2). However, the export of impersonal political institutions such as representative government has often created a misfit with local culture, leading to their poor function and replacement by relational specific forms (Henrich, 2020, 485). Moreover, Lockean or Rousseauian conceptions of society as a voluntary social contract only occurred in Europe a few hundred years ago. Most political systems relied on hereditary transfers of power until quite recently. Although royalist ambitions have waned, maintaining power structures in the hands of ‘true’ citizens is often dog whistling for power structures remaining in the hands of a historically significant ethnicity.

Finally, although we might have expected religious and linguistic categories to be strongly connected to history since they are heavily influenced by parental upbringing, each offers an established means for

generalizing participation. Foreigners can learn a new language (though may be treated differently according to accent). Religions offer conversion opportunities (at least in proselytizing religions), create anonymized descriptions of duties and obligations (e.g. Christian *caritas* prescribing fairness to Christian strangers), and form associative networks extending across territorial boundaries (e.g. Jewish trading posts along the Silk Road). There is also research on divine monitoring improving interactions between strangers (e.g. Norenzayan and Shariff, 2008). Such features pry open such institutions at least partially to outsiders, which provides a standardized description of duties and obligations for participation.

Ultimately, these results comment on frictions across political boundaries. We find a high correlation between pairwise genetic similarity and cultural similarity, meaning cultures that overlap further back in history have fewer barriers to cooperation. This paper then complements existing explanations and lays deeper footing to the role of impersonality on historical persistence whereby a modular architecture welcomes mobility between parts to undermine regional or personal lock-in effects.

Concluding remarks

Economic outcomes today depend heavily on their past even at long time horizons. Populations with similar cultures can share knowledge and communicate more easily, diffusing more complex technological and institutional innovations. But culture is challenging to define and even more challenging to measure. Other studies have used survey answers or economic experiments to correlate genetic distance with cultural distance along general attitudes towards trust, individualism, and perceptions of work and poverty (Alesina and Giuliano, 2015; Desmet *et al.*, 2011; Spolaore and Wacziarg, 2016). Instead, we develop a systemized measure from one source: the network of Wikipedia. Connections between articles on Wikipedia capture many features of cultural attention, and when aggregated across thousands of links for modern states, provide a standardized measure of cultural relatedness between country pairs. Moreover, given the size of the encyclopaedia, we can coarsen the network by categories, disentangling similarity along different dimensions such as politics, religion, language, cuisine, economics, and more. The results reinforce previous findings that genetic distance is correlated broadly and significantly with a range of differences in cultural factors. We then consider new categories of culture and find differences across regression regimes. These differences can be interpreted as variations in historical persistence, providing a glimpse towards where history matters and where it matters less. Overall, the Wikipedia data are consistent with previous results, demonstrating the encyclopaedia as a promising source for application to other phenomena, and analytical methods can be expanded to consider not only network statistics but also natural language processing of its content, too.

Additionally, these results suggest caution in the broad application of policies and institutions without recognizing long-term variables. That is, promoting certain institutions associated with economic growth will be more successful if rooted alongside careful considerations about cultural traditions (hence frequent malfunctions of representative government imposed on local cultures by foreign regimes). Policy might cross cultural differences more effectively by working to impersonalize social roles rather than flatten cultural differences broadly, which could improve the flow of innovations across nations. It is also worth pointing out that though there is historical persistence in modern outcomes, they are by no means deterministic. The R^2 values leave a large fraction of variation to be explained by other items, which reassures us against a purely Laplacian interpretation of society.

Acknowledgements. I am particularly grateful to Metin Coşgel and Thomas Miceli for numerous insightful discussions. I also thank four anonymous referees for their suggestions and improvements.

References

- Abramitzky R. (2015). Economics and the modern economic historian. *The Journal of Economic History* 75(4), 1240–1251.
- Acemoglu D. and Robinson J. A. (2012). *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. New York, NY: Crown Books.
- Acemoglu D., Johnson S. and Robinson J. A. (2001). The colonial origins of comparative development: an empirical investigation. *American Economic Review* 91, 1369–1401.

