Reconstructing history: Using language to estimate religious spread

Arthur Blouin
University of Toronto

Julian Dyer
University of Exeter

November 1, 2024*

We introduce a data-driven approach to use language to reconstruct history. We apply the methodology to estimate the geographic origins of religious spread. To validate the approach, we use language data to estimate origins of Islam and Buddhism that to within 500km of their true (and uncontested) origins. We then apply the methodology to the more complex (and contested) cases of Christianity, Judaism and Hinduism. We show that language-based estimates, in these cases, are significantly more aligned with the origin of scripture than to the origin of the religion.

*We acknowledge financial support from SSHRC and the Connaught fund. We thank Sascha Becker, Alberto Bisin, Andrew Oswald, Jared Rubin, Elliott Ash, and seminar participants at NYU and UBC for helpful comments. We also thank the editor Bishnupriya Gupta, and two anonymous referees for very helpful feedback, which improved the article greatly. Correspondence: Arthur Blouin, Ph.D. (corresponding author), 150 St. George Ave. room 305, Toronto, ON, Canada, M5S 3G7, a.blouin@utoronto.ca; Julian Dyer, Ph.D., room 1.21, University of Exeter Business School, Rennes Dr., Exeter, UK, EX4 4PU, j.dyer3@exeter.ac.uk

1. Introduction

Reconstructing the vast share of human history that remains unrecorded has long been a crucial, but challenging, task for historians. This task is made even more difficult when historians study contexts with incomplete survival of historical records, or from places and eras that did not keep easily interpreted records in the first place. The main approach to deal with this issue is to study archaeological evidence, which—while reliable—is costly and heavily localized. Another method of reconstructing history has been to consider the information contained in a society's language. This approach has been prevalent for centuries, having been touted at least as far back as 1765 as the one "that serve[s] best for determining the origin of peoples" (Leibniz (1996 translation), p. 285). Nearly 200 years later, using language to reconstruct history was called "one of the triumphs of nineteenth-century science" (Bloomfield (1939), p. 124).

However, while this approach remains heavily relied on today, its use is controversial, and its validity is vigorously debated. In fact, it has faced skepticism over "arbitrary and unrigorous methods" (Coleman (1988), p. 450), concerns that "semantic reconstruction lacks rigor" (Diebold (1994), p. 2909), and that it is "notoriously subject to individual interpretation" (Lehmann (1968), p. 404). That said, if a rigorous empirical approach to using semantic analysis to reconstruct history was available, it could open new opportunities for scholars to study unrecorded history.

The main goals of this article are twofold. First, to provide a proof-of-concept assessment of whether the practice of using linguistic clues to reconstruct history can be accurately applied in an objective and rigorous fashion. Second, if language can help to reconstruct history, we hope to shed some light on what parts of history language can help to identify. To accomplish these goals, we start by constructing a database with global scope that identifies *loanwords* and their source language using machine-learning techniques. Loanwords are words that, at some point in history, have been adopted from another society.[1] Using the loanwords data we construct topic-specific language-networks, and identify the most influential members of these networks.

We use the loanwords data to explore the geographic origins of the spread of the world's five major religions.[2] Religion is an apt application for our purposes because there are religious words in essentially all languages; religion is an important feature of the global landscape (Pascali (2016); Valencia Caicedo (2019); Becker and Pascali (2019); Valencia Caicedo, Dohmen, and Pondorfer (2021); Becker and Pfaff (2022)); and the potential origins of spread of each of the five major religions have been thoroughly studied.

We start by validating our methodology. To do so, we focus on the origins of Buddhism

[1] Loanwords are distinguished from cognates, which are words with common linguistic ancestry, and neologisms which are newly innovated words.

[2] These are: (1) Buddhism; (2) Hinduism; (3) Islam; (4) Judaism; (5) Christianity.

and Islam, and demonstrate that our approach can accurately estimate the geographic locations where the global spread of these religions originated. These religions have well-known and uncontested origins, allowing us to provide evidence that our methodology successfully identifies the correct locations.[3] This validation exercise suggests that loanwords do hold significant informational value. The historical account and our estimated origin of spread for Buddhism and Islam are each less than 500km away from each other (and on average about 370km away).[4] However, when linguistic information is excluded, the estimates are about 1,300km away. This suggests that methods that draw on etymology to make historical inferences are empirically valid.

After showing that language can help to trace the historical origins of religion, we apply the methodological approach to explore the more complex cases of Judaism, Christianity, and Hinduism, where there is more debate and uncertainty surrounding their origins. Much of this uncertainty stems from the fact that the global spread may have originated from canonical religious texts, or *scripture* (Rubin (2014)), rather than from the early adherents to a particular religion.[5] Accordingly, language-based estimates could reflect origin locations of words spread orally (via preaching) or in writing (via scripture). Understanding this nuance could be crucial to future applications of relying on language to reconstruct history, since these locations are often very different from each other.

The spread of Christianity, for example, could be seen as emanating from Greece, where the gospel was preached by Paul; Alexandria, where the first canonical Christian scripture was written; Constantinople, where the first Christian state was centred; or Jerusalem, where Jesus was born. For Judaism, the origin could be Jerusalem, or near Babylon, where Jews were exiled and first wrote scripture to preserve Jewish traditions. In the case of Hinduism, theories suggest an origin of the scripture in the Indus Valley or the Bactria–Margiana Archaeological Complex (BMAC) region, while the first practising Hindus are often thought to have originated from the Pontic Steppe.

Thus for each of Christianity, Judaism and Hinduism, the geographic origins of scripture are different from the origins of the religion itself, or of sacred religious figures. In each of these three cases, we find that the estimates are much nearer to the origin of the scripture than to the origin of the religion itself. This proof-of-concept evidence from religious spread suggests that methods based on linguistic change may primarily identify the *textual* or *canonical* origin of a historical phenomenon than identifying the geographic origin of the phenomenon itself.

[3]There is a large body of work using complementary applications of non-etymological forms of historical information in language to answer other questions, such as Yu and Huangfu (2019), Baledent, Hiebel, and Lejeune (2020), and Assael et al. (2022), to name a few.

[4]We calibrate our estimates using both Islam and Buddhism to avoid a mechanical estimate of either one. While the Buddhism estimate is slightly closer when we calibrate using Buddhism, and likewise for Islam, the estimates are still only 399km off on average if we rely just on the estimate of Buddhism calibrated using Islam, and the Islam estimate calibrated using Buddhism.

[5]Henceforth, we use *scripture* to reference the canonical sacred texts of any religion.

So, while language does appear to contain historically relevant information, some caution is certainly in order. As noted above, we should be careful about how to interpret language-based location estimates. Because of this, the methodological approach should not be viewed as a substitute for traditional historical analysis, nor is it suitable as such. Even beyond issues relating to interpretation and context, as one might expect in a completely automated approach that does not incorporate historical source information, the estimates are relatively noisy, and much less precise than traditional historical analysis. Accordingly, the specific implementation of the approach we investigate in this article may be less helpful for supporting traditional historical analysis when written records are plentiful than for situations where there is no historical scholarship, or where the historical scholarship that exists is heavily contested.[6] Second, we automate the entire process because it helps to "tie our hands," which from an empirical validation perspective is desirable, especially in light of the typical critiques that linguistic historical reconstruction is too "subject to individual interpretation" (Lehmann (1968), p. 404). However, there are trade-offs with this approach. For instance, it seems likely that integrating additional historical facts could help to greatly improve the accuracy of the approach, however doing so is beyond the scope of our analysis.

Our main contribution to the literature is to highlight that language can be helpful to reconstruct history when primary source data is missing. There is already a literature that aims to estimate the historical origins of various phenomena. For example, Nunn and Wantchekon (2011) demonstrate that slave trade hubs were the historical origin of mistrust in Africa. In the same vein, Lowes and Montero (2021) highlight the colonial roots of mistrust in medicine in the Democratic Republic of Congo. In these cases,[7] the object of historical reconstruction is a cultural feature of a society. In our case, we identify geographic origins of the diffusion of ideas. While we consider the case of religion as a demonstration of the approach, it seems possible that a similar approach could be used to study the spread of a variety of under-documented historical phenomena that may be of interest to economists. This includes a wide range of topics, from the spread of markets, to the diffusion of various technologies, to various cultural attributes, as in both Nunn and Wantchekon (2011) and Lowes and Montero (2021).

A second contribution to the literature relates to our construction of novel data using machine-learning methods. This approach, summarized in Abramitzky et al. (2021) and Bailey et al. (2020), has recently become more prevalent in economic history. For example, both Feigenbaum (2016) and Price et al. (2021) develop and validate the use of

[6] We believe that this approach, given that it does not require written sources, will provide the greatest benefit where such written records are unavailable. This may include applications crucial to the study of long-run economic development. This could include the emergence and spread of technology, states, and other social institutions in less-developed regions of the world, though this is beyond the scope of this paper.

[7] And many others, see for instance, Alesina and Giuliano (2015) for a review of the literature.

machine-learning methods to link individuals across administrative data sets to generate long-run historical panels. These methods, in addition to overlapping in their aim to construct better data for the purpose of research in Economic History, also are similar in methodology. Just as in our application, Feigenbaum (2016) and Price et al. (2021) rely on orthographic similarity measures in their matching algorithm. In our case, we augment this information with other linguistic features, such as phonetic similarity, which improves performance in our case, and may therefore have more general applications in the records-matching literature.