- Alesina A. and Giuliano P. (2015). Culture and institutions. *Journal of Economic Literature* **53**(4), 898–944.
- Alesina A., Devleeschauwer A., Easterly W., Kurlat S. and Wacziarg R. (2003). Fractionalization. *Journal of Economic Growth* **8**, 55–194.
- Ashraf Q. and Galor O. (2013). The ‘Out of Africa’ hypothesis, human genetic diversity, and comparative economic development. *American Economic Review* **103**(1), 1–46.
- Baldwin C. Y. (2008). Where do transactions come from? Modularity, transactions, and the boundaries of firms. *Industrial and Corporate Change* **17**(1), 155–195.
- Besley T. and Persson T. (2019). Democratic values and institutions. *American Economic Review: Insights* **1**(1), 59–76.
- Blaydes L., Grimmer J. and McQueen A. (2018). Mirrors for princes and sultans: advice on the art of governance in the medieval Christian and Islamic worlds. *Journal of Politics* **80**(4), 1150–1167.
- Cavalli-Sforza L. L., Menozzi P. and Piazza A. (1994). *The History and Geography of Human Genes*. Princeton, NJ: Princeton University Press.
- Chang H. (2010). Institutions and economic development: theory, policy, and history. *Journal of Institutional Economics* **7**(4), 473–498.
- Coase R. (1937). The nature of the firm. *Economica* **4**(16), 386–405.
- Colella F., Lalive R., Sakalli S. O. and Thoenig M. (2019). Inference with arbitrary clustering. *IZA Discussion Paper*, 12584, Institute of Labor Economics (IZA).
- Desmet K., Ortuño-Ortín I. and Wacziarg R. (2012). The political economy of linguistic cleavages. *Journal of Development Economics* **97**(2), 322–338.
- Desmet K., Le Breton M., Ortuño-Ortín I. and Weber S. (2011). The stability and breakup of nations: a quantitative analysis. *Journal of Economic Growth* **16**, 183–213.
- Duhaime E. P. (2015). Is the call to prayer a call to cooperate? A field experiment on the impact of religious salience on prosocial behavior. *Judgment and Decision Making* **10**(6), 593–596.
- Easterly W. and Levine R. (2003). Tropics, germs and crops: how endowments influence economic development. *Journal of Monetary Economics* **50**, 3–39.
- Engerman S. and Sokoloff K. L. (1997). Factor endowments, institutions, and differential growth paths among new world economies. In Haber S. (ed.), *How Latin America Fell Behind*. Stanford, CA: Stanford University Press, pp. 148–180.
- Everett J., Haque O. S. and Rand D. (2016). How good is the Samaritan, and why? An experimental investigation of the extent and nature of religious prosociality using economic games. *Social Psychological and Personality Science* **7**(3), 248–255.
- Giles J. (2005). Internet encyclopaedias go head to head. *Nature* **438**, 900–901.
- Glaeser E. L., La Porta R., Lopez-de-Silanes F. and Shleifer A. (2004). Do institutions cause growth? *Journal of Economic Growth* **9**(3), 271–303.
- Gorodnichenko Y. and Roland G. (2011). Which dimensions of culture matter for long-run growth? *American Economic Review* **101**(3), 429–498.
- Graeber D. (2011). *Debt: The First 5,000 Years*. New York, NY: Melville House Press.
- Grajzl P. and Murrell P. (2019). Toward understanding 17th century English culture: a structural topic model of Francis Bacon’s ideas. *Journal of Comparative Economics* **47**(1), 111–135.
- Greenstein S. and Feng Z. (2012). Collective intelligence and neutral point of view: The case of Wikipedia. *NBER, Working Paper* 18167.
- Guiso L., Sapienza P. and Zingales L. (2006). Does culture affect economic outcomes?. *Journal of Economic Perspectives* **20**(2), 23–48.
- Hayek F. A. (2011 [1960]). *The Constitution of Liberty: The Definitive Edition*. Chicago, IL: University of Chicago Press.
- Henrich J. (2000). Does culture matter in economic behavior? Ultimatum game bargaining among the Machiguenga of the Peruvian Amazon. *American Economic Review* **90**(4), 973–979.
- Henrich J. (2020). *The WEIRD People in the World*. New York, NY: FSG Publishing.
- Hodgson G. M. (2006). What are institutions? *Journal of Economic Issues* **40**(1), 1–25.
- Juma C. (2016). *Innovation and Its Enemies: Why People Resist New Technologies*. Oxford, UK: Oxford University Press.
- Kuhn T. (1962). *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.
- Langlois R. N. (2002). Modularity in technology and organization. *Journal of Economic Behavior & Organization* **49**(1), 19–27.
- Langlois R. N. and Garzarelli G. (2008). Of hackers and hairdressers: modularity and the organizational economics of open-source collaboration. *Industry and Innovation* **15**(2), 125–143.
- Msgari M., Okoli C., Mehdi M., Nielsen F. and Lanamki A. (2015). The sum of all human knowledge: a systematic review of scholarly research on the content of Wikipedia. *Journal of the Association for Information Science and Technology* **66**(2), 219–265.
- Meyer-Rochow V. (2009). Food taboos: their origins and purposes. *Journal of Ethnobiology and Ethnomedicine* **5**(18).
- Meyer R. (2013). 90% of Wikipedia’s editors are male – Here’s what they’re doing about it. *The Atlantic*, October 2013.
- Michalopoulos S. and Papaioannou E. (2016). The long-run effects of the scramble for Africa. *American Economic Review* **106**(7), 1802–1848.
- Michalopoulos S. and Xue M. (2021). Folklore. *Quarterly Journal of Economics* **136**(4), 1993–2046.
- Muller J. (2003). *The Mind and the Market: Capitalism in Western Thought*. New York, NY: Knopf Doubleday Publishing.