2. Historical Background

2.A. Loanwords as Historical Artefacts

This article explores whether the information contained in the etymology, or origin, of words in a society's vocabulary contains information about the evolution of important historical phenomena. This builds on the idea that words themselves contain important information about a group's past experience. The idea that there is informational content in language is not new. There is a long tradition in linguistics exploring how changes in the words a society uses are related to their history, and their evolution.

> *A community is known by the language it keeps, and its words chronicle the times. Every aspect of the life of a people is reflected in the words they use to talk about themselves and the world around them. As their world changes—through invention, discovery, revolution, evolution, or personal transformation—so does their language. Like the growth rings of a tree, our vocabulary bears witness to our past.* (Algeo (1993))

One aim of this article is to understand whether a linguistic measure of the intensity of cross-societal influence related to a given phenomenon, in our case religion, is useful for tracing the origins of these phenomena. Since we are interested in understanding the nature of cross-societal influence in the religious domain, we follow standard practice to interpret borrowed words relating to a given topic as an indicator of influence related to that topic.

Consider, for instance, the Lakhmid kingdom, which comprised parts of what is now Saudi Arabia (circa 300 - 600 C.E.), and for whom it has been notoriously difficult to reconstruct a history. Loanwords have helped to trace the roots of their formal institutions: "[...] the Lakhmids, while remaining Arab, inevitably picked up Persian influences: the prime symbol of their kingship, for example, the crown, was a Persian import, as is the loan word for it in Arabic, taj" (Mackintosh-Smith (2019)). Indeed, analyzing the etymology of certain types of words has long allowed researchers to make inferences about the introduction of certain ideas, technologies, institutions, beliefs, or cultural practices

to a particular society. In the quote, for instance, the presence of the new loanword, *crown*, indicates the source from which new ideas related to kingship had been introduced. However, it is important to note that while it is uncontroversial to interpret the presence of loanwords as—for example—evidence that the Lakhmid concept of kingship was influenced by Persian societies, this does not necessarily mean that they had no prior concept of kingship. Instead, it simply suggests that something new related to this concept had been introduced.

This example relies upon the field of etymology, which traces the history of words. Linguists define loanwords as words that have been adopted from another language group, unlike neologisms, which are invented within a given language,[8] and cognates, which are inherited from an ancestral language. Cognates have been most heavily studied by linguists, and in particular, the field of glottochronology—where differences in cognates are used to date the age of branches in linguistic family trees (Vansina (1990))—has received considerable attention.

This study of language ancestry is complicated by the possibility of horizontal transmission, which has led linguists in the field of glottochronology to try to exclude loanwords as much as possible. To do so, they compile lists of core meanings that are essentially required in all languages. These words are considered unlikely to have been borrowed, since each language would have very likely had to include some version of them prior to borrowing from another group. These Swadesh lists (first developed by Morris Swadesh) are used in many applications to identify distance between language groups (Swadesh (1950)).[9]

While glottochronology seeks to exclude horizontal language transmission, a literature on "wave-like" language evolution (originally proposed in Schmidt (1872)) stresses that horizontal transmission is a pervasive source of linguistic differences. The exclusion of horizontal borrowing has been identified as a major limitation of glottochronology—with its strictly "tree-like" models of language evolution. This critique has led researchers to consider new types of data, to allow for more complex models that incorporate cross-societal influences (Ben Hamed (2015)). Within this literature, historians and linguists regularly interpret the presence of loanwords as evidence of influence.

One notable example of this allowed historians to trace cross-societal contact between East and West dating as far back as the Parthian empire (circa 247 B.C.E. - 224 C.E.),

[8]Linguists use the term loanwords and refer to words as borrowed or loaned, even though they recognize that the lending metaphor is a poor one (e.g., words are non-rival and will obviously not be 'returned'). They do this because the terms have come to mean something very specific within the field. In fact, paradoxically, the persistence of this jargon has been attributed to the metaphor being terrible. Since nobody outside of linguistics would naturally refer to words in this way, the formal definitions have not been diluted or corrupted by laymen. We will interchangeably refer to loanwords as being *adopted* or *borrowed*.

[9]One such prominent application is the Automated Similarity Judgment Program (ASJP) (Wichmann, Holman, and Brown (2016)).

from an era in which written records are quite difficult to come by. That work concludes that "Buddhism made sizeable inroads along the principal trading arteries to the west [...] The rash of Buddhist loan words in Parthian also bears witness to the intensification of the exchange of ideas in this period" (Frankopan (2016), p. 32).

For economists, being able to directly measure the external influences on economic markets, or formal institutions, could represent an important opportunity to better understand how they evolve. One clear application of this is the work in economics on the impact of colonialism, and there are parallels in linguistics as well. Consider, for instance, the following quote about Swahili, a commonly spoken language across British-colonized East Africa: "English influence is concentrated on the semantic field Modern world, including (modern) clothing and the (modern) legal system" (Schadeberg (2009), p.87). While economists tend to exploit natural historical experiments to better understand the impact of colonialism, linguists are able to tackle the question more directly, by assessing the types of words that were borrowed from colonists. And through this complementary approach, they have been able to identify specific institutions and technologies that were particularly heavily influenced by colonists.

2.B. Religious Origins

To validate our empirical methodology requires an idea of the 'true' origin against which to compare, but it is worth keeping in mind that the notion of a single 'true' religious origin is already an oversimplification in many cases. This issue is further complicated by the fact that the global origin of a religion depends on whether we are considering largely localized oral spread through preaching by sacred figures, or global spread which predominantly took place via the creation of a canonical scripture.

In some cases, these locations are the same, or at least very similar (see section A for more detailed accounts of the various modes of early religious spread for the religions we consider). In the case of Buddhism, the preaching of the Buddha, Siddhattha Gotama, was centered in the Ganges river basin near his birthplace in Lumbini (near what is now the Nepal-India border). This region is marked in pink in figure 1, and the centroid of that region is listed in table 1, columns 3 and 4. The councils of disciples that decided the core scriptures of Buddhism were held near Rājagaha, also within the Ganges Valley. This region is depicted in green in figure 1, with centroid in columns 1 and 2 of table 1. While these regions are not identical, they are very nearby one another, and are close enough that we will have no chance to empirically distinguish between them.

Likewise, the early spread of Islam—both in terms of the preaching of Muhammad and the compilation of early manuscripts of the Qur'an—emanated from a similar place and time. Muhammad was based in Mecca and later Medina in the seventh century C.E.. We demarcate the historical Mecca and Medina Provinces with pink diagonal lines

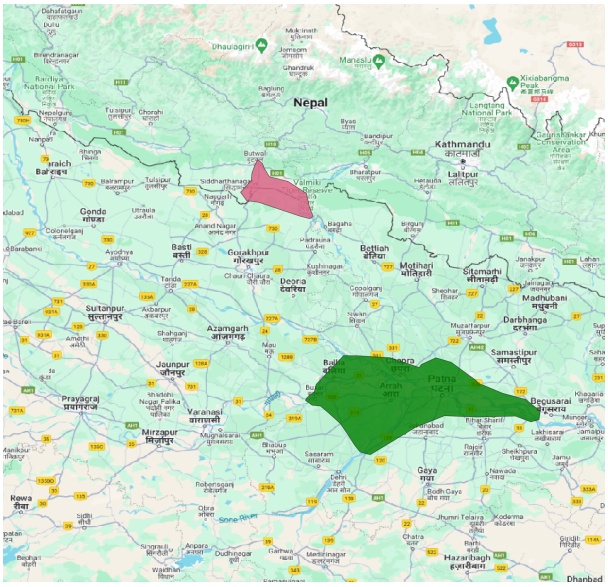


Figure 1: Map of the origins of Buddhism (pink) and its scripture (green)

Note: This map shows the historical account of the geographic origin of Buddhism itself, in pink as well as the geographic origin of Buddhist scripture, denoted in green. The centroids of these regions are listed in table 1.

in figure 2.[10] The Qu'ran, meanwhile, was collected into one volume after the death of Muhammad, by the first caliph, Abu Bakr (r. 632-634). By this time, the Rashidun caliphate comprised the majority of the Arabian peninsula, and its capatial had moved just east of Mecca and Medina, towards contemporary Riyadh (Campo (2009)). This is depicted in green in figure 2, and the eastward movement in the centroid is reflected in table 1. However, as with Buddhism, the origin of religious spread is not markedly different if we consider where Muhammed was based or where the first scripture was compiled.

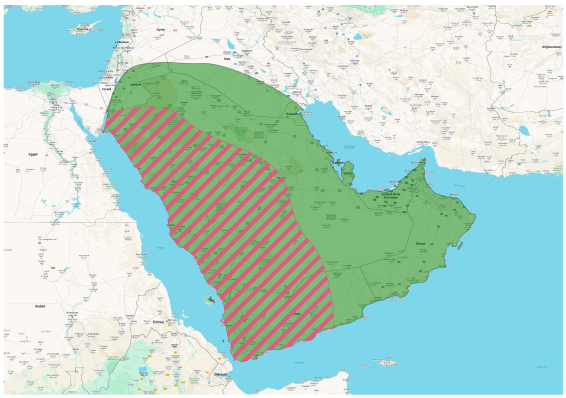


Figure 2: Map of the origins of Islam (pink lines) and its scripture (green)

Note: This map shows the historical account of the geographic origin of Islam itself, in pink as well as the geographic origin of Islamic scripture, denoted in green. The centroids of these regions are listed in table 1.

The same is not true of either Hinduism, Judaism or Christianity. In the case of

[10]These regions are based on the maps in Armstrong 2001.