- Newman M. (2010). *Networks: An Introduction*. Oxford, UK: Oxford University Press.
- Norenzayan A. and Shariff A. (2008). The origin and evolution of religious prosociality. *Science* **322**(5898), 58–62.
- North D. (1990). *Institutions, Institutional Change, and Economic Performance*. Cambridge, UK: Cambridge University Press.
- North D. and Thomas R. P. (1973). *The Rise of the Western World: A New Economic History*. Cambridge, UK: Cambridge University Press.
- North D., Wallis J. J. and Weingast B. R. (2009). *Violence and Social Orders: A Conceptual Framework for Interpreting Recorded Human History*. Cambridge, UK: Cambridge University Press.
- Nunn N. (2020a). History as evolution. *NBER Working Paper No. w27706*.
- Nunn N. (2020b) The historical roots of economic development. *Science* **367**, 1441.
- Petiška E. and Moldan B. (2019). Indicator of quality for environmental articles on Wikipedia at the higher education level. *Journal of Information Science*, **47**(2), 269–280.
- Putterman L. and Weil D. N. (2010). Post-1500 population flows and the long-run determinants of economic growth and inequality. *Quarterly Journal of Economics* **125**(4), 1627–1682.
- Romer P. (1986). Increasing returns and long-run growth. *Journal of Political Economy* **94**(5), 1000–11037.
- Rose C. (1986). The comedy of the commons: custom, commerce, and inherently public property. *The University of Chicago Law Review* **53**(3), 711–781.
- Rubin J. (2014). Printing and protestants: an empirical test of the role of printing in the reformation. *Review of Economics and Statistics* **96**(2), 270–286.
- Rubin J. (2017). *Rulers, Religion, and Riches: Why the West Got Rich and the Middle East Did Not*. Cambridge, UK: Cambridge University Press.
- Schumpeter J. A. (1980 [1912]). *Theory of Economic Development*. New York, NY: Routledge.
- Simon H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society* **106**(6), 267–282.
- Smith A. (1976 [1776]). *An Enquiry into the Nature and Causes of the Wealth of Nations*, Glasgow edition, Oxford, UK: Oxford University Press.
- Smith D. (2020). Situating Wikipedia as a health information resource in various contexts: a scoping review. *PLoS ONE* **15**.
- Smith B. and Gutafson A. (2017). Using Wikipedia to predict election outcomes: online behavior as a predictor of voting. *Public Opinion Quarterly* **81**(3), 714–735.
- Spolaore E. and Wacziarg R. (2009). The diffusion of development. *Quarterly Journal of Economics* **124**(2), 469–529.
- Spolaore E. and Wacziarg R. (2013). How deep are the roots of economic development? *Journal of Economic Literature* **51**(2), 325–369.
- Spolaore E. and Wacziarg R. (2016). Ancestry, language and culture. In Ginsburgh V. and Weber S. (eds.), *The Palgrave Handbook of Economics and Language*. New York, NY: Palgrave Macmillan, pp. 174–211.
- Spranz R., Lenger A. and Goldschmidt N. (2012). The relation between institutional and cultural factors in economic development: the case of Indonesia. *Journal of Institutional Economics* **8**(4), 459–488.
- Steinsson S. (2023). Rule ambiguity, institutional clashes, and population loss: how Wikipedia became the last good place on the internet. *American Political Science Review* **118**(1), 1–17.
- Viégas F. B., Wattenberg M. and Dave K. (2005). Studying cooperation and conflict between authors with history flow visualizations. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* **4**, 575–582.
- Wikistats. (2021). Wikimedia statistics 2. Wikimedia Foundation, Inc. Available at <https://stats.wikimedia.org/> (accessed 22 September 2021).
- Winchester S (2018). *The Perfectionists: How Precision Engineers Created the Modern World*. New York, NY: Harper.