Judaism, the establishment of the Kingdom of Israel and the confederation of the twelve tribes of Judaism occurred in the area west of the Jordan river near Jerusalem (denoted in pink in Figure 3). However, historians believe that canonical Jewish scripture was compiled during exile in Babylon to codify and preserve Jewish religious life and laws (denoted in green in Figure 3). In this case, the origins of religious spread via preaching and the origin of scripture would not be similar. We can see this in Table 1. Column 5 shows that while the origins of scripture and the religion itself are less than 500km away for each of Buddhism and Islam, they are over 1,000km away for each of the other three major religions.

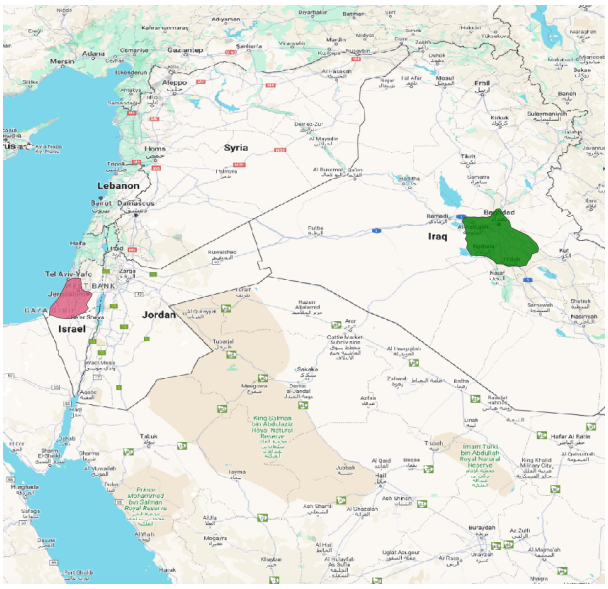


Figure 3: Map of the origins of Judaism (pink) and its scripture (green)

Note: This map shows the historical account of the geographic origin of Judaism itself, in pink as well as the geographic origin of Judaic scripture, denoted in green. The centroids of these regions are listed in table 1.

For Christianity, the origin of religious spread via preaching would have been centred on the events in the life of Jesus Christ, in and around Jerusalem (denoted in pink in Figure 4). The creation of codified Christian scripture, however, was not centred in the same region as the events depicted in the Bible. Instead, this was driven by later Greek-speaking early Christians, namely Paul, a Greek speaker from modern-day Turkey. Early Christian gospels were also written in Greek, not the Aramaic that would have been spoken by the original disciples. The bible, meanwhile was first compiled by in Alexandria, so the spread of scripture would have emanated from the historically Greek regions depicted in green in figure 4, well west of Jerusalem. Similar to Judaism, the origins of Christian preaching and the origins of Christian scripture are quite distinct.

The nature of the oral and written origins of Hinduism are less clear than the other religions we consider, which is unsurprising given it is, by far, the oldest. There is continuing debate on the origins of Hinduism that relate to the uncertainty about the origins of the Indo-European languages. While this is an incredibly complex issue, it is

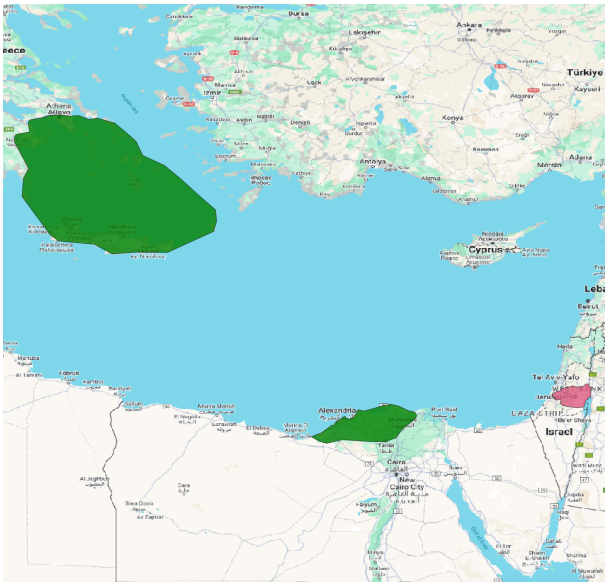


Figure 4: Map of the origins of Christianity (pink) and its scripture (green)

Note: This map shows the historical account of the geographic origin of Christianity itself, in pink as well as the geographic origin of Christian scripture, denoted in green. The centroids of these regions are listed in table 1.

notable for our purposes that given the age of Hinduism itself, its actual origins are tied to early Indo-European settlements. According to the predominant "steppe hypothesis" this traces back to Early Bronze Age migrants from the Pontic-Caspian steppe, north of modern-day Turkey. Accordingly, we denote this as the religious origin, denoted in pink in figure 5, and the centroid of that region is used for distance calculations thoughout, and reported in table 1. In terms of the origins of Hinduism's scripture, there are, broadly, two mainstream hypotheses. The first is that it originated in the Bactria–Margiana Archaeological Complex in present Afghanistan (the northern most region denoted in green in figure 5), and occurred before proto Indo-Europeans spread south to the Indus Valley. The second hypothesis is that Hindu scripture originates in the Indus Valley, and was adopted by proto Indo-Europeans after they had migrated to this region (the more southern region denoted in pink in figure 5). There is also a separate hypothesis that Hinduism originated within India, however, this has far less support among historians, and is outside of the mainstream view of scholars.

We take the centroids of each of the scripture and preaching origins for each of the five religions we consider and present them, along with the distance between these centroids, below in table1.

3. Data

The foundation of our approach is to quantify and analyze the the intensity and direction of religious language transfer among language groups.[11] To accomplish this, we build a

[11]Here we use the Ethnologue for our definition of language groups, shown on the map in figure C2.

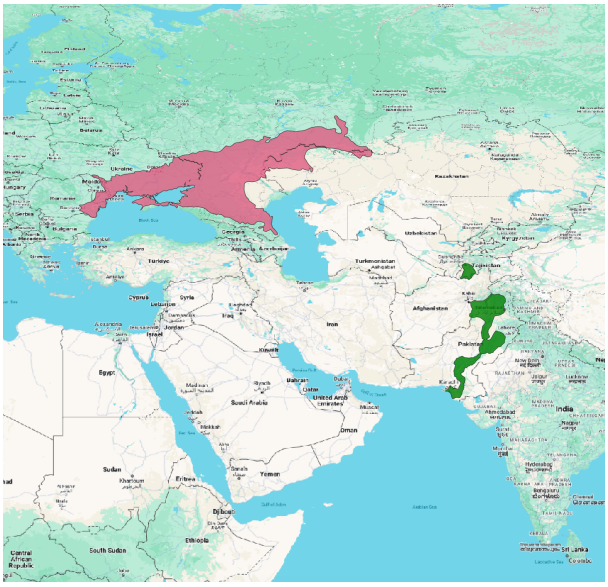


Figure 5: Map of the origins of Hinduism (pink) and its scripture (green)

Note: This map shows the historical account of the geographic origin of Hinduism itself, in pink as well as the geographic origin of Hindu scripture, denoted in green. The centroids of these regions are listed in table 1.

dataset on religious loanwords, which requires first to identify a set of religious words, and then to assess which ones were 'borrowed,' and from whom.

To do this we start by identifying words related to religion using a list of seed words based on a standard topic classification scheme. Next, we estimate which words were borrowed from other languages and identify the most likely source language. Finally, we aggregate this word-pair-level data to the language-pair level. This process is based on the methodology described in Blouin and Dyer (2022). We will outline how we identify religious words, and then describe the algorithm for identifying loanwords among these religious words.

3.A. Identifying Religious Words

To identify language transfer related to religion, we first need to identify a set of words broadly related to religion. This is a multi-step process, whereby we first identify a set of seed-words in English, then expand this set of seed words to capture all semantically similar concepts in all other languages, and then codify this set of concepts to be able to estimate loanwords. We will describe each of these steps in turn.

i) Seed-words in English The task of identifying religious words begins with a small number of seed words in English. We identified seed words by starting from the Library of Congress Classifications (LCC) system as an external, objective guide of words and concepts that represent the topic of religion. These words represent the concepts, people, and places of worship in the major religions we are aiming to represent. They were deliberately selected to cover religious concepts, without prioritizing the means of religious

Table 1: Religious Origins

	Coordinates of possible origin of religious spread				
	Origin of Scripture (centroid)		Origin of Religion (centroid)		Scripture - Religion Difference
	Latitude (1)	Longitude (2)	Latitude (3)	Longitude (4)	Distance (km) (5)
Buddhism	84.80	25.57	83.63	27.43	237.34
Islam	46.15	23.18	43.34	21.54	338.22
Hinduism	69.82	34.34	42.79	48.51	3,111.85
Judaism	44.35	32.95	34.84	31.60	1,063.72
Christianity	25.44	35.81	35.18	31.79	1,149.41

Note: This table presents the centroids of the possible origins of religious spread presented in figures fig:buddhistOrigin-5.

spread or specifically including religious texts.

Our primary motivations for using the LCC were to tie our hands and to be as transparent as possible. An alternative option would have been to compile our own list of seed words tailored to the context, but this would leave a large degree of methodological freedom to search over plausible lists until the desired result is obtained. The LCC is a reasonably objective and widely known classification system, with a relatively complete, neutral and objective set of classification categories.

We started from the LCC Subclass BL (Codes BL1-2790, Religions, Mythology, and Rationalism), summarized in table B1.[13] We then removed headings related to Mythology and Rationalism, as well as to the study or classification of religions. We also removed headings related to the history of specific religions and specific religious doctrines. We dropped any references to technical classification words such as *General*, as shown in table B2. What was left over after these removals was used as our list of seed words. We cleaned this data by replacing some of the more esoteric terms with synonyms more likely to be found in common language, or ones less likely to have non-religious connotations. In both cases, this was done to facilitate the expansion of the seed words in the next step of the process.[14] The resulting seed words, as well as the justifications for any such data cleaning, are in table 2.

[13]Original classification schema sourced from https://www.loc.gov/aba/cataloging/classification/lcco/lcco_b.pdf.

[14]Since the seed-word expansion searches Wikipedia for synonyms, it is important that (a) our seed words are common enough to appear on Wikipedia; and (b) are unambiguously religious.

ii) Expanding to other languages With these English seed words in hand, the next priority was to propagate this list across the languages in our sample. We use the English seed-words to identify related words in nearly three hundred languages from across the world, based on semantic similarity. For an overview of this process, the entire routine is presented graphically in section B.1.1 and figure B2. The intuition of this procedure is to look for similar sentence structures across languages, to see which words in these other languages are often used.[15] Doing so allows us to mitigate any bias introduced through the English seed-words.

The goal is to propagate the initial list of the seed words across each language group. The data source for language groups throughout, is the well known Ethnologue (Lewis (2009)).[16] To expand our seed-words to each of these language groups, we start from data on the words that exist in each language - the *lexicon* of the language. These lexicons come from PanLex, a single coherent lexical database built from thousands of translation dictionaries and including over twenty-five million words.[17] PanLex includes most living languages and can be directly matched to the ISO 639-3 codes used in the Ethnologue. These combined word lists include as close as is possible to all known words in all known languages. PanLex includes meaning IDs for each word, so as a first step we can match our English seed words to translations in each other language using the meaning identifier. Each of these words is converted into the International Phonetic Alphabet (IPA) using data from Ager (2019) and Mortensen, Dalmia, and Littell (2018) so we can compare words across different scripts.[18]

However, if we stopped at direct translations we would risk the list of religious words capturing a large western bias. So, it was important to identify religious concepts in each of these languages, as they are typically used in those languages, rather than being restricted only to direct translations of the English seed words. To do this, we implemented a well-established semantic analysis routine trained on Wikipedia data (see Bojanowski et al. (2017)) for two hundred ninety-four languages.

The logic is, for each language, to represent words numerically in a way that captures the meanings of words and how they are associated with each other. The similarity in the

[15]For instance for place of worship, we might find 'Temple' in some languages or 'Mosque' in others, which are not direct translations of each other

[16]The goal is to identify religious words in all languages. Throughout the study, a language group, as defined by the digitized Ethnologue map of ethnolinguistic societies, is the unit of observation. The Ethnologue provides the locations of each language, and it includes both contemporary languages as well as recently extinct and vulnerable languages. In the Ethnologue, borders for each group are provided, which allows us to compute the centroid of each group.

[17]PanLex is a non-profit with the mission of improving resources available to underserved languages. To do this, they have attempted to build the largest possible lexical translation database. See https://panlex.org. The database is constantly being updated to include new sources and for our analysis we the dataset as it was on October 1, 2018.

[18]For further information on how we filtered out phrases and expressions that are not words, see section B.1.

context in which words are used allows us to compute the 'distance' between two words.[19] To do this we represent words as vector values in a 300-dimensional vector space, where each of these dimensions is intuitively related to a 'feature' that captures the relationship between two words. For example, the word 'Queen' can be represented as being quite similar to the representation 'King - Man + Woman' (Mikolov, Yih, and Zweig (2013)).

After finding direct translations of the seed words (i.e., those assigned identical meaning identifiers in PanLex) among the covered languages in PanLex, we use this routine to identify words that are not direct translations, but are similar. To consider a broad range of associations, we consider two meanings to be similar if their word-vector representations are similar to a seed word or its direct translation in any of the covered languages. This means that even if a concept is not closely related to religion in English, but is semantically similar in another language, we are able to include this association in our list identification of religious words. Therefore, the concepts we identify as related to the initial seed words are not purely based on English worldviews. We take these 'similar meanings' and again translate the expanded word set using the PanLex meaning IDs, to get a large list of words in each language that are related to religious seed words.[20]

There are several important advantages to this method. The first is that it allows for broader coverage. Some of the languages in PanLex have more coverage than others, and expanding the set of words that we examine increases the odds that one or more of them is included in the less heavily documented languages. Second, it is important not to narrow in too closely on the loanwords data. Our intention was to develop a way to examine *global* patterns in language transmission. Rather than getting into the process of defending the loanword status of specific word pairs - which is the focus of linguists[21] - our approach is to acknowledge that any automated approach will come with error, and we should accordingly manage that error to the best of our ability. One way of doing this is by exploring averages of larger sub-samples, whenever possible. Finally, the procedure aims to minimize the likelihood that - despite the relatively objective nature of the LCC - our identification of religious words is driven by word associations in English, and hence reflects solely Western worldviews.

Once we have identified all similar words in all languages in the Ethnologue, both the original English seed-words and the much larger set of semantically similar words in the other languages are all matched to the meaning IDs described above. This comprises our final list of religious words.

[19] This has been used in economics as a way to measure worldviews and cultural discourse (Giorcelli, Lacetera, and Marinoni (2022)).

[20] This produces a list of over 8,000 meanings that are associated with our original English seed words. The vast majority do not have direct English equivalent, but we present in table B3 the English words associated with these additional meanings.

[21] We view our approach as complementary to the work that linguists do. It is certainly not a substitute, since we cannot claim with anywhere near the same level of certainty that any particular word pair is, or is not, a loanword pair.

3.B. Machine Learning Algorithm: Identifying Loanwords

Having algorithmically identified a set of religious words across the world's languages, the next step is to identify which of these words were borrowed from other languages, and to identify the source language. While Panlex is a near-complete list of words in the world's languages, it does not contain the necessary information on borrowing. To generate this data, we use a standard machine learning algorithm to predict loanword status, and identify the most likely source. Our approach was to automate the procedure used by linguists to identify loanwords as closely as possible. To this end, we follow the discussion of this process in the section *Recognizing Loanwords* from the authoritative guidebook *Loanwords in the World's Languages: A Comparative Handbook* (Haspelmath and Tadmor (2009)). To the extent that is possible, we aimed to create computational analogues based on Haspelmath and Tadmor 2009, to generate features in our data set that approximate the features that linguists typically consider.

However, to do this, we needed a validated set of loanwords we could use to train the classifier. This data does exist, in the form of the World Loanword Database (WoLD), which is the largest data-set of consistently compiled loanwords identified by linguistic experts. To be more precise, WoLD includes "vocabularies (mini-dictionaries of about 1000-2000 entries) of 41 languages from around the world, with comprehensive information about the loanword status of each word" and identifies the source words for these borrowings from three hundred sixty-nine other languages. We used this data set to train our machine learning algorithm on the word-pairs in PanLex that can be matched to WoLD. We then applied the classifier to all of the word-pairs in PanLex that are potential loanwords.

To do this, we started by creating a word-pair level database of words that are semantically similar and thus may have been transferred from one language to another. An overview of the process, and the databases and tools used at each stage, is presented in figure B1. To build the training set, we drew a stratified sample from the subset of PanLex word-pairs that are also included in WoLD.[22] We had to address the fact that the training set is heavily imbalanced, with many fewer true loanword word-pairs than non-loanword word-pairs. This poses a problem, because it could result in high accuracy by drastically under-estimating loanwords. We dealt with this by selecting only a random sub-sample of the heavily over-represented categories, and then augmenting the under-represented categories with synthetic oversampling (Chawla et al. (2002); Lemaitre, Nogueira, and Aridas (2017)).[23] Based on this training set, we predicted loanword status using a ran-

[22]This stratified sample included some word-pairs that were actual loanwords, and different types of non-loanword word pairs including: non-borrowed words, borrowed words but matched to the wrong source word, and borrowed words but where the direction of borrowing is inverted.

[23]We implemented the classification procedure in two stages, with a coarse first-pass to remove obvious non-loanwords, and a second-stage refined classifier that focused on the less-obvious cases, such as cognates v. loanwords or loanwords with the direction of transfer being inverted. We then applied this

dom forest classifier. Estimation details are all in appendix section B.1.2. Overall, the accuracy of the classifier was approximately 98%.[24]

After training the classifier, we applied it to the full set of potential loanword word-pairs in PanLex, selecting the highest-probability source word for each.[25] We then restricted to the set of words identified as religious words (as described in section B.1.1) and constructed measures of intensity of religious borrowing between language pairs. This aggregated variable represents language adoption by group i from group j, and is defined as follows:

$$\mathcal{L}_{ij} = \frac{\#ReligiousLoanword_{ij}}{\#ReligiousWord_i} \tag{1}$$

We define $\#ReligiousWord_i$ as the number of religious words in the language of society i. Similarly, $\#ReligiousLoanword_{ij}$ is the number of religious loanwords in the language of society i originating from j. $\mathcal{L}_{ij}$ is therefore the share of religious words in society i that were adopted from society j, or equivalently, a measure of the religious linguistic influence of j over i. It is worth noting that $\mathcal{L}_{ji}$ is a separate observation indicating religious linguistic influence in the opposite direction, of group i over j.

Summary statistics are in table 3. They show that conditional on there being any language adoption, borrowing between a typical language-pair accounts for approximately 3% of religious words. We use this pairwise data to construct a directed network of religious language transfer among Ethnologue groups. Details of how we construct the networks and associated measures of network centrality are in appendix section B.2.