Appendix A: Category tags

Wikipedia employs a classification system based on contributor topic tags. Each category tag is itself a special page in Wikipedia, meaning they're also assigned categories of even broader topics. This leads to a category tree that repeatedly groups more obscure topics into more general ones, concentrating towards a general taxonomy of Wikipedia's main topics. For example, the Wikipedia article for Adam Smith lists categories (inter alia): '18th-century economists', 'Enlightenment philosophers', 'Fellows of the Royal Society of Edinburgh', 'British classical liberals', 'Capitalism', 'British male non-fiction writers', and 'People from Kirkcaldy'. In total, his article contains more than 50 tags. They range in specificity and are not necessarily mutually exclusive. Some tags relate to his geography, period, or attributes, while others fall under broad concepts or social eras. Each category leads to its own page with broader classifications. To further the example, 'Enlightenment philosophers' groups hundreds of biographies under several new categories, including 'Enlightenment philosophy', 'Early Modern philosophy', and 'People of the Age of Enlightenment'. The goal of categorization is to climb the category tree to the appropriate main topic classification, an unfortunately non-trivial task. We are interested in classifying pages under Wikipedia's main topic taxonomy (listed in footnote 6), which would classify Adam Smith under 'People', but following the first tag of 'Enlightenment Philosophers' leads to '18th Century', then 'Millennia' and 'Chronology' more broadly, ultimately ending up in 'Humanities' at the main topic level. Another problem occurs when climbing gets caught in a loop. One topic leads to another topic which circles back to the original. There is seldom a clean route to the top.

To handle this, we develop a climbing algorithm that searches vertically with a shallow horizontal search occurring simultaneously. To find the most appropriate main classification tag, the algorithm assumes the nearest topic node-wise is most relevant. When rolling up a category such as 'Enlightenment philosophers', the algorithm continuously climbs the tree vertically but searches branches along each page. It will take a few steps along 'Early Modern philosophy' and 'People of the Age of Enlightenment' to see if those are near a main topic (which in this case, the latter quickly approaches 'People'). This method repeats for all 50 or so categories for Adam Smith to generate a list of unique main topics for his page.

After performing this for all articles in the matrix, we coarsen the network into specific categories (e.g. drop all pages not related to religion). This generates a smaller network representing the connectivity of a specific topic. We generate two of these submatrices for each topic. Category tags are applied (on the html source page) in order of relevance, meaning the leading topic is generally a more accurate classification of the page than the last. To collapse the network, then, a first approach coarsens it to any page containing some category tag. These subnetworks are relatively large because nearly all pages have several main topic classifications. They also overlap. Adam Smith would appear in a coarsened network for 'People', 'Economy', 'Philosophy', 'Ethics', and 'Government'. A second approach only pulls pages by the first category tag, meaning Adam Smith is only a node in the network for 'People'. This is more precise but loses many of the connections in the network.

Appendix B: Detailed results

Table B1. Regression results for the full network from Table 4 with connectivity as the dependent variable

	(1) All	(2) Old World
Weighed F_{ST} distance	-7.208*** (0.258)	-7.434*** (0.384)
Contiguity controls		
Shares land border	0.171*** (0.041)	0.135*** (0.044)
Western Europe	0.345*** (0.016)	0.352*** (0.017)
Eastern Europe	0.359*** (0.015)	0.373*** (0.018)
North America	0.570*** (0.048)	
South America	0.256*** (0.031)	
Latin America	-0.118*** (0.034)	
Caribbean	-0.028 (0.032)	
Southern Africa	-0.122*** (0.014)	-0.175*** (0.019)
Northern Africa	0.235*** (0.020)	0.202*** (0.023)
Southeast Asia	0.292*** (0.014)	0.293*** (0.017)

(Continued)

Table B1. (Continued.)

	(1) All	(2) Old World
Southern Asia	0.315*** (0.019)	0.295*** (0.022)
Other Asia	0.273*** (0.016)	0.287*** (0.019)
Middle East	0.125*** (0.014)	0.128*** (0.016)
Geography controls		
Latitude	-0.001** (0.000)	-0.001** (0.000)
Longitude	0.000 (0.000)	-0.000** (0.000)
Geodesic	0.010*** (0.002)	0.011*** (0.002)
Water controls		
Shares water body	0.075*** (0.018)	0.069*** (0.021)
Island	-0.029** (0.011)	-0.023* (0.013)
Landlock	-0.197*** (0.009)	-0.220*** (0.012)
History controls		
Shares empire	0.013 (0.026)	0.015 (0.028)
Former colony	0.147*** (0.027)	0.186*** (0.031)
Shares language	-0.095*** (0.027)	-0.061** (0.030)
Shares linguistic family	-0.028*** (0.011)	-0.043*** (0.015)
Years independent	0.015*** (0.002)	0.016*** (0.003)
Constant	0.921*** (0.034)	0.942*** (0.039)
Observations	15,576	10,011
Centred R^2	0.31	0.323

For controls, 'Geodesic' is distance in thousands of kilometres, 'Latitude', 'Longitude', and 'Years independent' are in differences, 'Island', 'Landlock', and 'Former colony' are counts in each pair, and the remaining variables are dummies where 1 = true. Robust standard errors in parentheses.

*** $P < 0.01$, ** $P < 0.05$, * $P < 0.1$.

Table B2. Regression results for the 'All tags' submatrices from Table 5 with connectivity as the dependent variable without controls

	(1) Business	(2) Education	(3) Government	(4) Language	(5) Military	(6) Religion	(7) Social	(8) Technology
Weighed F_{ST} distance	-1.71*** (0.041)	-0.82*** (0.023)	-4.39*** (0.096)	-1.93*** (0.043)	-1.89*** (0.044)	-1.97*** (0.041)	-4.11*** (0.095)	-1.46*** (0.037)
Contiguity controls	n	n	n	n	n	n	n	n
Geography controls	n	n	n	n	n	n	n	n
Water controls	n	n	n	n	n	n	n	n
History controls	n	n	n	n	n	n	n	n
Constant	0.226*** (0.002)	0.119*** (0.001)	0.558*** (0.005)	0.238*** (0.002)	0.236*** (0.002)	0.219*** (0.002)	0.499*** (0.005)	0.186*** (0.002)
Observations	15,576	15,576	15,576	15,576	15,576	15,576	15,576	15,576
Centred R^2	0.095	0.075	0.118	0.121	0.107	0.138	0.107	0.088

Coefficients and R^2 orderings are similar to the results above.

Robust standard errors in parentheses.

*** $P < 0.01$, ** $P < 0.05$, * $P < 0.1$.

Table B3. Regression results for the ‘First Tag Only’ submatrices from Table 6 with connectivity as the dependent variable without controls

	(1) Business	(2) Education	(3) Government	(4) Language	(5) Military	(6) Religion	(7) Social	(8) Technology
Weighed F_{ST} distance	−1.66*** (0.078)	−2.88*** (0.095)	−12.62*** (0.310)	−6.39*** (0.158)	−5.50*** (0.146)	−5.18*** (0.144)	−13.57*** (0.364)	−4.85*** (0.141)
Contiguity controls	n	n	n	n	n	n	n	n
Geography controls	n	n	n	n	n	n	n	n
Water controls	n	n	n	n	n	n	n	n
History controls	n	n	n	n	n	n	n	n
Constant	0.50*** (0.003)	0.58*** (0.004)	1.901*** (0.015)	1.04*** (0.007)	0.86*** (0.007)	0.88*** (0.007)	1.88*** (0.017)	0.77*** (0.007)
Observations	15,576	15,576	15,576	15,576	15,576	15,576	15,576	15,576
Centred R^2	0.026	0.055	0.094	0.097	0.084	0.071	0.084	0.069

Coefficients and R^2 orderings are similar to the results above.

Robust standard errors in parentheses.

*** $P < 0.01$, ** $P < 0.05$, * $P < 0.1$.

Table B4. Regression results for the ‘All tags’ submatrices from [Table 5](#) with connectivity as the dependent variable