4. Empirical Methodology

Our empirical approach is inspired by Barjamovic et al. (2019), who collect exceptionally rich historical data on inter-city trade flows to reconstruct the probable locations of 'lost' ancient cities. In many cases, collecting such data is not feasible or even possible. One insight of this article is to show that data on language can help for geolocation as well, albeit for slightly different purposes. However, accommodating this broader range of settings introduces various challenges that require non-trivial adaptations of the

classifier to a much larger sub-sample and trained a second more refined classifier on those identified as plausible potential loanwords by the first classifier.

[24]The vast majority of potential loanword word-pairs were rejected by the first-stage coarse classifier. The refined classifier was approximately 92% accurate on the less-obvious cases that were not rejected in the first pass and made it to the refined second-stage classifier. We present further details on classifier performance in the confusion matrix in figure B3.

[25]Please see appendix B.1 for further details on the classification procedure and the features used at each stage.

methodology, so the two approaches should be considered complementary.[26]

4.A. Calibration

The empirical exercise begins with the constructed measure of influence (or centrality) in the network of adoption of religious words based on the loanwords data described above (again, see appendix B.2). Using this measure, we estimate the relationship between religious language influence and distance to a *known* origin of spread. We then use this information to make inferences about the geographic locations of *unknown* origins of spread.

For the purpose of validation, this means that we would like to understand, for each of Islam and Buddhism - the two religions with clear and uncontested origins - if we can use what we know about one to estimate the location of the other. For the purpose of better understanding what the methodology is capturing, we use calibrations from both Buddhism and Islam, to estimate whether the resulting estimates for each of Christianity, Jusaism and Hinduism are nearer to the origins of the religions themselves, or to the origins of the scripture. Across both exercises the results using either Buddhism or Islam to calibrate are not materially different.

Starting with the validation exercise, we calibrate first, using Islam, and use this information to estimate the location of Buddhism, and then, calibrate using Buddhism, and estimate the location of Islam. To generate the estimates used for calibration, we proceeded with the regression model in equation 2. Throughout this paper we refer to language influencers, the group that is the source of loanwords, and language adopters, the group who adopts the loanword from another language. As before, we denote this using subscript i to indicate a language in its role as an adopter, and j to indicate a language as an influencer.[27]

$$log(d_j) = \beta \mathbf{c} + \gamma LexiconSize_i + f(DistanceBetweenGroups_{ij}) + \epsilon_{ij} \tag{2}$$

In equation 2, $\mathbf{c}$ is a matrix containing some polynomial of c_j, which is a measure of linguistic influence. We consider a cubic specification in the main results, but all results are consistent using linear and quadratic specifications as well, and estimates from these models are presented in the appendix throughout.[28] c_j measures influence within

[26] Barjamovic et al. (2019) estimate a gravity trade model with commercial records from 12,000 clay-tablets dating back to 19thC BCE, which required an understanding of an Old Assyrian dialect of ancient Akkadian. Without this information, we rely on unsupervised machine learning to separate estimated source points into clusters corresponding to specific religions.

[27] Given that our data is at the directional pair level, each language will appear both as lender and borrower.

[28] c_j is included as a cubic polynomial in the main specification in order to account for the expected pattern of non-linearities in the relationship between distance, lending and borrowing. For instance, we

a directed network of religious word spread. For the main results we use eigenvector centrality, which is defined formally in equation 9 in appendix B.2. Again though, results are robust to using alternate measures of network influence, which are also presented in the appendix throughout. $f(DistanceBetweenGroups_{ij})$ is the distance between the influencing and adopting language groups. We control for the size of the lexicon included in the source data ($LexiconSize_i$) to account for the possibility that centrality is artificially low when data is more sparse.[29] $log(d_j)$ is the natural logarithm of the distance from the centroid of language group j to either Mecca or Lumbini, and results are all robust to modelling this linearly as well.

We do the same for adopters in the network (i.e. those being influenced). In this case we have a regression equation as follows:

$$log(d_i) = \beta \mathbf{c} + \gamma LexiconSize_j + f(DistanceBetweenGroups_{ij}) + \epsilon_{ij} \tag{3}$$

Everything is defined as before, but the subscripts are swapped. In this case, because the focus is on adopters, the matrix $\mathbf{c}$ contains elements c_i to measure a language group's propensity for adoption within the network of religious word spread. An observation is a language pair ij. Of course, for all i or j in these regressions, both the network centrality and the distance to Mecca / Lumbini only vary at the group-level, and not the group-pair level.[30] This has implications for the standard errors, so to account for this they are two-way clustered by groups i and j.

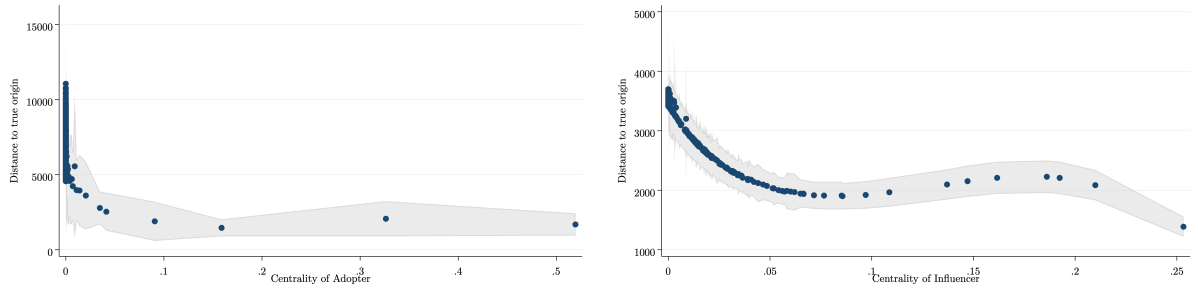
The resulting estimates are in table 4. We show estimates using Buddhism in columns 1 and 2, and using Islam in columns 3 and 4. Importantly, across all specifications we see significant non-linearities, which partly justifies the non-linear specifications in equations 2 and 3. Again though, estimates are robust to alternative specifications as

expect that very nearby an origin is likely to almost exclusively lend, and therefore have high out-group centrality, but we expected that this may likely to trail off quickly, and those beyond even relatively small radii from the origins (relative to the study region) may almost exclusively borrow. Beyond this, borrowing too would dwindle as religious influence decreases with distance to the given origin. We also wanted to keep the specification consistent for both borrowers and lenders, and felt that including a more flexible specification would make that more sensible.

[29] As described in appendix B.1 our borrowing/lending data is based on the wordlists in the PanLex lexicon for each language, from which we calculate $LexiconSize_i$ (the number of single-word expressions) to control for data availability. We discuss the potential bias from the sources used to construct our data in B.6.

[30] Another valid option would have been to aggregate the data to the group level prior to running the regressions instead of after. The two options are essentially equivalent. But the next step of converting the predicted distances from these regressions to origin co-ordinates necessarily takes place at the pair level. So, in this case we would have to aggregate the data for this step, dis-aggregate for the next step, and then re-aggregate again after that, which seemed unnecessarily complicated. However, the clear trade-off is that in this case we have a group-pair data-set with primarily group level variation. There are the same number of observations for each group in our 'stacked' data-structure (i.e. all observations are equally weighted regardless), so the only implication is for the standard errors.

well.[31]



(a) Relationship between network centrality and distance for borrower

(b) Relationship between network centrality and distance for lender

Figure 6: Scatterplot of relationship between Distance to Origin and Network Centrality

Note: The figure displays binned scatterplots to show the relationship between the network centrality measures for each language group, and their distance from the religious origin. The plots are constructed based on 1500 bins in each case.

Overall, as one might expect, those that are influential within the religious network for Buddhism are nearer to the origins. This can be seen in figure 6, which presents the scatterplot between network centrality and distance to origin for both lenders and borrowers. The graphs for each show the heavily non-linear relationship that implies that by far most linguistic exchange takes place nearby the religious origins or either the scripture or the religion.

4.B. Solving for the Origins of Religious Spread: Euclidian formula

The next step to solving for the origins of religious spread is to use the estimates from equations 2 and 3 - which are shown in table 4 - to compute the predicted distance to the origin for each observation. Intuitively, this represents a weighted average over religions, so that, for example, heavy Buddhist influence 'pulls' the predicted origin to the east, and heavy Islamic influence 'pulls' it to the west. Each predicted distance value minimizes the error from equations 2 and 3.

It is simple to compute these predicted distances for for each influencer and adopter in the data (i.e. using columns 1 and 2 of table 4, in the case of Buddhism), however what we are interested in is geographic coordinates, not distances. These origin coordinates are relatively straightforward to derive from the distances. For each language-pair in the data, the distance represents the radius of a circle emanating from their own language group geographic centroid. Along the circle formed by this radius lies the estimated religious origin centroid described above. To convert our radii into a latitude and longitude of this origin centroid, we solve for the geographic coordinates that best rationalizes the two circles (i.e. one estimated for group i and the other for group j, of pair ij). Given our distance estimates ($\hat{d}$) from equations (2) and (3), these radii are already estimated. The

[31] We show plots of actual and estimated distance to Mecca for lenders and borrowers in figure C5 and show the relationship exhibits the expected pattern.

associated coordinates are directly implied by the Euclidian distance formulas. These are:

$$\hat{d}_j = (10000/90)\sqrt{(\phi_j - \phi_o)^2 + (cos(\frac{37.9\pi}{180})^2(\lambda_j - \lambda_o))^2} \tag{4}$$

and

$$\hat{d}_i = (10000/90)\sqrt{(\phi_i - \phi_o)^2 + (cos(\frac{37.9\pi}{180})^2(\lambda_i - \lambda_o))^2} \tag{5}$$

In these equations $\hat{d}_j$ and $\hat{d}_i$ are the predicted distances based on table 4. ϕ represents longitude, so that ϕ_j is the longitude of the influencing group (which is known from the Ethnologue), and ϕ_i is the longitude of the adopting group (also known from the Ethnologue). ϕ_o is the longitude of the origin, which is what we would like to solve for. Likewise, λ represents latitude for either group i or j (both known from the Ethnologue), or origin o (which we aim solve for).