	(1) Business	(2) Education	(3) Government	(4) Language	(5) Military	(7) Religion	(6) Social	(8) Technology
Weighted F_{ST} distance	-1.247***	-0.595***	-3.083***	-1.322***	-1.197***	-1.150***	-2.799***	-1.050***
	(0.047)	(0.026)	(0.104)	(0.046)	(0.046)	(0.040)	(0.103)	(0.042)
Contiguity controls								
Shares land border	0.027***	0.016***	0.074***	0.034***	0.033***	0.032***	0.068***	0.029***
	(0.008)	(0.004)	(0.018)	(0.007)	(0.008)	(0.007)	(0.016)	(0.007)
Western Europe	0.060***	0.034***	0.139***	0.069***	0.057***	0.070***	0.132***	0.054***
	(0.003)	(0.002)	(0.006)	(0.003)	(0.003)	(0.003)	(0.006)	(0.003)
Eastern Europe	0.056***	0.029***	0.156***	0.070***	0.071***	0.068***	0.141***	0.047***
	(0.003)	(0.002)	(0.006)	(0.003)	(0.003)	(0.002)	(0.006)	(0.003)
North America	0.106***	0.057***	0.244***	0.099***	0.100***	0.094***	0.217***	0.096***
	(0.009)	(0.005)	(0.021)	(0.008)	(0.009)	(0.008)	(0.019)	(0.008)
South America	0.041***	0.021***	0.109***	0.049***	0.042***	0.041***	0.094***	0.032***
	(0.005)	(0.003)	(0.013)	(0.005)	(0.006)	(0.005)	(0.013)	(0.004)
Latin America	-0.021***	-0.009***	-0.062***	-0.023***	-0.019***	-0.033***	-0.043***	-0.021***
	(0.006)	(0.003)	(0.014)	(0.006)	(0.007)	(0.005)	(0.014)	(0.005)
Caribbean	-0.001	0.006**	0.013	-0.010*	0.002	0.014***	-0.009	0.001
	(0.005)	(0.003)	(0.013)	(0.005)	(0.006)	(0.005)	(0.013)	(0.004)
Southern Africa	-0.024***	-0.002	-0.034***	-0.019***	-0.010***	0.000	-0.075***	-0.015***
	(0.002)	(0.001)	(0.005)	(0.002)	(0.002)	(0.002)	(0.005)	(0.002)
Northern Africa	0.029***	0.027***	0.123***	0.045***	0.073***	0.077***	0.067***	0.026***
	(0.003)	(0.002)	(0.008)	(0.004)	(0.004)	(0.004)	(0.008)	(0.003)
Southeast Asia	0.058***	0.037***	0.139***	0.041***	0.057***	0.054***	0.100***	0.061***
	(0.003)	(0.001)	(0.006)	(0.002)	(0.002)	(0.002)	(0.006)	(0.002)
Southern Asia	0.050***	0.036***	0.146***	0.056***	0.054***	0.079***	0.120***	0.051***

	(0.003)	(0.002)	(0.008)	(0.003)	(0.004)	(0.004)	(0.008)	(0.003)
Other Asia	0.044***	0.016***	0.114***	0.064***	0.041***	0.066***	0.117***	0.036***
	(0.003)	(0.002)	(0.007)	(0.003)	(0.003)	(0.003)	(0.007)	(0.003)
Middle East	0.018***	0.014***	0.075***	0.015***	0.049***	0.063***	0.026***	0.013***
	(0.003)	(0.001)	(0.006)	(0.003)	(0.003)	(0.003)	(0.006)	(0.002)
Geography controls								
Latitude	0.000	0.000	-0.000***	-0.000***	-0.000***	-0.001***	-0.000**	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Longitude	-0.000*	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Geodesic	0.002***	0.000	0.004***	0.001***	0.001**	0.001**	0.002***	0.001***
	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)
Water controls								
Shares water body	0.012***	0.006***	0.024***	0.012***	0.013***	0.017***	0.024***	0.012***
	(0.003)	(0.002)	(0.007)	(0.003)	(0.003)	(0.003)	(0.007)	(0.003)
Island	-0.003	-0.008***	-0.029***	-0.002	-0.014***	-0.007***	-0.006	-0.006***
	(0.002)	(0.001)	(0.005)	(0.002)	(0.002)	(0.002)	(0.005)	(0.002)
Landlock	-0.031***	-0.019***	-0.071***	-0.040***	-0.041***	-0.032***	-0.086***	-0.027***
	(0.002)	(0.001)	(0.004)	(0.002)	(0.002)	(0.001)	(0.004)	(0.001)
History controls								
Shares empire	-0.003	0.000	-0.002	0.000	0.000	0.002	0.009	0.000
	(0.005)	(0.003)	(0.012)	(0.005)	(0.005)	(0.004)	(0.011)	(0.004)
Former colony	0.015***	0.009***	0.045***	0.024***	0.018***	0.027***	0.061***	0.012***
	(0.005)	(0.003)	(0.012)	(0.005)	(0.005)	(0.004)	(0.012)	(0.004)
Shares language	-0.011**	-0.009***	-0.036***	-0.012**	-0.020***	-0.015***	-0.037***	-0.013***
	(0.005)	(0.003)	(0.012)	(0.005)	(0.005)	(0.004)	(0.012)	(0.004)

(Continued)

Table B4. (Continued.)

	(1) Business	(2) Education	(3) Government	(4) Language	(5) Military	(7) Religion	(6) Social	(8) Technology
Shares linguistic family	-0.003 (0.002)	-0.002** (0.001)	-0.005 (0.004)	-0.004** (0.002)	-0.003 (0.002)	0.001 (0.002)	-0.013*** (0.004)	-0.004** (0.002)
Years independent	0.002*** (0.000)	0.002*** (0.000)	0.004*** (0.001)	0.002*** (0.000)	0.003*** (0.000)	0.002*** (0.000)	0.006*** (0.001)	0.003*** (0.000)
Constant	0.161*** (0.006)	0.082*** (0.003)	0.379*** (0.015)	0.162*** (0.007)	0.160*** (0.007)	0.122*** (0.006)	0.344*** (0.015)	0.131*** (0.006)
Observations	15,576	15,576	15,576	15,576	15,576	15,576	15,576	15,576
Centred R^2	0.287	0.254	0.315	0.333	0.302	0.346	0.324	0.289

For controls, 'Geodesic' is distance in thousands of kilometres, 'Latitude', 'Longitude', and 'Years independent' are in differences, 'Island', 'Landlock', and 'Former colony' are counts in each pair, and the remaining variables are dummies where 1 = true.

Robust standard errors in parentheses.

*** $P < 0.01$, ** $P < 0.05$, * $P < 0.1$.

Table B5. Regression results for the ‘First Tag Only’ submatrices from Table 6 with connectivity as the dependent variable

	(1) Business	(2) Education	(3) Government	(4) Language	(5) Military	(6) Religion	(7) Social	(8) Technology
Weighted F_{ST} distance	-1.447*** (0.095)	-2.315*** (0.113)	-9.335*** (0.350)	-4.679*** (0.178)	-3.532*** (0.159)	-3.191*** (0.155)	-8.896*** (0.401)	-3.579*** (0.162)
Contiguity controls								
Shares land border	0.055*** (0.014)	0.051*** (0.015)	0.252*** (0.059)	0.130*** (0.027)	0.103*** (0.027)	0.108*** (0.025)	0.220*** (0.061)	0.105*** (0.027)
Western Europe	0.082*** (0.005)	0.122*** (0.006)	0.441*** (0.021)	0.276*** (0.011)	0.163*** (0.010)	0.173*** (0.009)	0.403*** (0.024)	0.190*** (0.010)
Eastern Europe	0.054*** (0.005)	0.100*** (0.006)	0.459*** (0.021)	0.240*** (0.010)	0.179*** (0.009)	0.104*** (0.009)	0.385*** (0.024)	0.173*** (0.010)
North America	0.183*** (0.017)	0.237*** (0.021)	0.808*** (0.072)	0.380*** (0.030)	0.336*** (0.030)	0.336*** (0.030)	0.810*** (0.075)	0.370*** (0.032)
South America	0.076*** (0.014)	0.081*** (0.014)	0.353*** (0.043)	0.177*** (0.021)	0.117*** (0.020)	0.077*** (0.018)	0.235*** (0.048)	0.104*** (0.019)
Latin America	-0.029** (0.014)	-0.029* (0.015)	-0.217*** (0.046)	-0.096*** (0.022)	-0.066*** (0.022)	-0.062*** (0.020)	-0.115** (0.051)	-0.060*** (0.020)
Caribbean	0.050*** (0.014)	0.043*** (0.014)	0.043 (0.043)	0.074*** (0.021)	0.048** (0.020)	0.032* (0.018)	-0.068 (0.048)	0.072*** (0.019)
Southern Africa	-0.015*** (0.005)	-0.010* (0.006)	-0.028 (0.018)	0.030*** (0.009)	-0.022*** (0.008)	-0.022** (0.009)	-0.365*** (0.020)	0.003 (0.008)
Northern Africa	0.049*** (0.008)	0.058*** (0.009)	0.410*** (0.028)	0.212*** (0.015)	0.214*** (0.014)	0.222*** (0.015)	0.071** (0.029)	0.109*** (0.013)
Southeast Asia	0.105*** (0.005)	0.137*** (0.006)	0.412*** (0.019)	0.206*** (0.010)	0.199*** (0.009)	0.223*** (0.009)	0.316*** (0.022)	0.244*** (0.009)

(Continued)

Table B5. (Continued.)