Equations 4 and 5 therefore represent a system of two equations and two unknowns. The two unknowns are the latitude and longitude of the origin $\{\phi_o, \lambda_o\}$. The solution would be trivial if the radii intersected at only a single point (i.e. they were always exactly tangential) since there would be a unique analytical solution. But of course this is not always the case, due to measurement error in each of $\{\phi_i, \lambda_i\}$, $\{\phi_j, \lambda_j\}$, and c_i and c_j. Accordingly, we solve numerically for the latitude and longitude that best fits this system using the non-linear estimation procedure outlined in Ross (1990).[32]

This provides us an estimate of the coordinates of the centre of religious influence for each language pair. The estimation procedure converts radii into coordinates, but these coordinates have a similar interpretation to the predicted distance measures we described above. In other words, conceptually, neither $\hat{d}_j$ and $\hat{d}_i$, nor the associated implied coordinates, identify any particular religious origin. Instead they identify a centroid of origins. Intuitively this means that if a language was equally influenced by only Islam and Buddhism, both $\hat{d}$ and the associated $\{\phi_o, \lambda_o\}$ would represent a convex combination of each origin - which may be far away from both. The more that influence or adoption is confined to a single religion, the closer these distances will get to a true religious origin. Even if influence / adoption within a language pair is mostly skewed towards a single religion, we will end up with clusters of coordinates near the religious origins, rather than the goal of a single point-estimate.

To resolve this issue, we aggregate the estimated coordinates using k-means clustering.

[32]For computational efficiency we implemented this with a 10% random sample of the data, which took about 3 days.

We use several other aggregation methods as well, and these produce similar results, they are shown in the appendix throughout. We specify that there should be five origins of spread corresponding to the five global religions (details are in appendix B.3).[33] Figure C6 shows the efficacy of the k-means clustering routine when we specify a number of clusters different from five. That analysis suggests that specifying five clusters performs best, as it features the lowest rates of mis-assignment of observations to clusters.[34] This implies that even if we had not known to look for five religious origins, and instead used an algorithm to search for the optimal number of clusters, we would have arrived at the same set of five estimates. In addition to this, one of the robustness checks we use is to aggregate using Ward clustering, which is computationally demanding, but does not require a prespecified number of clusters. This method also produces 5 centroids associated with the 5 major religions (figure C20b). In all cases, the mean coordinates within each cluster produce five sets of latitude-longitude pairs that correspond to the origins of religious spread for each of the five religions that we are interested in.

To benchmark these estimates for the purpose of validation - the exercise using Islam and Buddhism - we follow the exact same procedure outlined above, but we replace the language network data with a random number on the same scale.[35] This procedure helps to ensure that we do not accidentally induce a mechanical relationship either through the clustering routine, or the choice of study region. If the loanwords-based estimates are systematically closer to the historical account than this benchmark, this can be interpreted as evidence that there is historically relevant information encoded within a society's language.

For the empirical test using Christianity, Judaism and Hinduism, we employ a similar framework, however, we compare the estimated distances to the origins of scripture to the origins of the religion itself. These are essentially the same for Islam and Buddhism, so this exercise is not possible for those (see Table 2). Likewise, validation using Christianity, Judaism and Hinduism is not possible since the origins of spread for those three religions are not straightforward.

We proceed first with the validation exercise, and then move onto trying to understand whether the language-estimates are capturing the origins of religions themselves, or the origins of scripture.

[33] We also use alternate clustering methods, as described in Section B.4.

[34] An observation is defined as mis-assigned in the conventional way - when the clustering algorithm assigns it to a cluster that it is not nearest to.

[35] The clustering algorithm is restricted to latitudes between 17.5 and 42.5; and longitudes between 20 and 95. This is to avoid the confounding effects of religions for which we are not trying to pinpoint an origin.

5. Validation: Is there informational content embedded in language?

5.A. Empirical test

We are interested in two main empirical validation exercises. The first is to compare the location taken from historian accounts of the origins of religious spread to the model estimate for the same location (and associated confidence regions). For this comparison, if our model is valid, we expect to be unable to reject the null-hypothesis that these locations are the same. The second exercise is to compare the model estimates that rely on language information to the benchmark estimates that do not. In this case if we can reject the null hypothesis that the distances of each to the historian accounts of the origins are the same, then we can conclude that there is relevant information contained in language. Both exercises are important to validating the practice of inferring history from etymology.

Our aim is to see if a purely data-driven approach will correctly fail to reject these reasonable hypotheses.[36] Since a core element of our empirical approach is the *failure* to reject the null, we follow Barjamovic et al. (2019) by reporting confidence areas that are much tighter than the standard 95%. In this case, we simply follow Barjamovic et al. (2019) and report 75% confidence areas, which makes the region much smaller, and therefore makes it more likely that whenever we do fail to reject the null, that we do so because the estimates are indeed quite similar, and not due to noisy estimates.

Another implication of the empirical approach is that we must accept some error. There are a few obvious sources of error, and likely more. One example is that the estimated origins are based on language group regions. The measured locations of these language groups are centroids of geographic polygons and not population hubs,[37] so we should expect this to introduce some measurement error. A second example is that we are unable to observe language dynamics over time. Again, each of these sources of error reduce the precision of the results, but would only represent a source of bias if our inability to account for them systematically moved the estimates nearer to the mainstream hypotheses of religious origins in the history literature. It is difficult to see how this would be the case.

5.B. Results

For the validation exercise, we focus on estimates of Buddhism and Islam. In our examination of Buddhism, we rely on the calibration exercise using Islam, and vice-versa in our examination of Islam. This is to avoid a mechanically precise estimate of a location

[36] Admittedly, what it can reject may often be more interesting, and we will discuss some of this as well. But, again, our main goal is validation, and getting close to well established hypotheses is arguably the most convincing way to do this.

[37] i.e. the centroid of the language polygon could be a location where nobody lives

based on its own calibration. A map of the results can be seen in figure C7.

The origins of Buddhism, and its spread, are uncontested historically. It began in Lumbini in Nepal and spread geographically from Rājagaha near the India-Nepal border, where Buddhist scripture was first compiled (appendix A.1). The map in figure C7 displays the historiography-based origin of the religion and scripture in pink and green respectively, and the estimated 75% confidence area with a circle - computed as in Barjamovic et al. (2019). For Buddhism, the confidence area completely overlaps with the areas of historical consensus, indicating that the estimated locations are not significantly different from the true locations.[38][39] In table 5 we present the estimated distances to the 'actual' origins based on the history literature. We see the estimates for Buddhism in columns 1 and 2. When we estimate the origins of Buddhism by calibrating with Buddhism (column 1) we estimate a difference in only 393km, however this may obviously be a mechanical relationship. Indeed, in column 2, where we estimate the origins of Buddhism calibrated using Islam, the estimate is further away, but only slightly. In this case the estimate remains only 405km away from the true origin. This is much closer than the comparable estimate that excludes linguistic information (nearly 1,400km away). Using the more reasonable Islam-based calibration, the language-based estimate is more than three times closer to the true origin, a difference that is significant well beyond the 1-percent level.

Second is Islam. The origin of the religion itself is the portion of the Arabian peninsula under the rule of Muhammad at the time of his death, while we take the area under the rule of Abu Bakr as the region of origin of the written scripture. The maps in figure C7, just as with Buddhism, show an almost complete overlap between our estimated regions and the histiography-based consensus regions. This implies that there is no significant difference between our estimates, and the historical account.[40][41] Furthermore, in columns 3 and 4 of table 5 we present the precise distances between our estimates, and the historical concensus. These estimates for Islam paint a very similar picture to the

[38] We present a series of robustness checks in the appendix figure C8 as well. In figure C8a we show that the estimated coordinate is in essentially the same place when we calibrate using a linear specification. In figure C8b we show robustness to a quadratic specification. In figure C8c we calibrate using a linear dependent variable instead of the log-dependent variable. In figure C8d we calibrate using Betweenness Centrality instead of Eigenvector Centrality, while in figure C8e we examine Degree Centrality instead of Eigenvector Centrality. In all cases, the estimated origin location is essentially unchanged.

[39] In figure C10 we also demonstrate robustness to various alternate clustering algorithms. Again, the estimated origin location is essentially unchanged.

[40] We present a series of robustness checks in the appendix figure C9 as well. In figure C9a we show that the estimated coordinate is in essentially the same place when we calibrate using a linear specification. In figure C9b we show robustness to a quadratic specification. In figure C9c we calibrate using a linear dependent variable instead of the log-dependent variable. In figure C9d we calibrate using Betweenness Centrality instead of Eigenvector Centrality, while in figure C9e we examine Degree Centrality instead of Eigenvector Centrality. In all cases, the estimated origin location is very nearby the original estimate and not significantly different from it.

[41] In figure C11 we also demonstrate robustness to various alternate clustering algorithms. Again, the estimated origin location is essentially unchanged.