	(1) Business	(2) Education	(3) Government	(4) Language	(5) Military	(6) Religion	(7) Social	(8) Technology
Southern Asia	0.067*** (0.007)	0.122*** (0.008)	0.515*** (0.027)	0.221*** (0.013)	0.226*** (0.013)	0.337*** (0.015)	0.372*** (0.029)	0.178*** (0.012)
Other Asia	0.050*** (0.005)	0.049*** (0.006)	0.265*** (0.022)	0.308*** (0.011)	0.133*** (0.010)	0.243*** (0.010)	0.298*** (0.025)	0.150*** (0.010)
Middle East	0.026*** (0.005)	0.057*** (0.006)	0.242*** (0.019)	0.147*** (0.010)	0.166*** (0.009)	0.226*** (0.010)	-0.019 (0.021)	0.075*** (0.008)
Geography controls								
Latitude	0.000* (0.000)	0.000 (0.000)	0.000 (0.000)	-0.001*** (0.000)	-0.001*** (0.000)	-0.001*** (0.000)	0.001*** (0.000)	0.000 (0.000)
Longitude	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	-0.000** (0.000)	0.000 (0.000)	0.000 (0.000)
Geodesic	0.002*** (0.001)	0.00258*** (0.001)	0.010*** (0.003)	0.005*** (0.001)	0.004*** (0.001)	0.003** (0.001)	0.003 (0.003)	0.002 (0.001)
Water controls								
Shares water body	0.018*** (0.006)	0.022*** (0.007)	0.073*** (0.024)	0.035*** (0.012)	0.031*** (0.011)	0.049*** (0.012)	0.063** (0.027)	0.026** (0.011)
Island	-0.046*** (0.004)	-0.028*** (0.005)	-0.105*** (0.015)	-0.022*** (0.008)	-0.069*** (0.007)	-0.015** (0.007)	0.001 (0.017)	-0.036*** (0.007)
Landlock	-0.073*** (0.003)	-0.085*** (0.004)	-0.229*** (0.013)	-0.148*** (0.006)	-0.139*** (0.006)	-0.130*** (0.006)	-0.275*** (0.013)	-0.100*** (0.006)
History controls								
Shares empire	-0.010 (0.010)	-0.002 (0.011)	0.001 (0.039)	-0.009 (0.020)	-0.004 (0.015)	-0.010 (0.013)	0.049 (0.045)	-0.011 (0.017)
Former colony	0.025**	0.029***	0.146***	0.114***	0.056***	0.102***	0.146***	0.028

	(0.010)	(0.011)	(0.040)	(0.021)	(0.016)	(0.015)	(0.047)	(0.017)
Shares language	-0.027***	-0.027**	-0.107***	-0.045**	-0.071***	-0.038***	-0.147***	-0.060***
	(0.010)	(0.011)	(0.040)	(0.021)	(0.016)	(0.014)	(0.047)	(0.017)
Shares linguistic family	0.001	-0.011**	-0.015	0.001	-0.016**	0.000	-0.047***	-0.019***
	(0.004)	(0.005)	(0.015)	(0.007)	(0.006)	(0.006)	(0.016)	(0.006)
Years independent	0.005***	0.006***	0.013***	0.009***	0.012***	0.006***	0.021***	0.012***
	(0.001)	(0.001)	(0.003)	(0.002)	(0.001)	(0.001)	(0.004)	(0.002)
Constant	0.430***	0.455***	1.335***	0.689***	0.640***	0.607***	1.473***	0.567***
	(0.013)	(0.014)	(0.049)	(0.025)	(0.020)	(0.020)	(0.057)	(0.022)
Observations	15,576	15,576	15,576	15,576	15,576	15,576	15,576	15,576
Centred R^2	0.168	0.205	0.268	0.251	0.263	0.277	0.291	0.219

For controls, 'Geodesic' is distance in thousands of kilometres, 'Latitude', 'Longitude', and 'Years independent' are in differences, 'Island', 'Landlock', and 'Former colony' are counts in each pair, and the remaining variables are dummies where 1 = true.

Robust standard errors in parentheses.

*** $P < 0.01$, ** $P < 0.05$, * $P < 0.1$.