Buddhism estimates. We show the calibration based on Buddhism in column 3 and the one based on Islam in columns 4. As before, we include both for completeness, but there is something mechanical about estimating Islam's origins using Islam-based calibrations. This is reflected in the estimated distance, just as with Buddhism, so we focus on the larger column 3 estimate, which presents the estimate of the origin of Islam, calibrated using Buddhism. This estimate is actually very similar to the Buddhism estimates we saw in columns 1 and 2, and off from the true origin by only 392km. In contrast, the estimate based on an identical procedure with the exception that we omit information on language leads to an analogous distance of over 850km. The difference between these estimates is significantly different from 0 well beyond the 1% level.

Overall, both the Buddhism and Islam estimates are very nearby the historical account, and in both cases the estimates can be statistically assessed as more informative than estimates lacking any linguistic information. This implies that there is historically relevant information embedded in language, and this information can be leveraged to make inferences about history when records are lacking.

6. Application: Is global spread driven by religious figures or scripture?

While there appears to be important information embedded within a society's language, what that information reflects remains unclear. This question is crucial since while we are able to accurately estimate religious origins in the two most straightforward cases, Islam and Buddhism, we still should acknowledge that there has historically been divergent conclusions based on linguistic and archaeological evidence. So far, we have no way of providing insight into whether whether these discrepancies are due to inherent bias in the analysis of linguistic data (Coleman (1988); Diebold (1994); Lehmann (1968)), or because the two approaches are inherently measuring the origins of different phenomena.

The intuition behind the latter possibility is that the linguistic approach focuses on *spread*, while archaeological evidence identifies the presence of the societies themselves. These may be the same locations, but may not be. It seems possible that any loanwords-based approach is more likely to estimate the origin of this spread rather than the origin of the religion itself. In the case of religion, these locations happen to be very nearby in the two cases we have looked at so far, but this is not the case for Judaism, Christianity or Hinduism. Accordingly, we now apply our approach to the origins for these three religions, with an eye towards whether they are picking up the origin of scripture, or of the religion itself.

Starting with Judaism, it is widely agreed that the religion itself developed in Jerusalem. However scripture was either conceived and written (in the case of the Talmud) or codified (in the case of the Torah) near Babylon - the capital of ancient Babylonia. Babylon

is where Jewish aristocracy were exiled by Nebuchadnezzar (appendix A.4), and was a central hub of Jewish life for over 1,000 years since. These locations are denoted in figure C12, where the region around Babylon is in green, and Jerusalem is in pink.

The distance between our estimated location for Judaism, and the locations of both Babylon and Jerusalem can be seen in table 6, columns 1 and 2. In column 1 we present the distances calibrated using Buddhism, and in column 2 they are based on the Islam estimates. The distances are consistently quite close to each other (within about 200km), which reflects that the coordinates estimated using each are quite similar (figure C12). In both cases the estimates are much closer to ancient Babylon than to Jerusalem. When we calibrate with Buddhism, in column 1, the distance to Babylon is more than 4 times closer to our estimate than the distance to Jerusalem, and more than twice as near when we calibrate using Islam. In both cases, therefore the estimates favour the origin of the scripture over the origin of the religion itself. The difference between these two distance estimates is statistically significant well beyond the 1% level. That said, the difference between the history-literature consensus origin of scripture and our estimate is not significantly different. While our 75% confidence areas are much larger than the true regions, the two areas completely overlap with both calibrations (figure C12). This is not the case for the origin of the religion itself, where there is no overlap at all when we calibrate with Buddhism (figure C12a), and only partial overlap when we calibrate with Islam (figure C12b).[42][43]

Next we turn to Christianity, which presents a similar dilemma to Judaism. Did Christianity primarily spread from Jerusalem, where Jesus lived? From Constantinople (i.e. Istanbul) where Christianity was institutionalized? Or from north Africa, where the New Testament was written and canonized (appendix A.5)? Christianity is even slightly more difficult to deal with than Judaism because even if we only consider the origin of scripture, it is not entirely clear what the appropriate origin location should be. For instance, Alexandria appears to be a reasonable choice, as the location where the New Testament was compiled. But equally reasonable could be Greece, where Paul proselytized, and wrote the majority of the early chapters of the New Testament. Because of this, in figure C15 we represent the region surrounding the Mediterranean in green to represent the origin of scripture, while we denote Jerusalem, the origin of the religion, in pink.

[42]We present a series of robustness checks in the appendix figure C13 as well. In figure C13a we show that the estimated coordinate is in essentially the same place when we calibrate using a linear specification. In figure C13b we show robustness to a quadratic specification. In figure C13c we calibrate using a linear dependent variable instead of the log-dependent variable. In figure C13d we calibrate using Betweenness Centrality instead of Eigenvector Centrality, while in figure C13e we examine Degree Centrality instead of Eigenvector Centrality. In all cases, the estimated origin location is essentially unchanged.

[43]In figure C14 we also demonstrate robustness to various alternate clustering algorithms. The estimated origin location is essentially unchanged across different clustering methods.

Regardless of how we define the 'true origin' of scripture, it does not make a difference for the analysis. Both Greece and Northern Africa, the two predominant monastic centres of early Christianity, are entirely overlapping the estimated confidence areas, which are centred roughly halfway between them (figure C15). Notably, Constantinople (Istanbul), the historical administrative capital of Byzantium after Constantine adopted Christianity is also inside the estimated confidence regions. Perhaps because of the estimated area had expanded to include Constantonople (Istanbul), Jerusalem is also inside the estimated region, unlike with Judaism, where it lay outside the confidence area.[44][45] Nevertheless, the estimate remains much closer to the monastic centres at the time of Christianity's spread than it does to the religion itself. This can be seen most clearly in table 6, which shows the distances from our estimate to each of the religious origin, and the origin of the scripture (columns 3 and 4). When we use the Buddhism calibration (column 3), the distance to the scripture's origin is just over half the distance to the religion's origin, whereas when we use the Islam calibration (column 4) the distance to the origin of the scripture is about 2.5 times closer. Both of these differences are statistically significant beyond the 1% level, as they were in the case of Judaism.

Finally, we move to Hinduism. In the case of Hinduism the historical account is far from resolved (appendix A.2). The ongoing debate attributes the origins of Hindu scripture either to the Indus Valley civilization (in the Indus Valley), where archaeological evidence has found similarities with iconography in modern Hindu scripture, or to central Asia, where the oldest known Hindu scripture, the *Rg veda*, has been attributed. The origin of Hinduism itself though, is incredibly old, by far the oldest of the five religions. The debate is contentious because it is tied into the origin of Indo-European people, which itself remains a heavily-debated academic question, though the most dominant hypothesis places the origin in the Pontic Steppe.

Both estimates, calibrated using either Buddhism or Islam, are located in south-central Asia, consistent with the hypothesized location of the origin of Hindu scripture (figure C18). The Islam estimate is slightly farther east than the Buddhism estimate, and narrowly leaves out the BMAC, which is one of the hypothesized regions of Hindu scripture, but does fully overlap with the Indus Valley region, which is the other main hypothesis (figure C18a). However, the estimate calibrated with Buddhism fully overlaps with both regions (figure C18b). Regardless of the calibration used, the estimates rule

[44]We present a series of robustness checks in the appendix figure C16 as well. In figure C16a we show that the estimated coordinate is in essentially the same place when we calibrate using a linear specification. In figure C16b we show robustness to a quadratic specification. In figure C16c we calibrate using a linear dependent variable instead of the log-dependent variable. In figure C16dwe calibrate using Betweenness Centrality instead of Eigenvector Centrality, while in figure C16e we examine Degree Centrality instead of Eigenvector Centrality. In all cases, the estimated origin location is essentially unchanged.

[45]In figure C17 we also demonstrate robustness to various alternate clustering algorithms. Again, the estimated origin location is essentially unchanged.

out the Pontic Steppe region that is typically thought of as the origin of Hinduism itself, there is no overlap in either case.[46][47]

Given these patterns, it is not surprising that the distances from our estimates to the history-literature based estimates are smaller in the case of the origin of scripture compared to the origin of the religion itself. This can be seen in table 6, columns 5 and 6. Indeed, the distance to the origin of scripture is 470km if we rely on the Buddhism calibration, and 1,100km if we rely on the Islam calibration. These estimates are larger than for each of the other religions, perhaps reflecting both the greater uncertainty associated with the history literature, and surely more measurement error associated with loanwords that would have had to have been borrowed so far into the distant past. In any case, despite these distances being larger for Hinduism, they remain much smaller than the comparable distances to the origin of the religion itself. With the Buddhism estimate the distance to scripture is more than 6 times closer, and for Islam it remains more than two and a half times closer. As before, in both cases the difference is significant well beyond the 1% level.

Our conclusion, therefore, is consistent across each of Christianity, Judaism and Hinduism. In each case we find that the language-based estimates are significantly closer to the origin of the religion's scripture than to the origins of the society in which the religion started. While this nuance may help to explain some of the discrepancy that has caused disagreements in the history literature, it also stands in stark contrast to early proponents of using etymology to trace historical phenomena, who argued explicitly that linguistic analyses "serve best for determining the origin of peoples" (Leibniz (1996 translation), p. 285).

7. Conclusion

This article empirically assesses the validity of using language etymology to make inferences about the origins of historical phenomena, and provides some suggestive evidence that the methodology serves better to identify *spread* of a phenomena than the origin of the phenomena itself. To do this we implement two empirical tests, applied to the historical origins of religion. The first is to test, in the case of Islam and Buddhism, which have straightforward and uncontested origins, whether a fully automated analysis can locate the latitude and longitude of the origin of these religions in the correct places.

[46] We present a series of robustness checks in the appendix figure C19 as well. In figure C19a we show that the estimated coordinate is in essentially the same place when we calibrate using a linear specification. In figure C19b we show robustness to a quadratic specification. In figure C19c we calibrate using a linear dependent variable instead of the log-dependent variable. In figure C19d we calibrate using Betweenness Centrality instead of Eigenvector Centrality, while in figure C19e we examine Degree Centrality instead of Eigenvector Centrality. In all cases, the estimated origin location is essentially unchanged.

[47] In figure C20 we also demonstrate robustness to various alternate clustering algorithms. Again, the estimated origin location is essentially unchanged.

The second is to test, in cases where the origin of the religion differs from the origin of scripture, whether etymology-based estimates are closer to the former than the latter.

We are able to reasonably accurately estimate the origin of each of Islam and Buddhism using only information on how words sound and what they mean. In doing so, we present the first quantitative evidence that linguistic analysis can be used in an empirically rigorous way to reconstruct history. Since our approach is entirely empirical – from the identification of religious words, to the estimation of their etymology, to their link with geographic coordinates – we avoid the main critique associated with using language to reconstruct history. Namely that it is too open to interpretation by researchers. Furthermore, the estimates for each of Judaism, Christianity and Hinduism suggest that, at least in the case of religion, language captures the origin of a standardized body of thought more accurately than sacred figures or religious origins. This stands in contrast to the traditional argument in favour of etymology-based historical reconstruction.

While the article is focused on religion, the ability to reconstruct history - at scale - in the absence of detailed primary sources may be able to make the study of questions/contexts that were previously impossible to explore more feasible. That said, there may be important contextual details that are important for the success of the methodology that cause the estimates to be particularly accurate in the case of religion. While we leave the generalizability of the methodology to future work, out approach may be applicable to other questions economic history where identifying the origin of spread of an idea or innovation is of interest. Our approach uses a single measure of linguistic transmission and is therefore most applicable to location of origin rather that time of origin. [48] One potential example would be to understand whether slavery or other social institutions had origins within colonized regions, or whether they were colonial imports. To the extent that this can be applied more generally, it could help illuminate the histories of peoples, places, and phenomena for which records have been ignored or destroyed.

References

Abramitzky, Ran et al. (2021). "Automated linking of historical data". In: *Journal of Economic Literature* 59.3, pp. 865–918.

Ager, Simon (2019). *Omniglot*.

Alesina, Alberto and Paola Giuliano (2015). "Culture and Institutions". In: *Journal of Economic Literature* 53.4, pp. 898–944. DOI: 10.1257/jel.53.4.898. URL: http://www.aeaweb.org/articles.php?doi=10.1257/jel.53.4.898.

[48] This would require a list of seed words relevant to the topic in question, similar to our list of religious seed words. This would also require a known origin that follows a process of spread similar to spread from the unknown origin. Since the approach is based on linguistic transmission, it is particularly suited to applications where the topic involves new words, rather than re-interpretations of existing words.

Algeo, John, ed. (1993). *Fifty years among the new words: a dictionary of neologisms, 1941 - 1991*. Centennial series of the American Dialect Society. Cambridge: Cambridge Univ. Press.

Armstrong, Karen. (2001). *Islam : a short history*. eng. Publication Title: Islam : a short history. London: Phoenix. ISBN: 1-84212-462-5.

Assael, Yannis et al. (Mar. 2022). "Restoring and attributing ancient texts using deep neural networks". en. In: *Nature* 603.7900, pp. 280–283. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-022-04448-z. URL: https://www.nature.com/articles/s41586-022-04448-z (visited on 08/17/2022).

Bailey, Martha J et al. (2020). "How well do automated linking methods perform? Lessons from US historical data". In: *Journal of Economic Literature* 58.4, pp. 997–1044.

Baledent, Anaëlle, Nicolas Hiebel, and Gaël Lejeune (May 2020). "Dating Ancient texts: an Approach for Noisy French Documents". English. In: *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*. Marseille, France: European Language Resources Association (ELRA), pp. 17–21. ISBN: 979-10-95546-53-5. URL: https://aclanthology.org/2020.lt4hala-1.3.

Barjamovic, Gojko et al. (2019). "Trade, merchants, and the lost cities of the bronze age". In: *The Quarterly Journal of Economics* 134.3, pp. 1455–1503.

Becker, Sascha O and Luigi Pascali (2019). "Religion, division of labor, and conflict: Anti-Semitism in Germany over 600 years". In: *American Economic Review* 109.5, pp. 1764–1804.

Becker, Sascha O and Steven Pfaff (2022). "Church and State in Historical Political Economy". In: *The Oxford Handbook of Historical Political Economy*. Ed. by Jeffrey Jenkins and Jared Rubin.

Ben Hamed, Mahé (2015). "Phylo-linguistics: Enacting Darwin's Linguistic Image". en. In: *Handbook of Evolutionary Thinking in the Sciences*. Ed. by Thomas Heams et al. Dordrecht: Springer Netherlands, pp. 825–852. ISBN: 978-94-017-9013-0 978-94-017-9014-7. DOI: 10.1007/978-94-017-9014-7_39. URL: http://link.springer.com/10.1007/978-94-017-9014-7_39 (visited on 03/14/2022).

Bloomfield, Leonard (1939). "Linguistic aspects of science." In: *International encyclopedia of unified science*.

Blouin, Arthur and Julian Dyer (2022). "How Cultures Converge: An Empirical Investigation of Linguistic Exchange, Trade, and Power". In: *Working paper*.

Bojanowski, Piotr et al. (2017). "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146.

Campo, Juan E. (2009). *Encyclopedia of Islam*.

Chawla, N. V. et al. (2002). "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16, pp. 321–357.

Coleman, Robert (1988). "Book review of Archaeology and Language by Colin Renfrew". In: *Current Anthropology* 29.3. Publisher: [University of Chicago Press, Wenner-Gren Foundation for Anthropological Research], pp. 437–468. ISSN: 00113204, 15375382. URL: http://www.jstor.org/stable/2743460 (visited on 09/16/2022).

Diebold, R.E (1994). "Linguistic paleontology". In: *The Encyclopedia of Language and Linguistics*. Pergamon Pres, pp. 2906–13.

Feigenbaum, James J (2016). "Automated census record linking: A machine learning approach". In:

Frankopan, P. (2016). *The Silk Roads: A New History of the World*. Knopf Doubleday Publishing Group.

Giorcelli, Michela, Nicola Lacetera, and Astrid Marinoni (Sept. 2022). "How does scientific progress affect cultural changes? A digital text analysis". en. In: *Journal of Economic Growth* 27.3, pp. 415–452. ISSN: 1381-4338, 1573-7020. DOI: 10.1007/s10887-022-09204-6. URL: https://link.springer.com/10.1007/s10887-022-09204-6 (visited on 01/18/2023).

Haspelmath, M. and U. Tadmor (2009). *Loanwords in the World's Languages: A Comparative Handbook*. De Gruyter Mouton.

Lehmann, WP (1968). *The System of Sonants and Ablaut in Kartvelian Languages: A Typology of Common Kartvelian Structure*.

Leibniz, Gottfried Wilhelm (1996 translation). *New essays on human understanding*. Cambridge University Press.

Lemaitre, Guillaume, Fernando Nogueira, and Christos K. Aridas (2017). "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning". In: *Journal of Machine Learning Research* 18.17, pp. 1–5.

Lewis, Paul M. (2009). *Ethnologue : languages of the world*. Texas: SIL International.

Lowes, Sara and Eduardo Montero (Apr. 2021). "The Legacy of Colonial Medicine in Central Africa". In: *American Economic Review* 111.4, pp. 1284–1314. DOI: 10.1257/aer.20180284. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20180284.

Mackintosh-Smith, T. (2019). *Arabs: A 3,000-Year History of Peoples, Tribes and Empires*. Yale University Press.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (June 2013). "Linguistic Regularities in Continuous Space Word Representations". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, pp. 746–751. URL: https://aclanthology.org/N13-1090.

Mortensen, David R, Siddharth Dalmia, and Patrick Littell (2018). "Epitran: Precision G2P for many languages". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Nunn, Nathan and Leonard Wantchekon (2011). "The Slave Trade and the Origins of Mistrust in Africa". In: *American Economic Review* 101.7, pp. 3221–3252.

Pascali, Luigi (2016). "Banks and development: Jewish communities in the Italian Renaissance and current economic performance". In: *Review of Economics and Statistics* 98.1, pp. 140–158.

Price, Joseph et al. (2021). "Combining family history and machine learning to link historical records: The Census Tree data set". In: *Explorations in Economic History* 80, p. 101391.

Ross, Gavin JS (1990). "A Program for Fitting Nonlinear Models, MLP". In: *Nonlinear Estimation*. Springer, pp. 143–173.

Rubin, Jared (2014). "Printing and Protestants: an empirical test of the role of printing in the Reformation". In: *Review of Economics and Statistics* 96.2, pp. 270–286.

Schadeberg, Thilo C. (2009). "Loanwords in Swahili". In: *Loanwords in the World's Languages: A Comparative Handbook*. Ed. by M. Haspelmath and U. Tadmor. De Gruyter Mouton, pp. 77–102.

Schmidt, Johannes (1872). *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen*. German. Weimar: Böhlau.

Swadesh, Morris (1950). "Salish Internal Relationships". In: *International Journal of American Linguistics* 16.4, pp. 157–167.

Valencia Caicedo, Felipe (2019). "The mission: Human capital transmission, economic persistence, and culture in South America". In: *The Quarterly Journal of Economics* 134.1, pp. 507–556.

Valencia Caicedo, Felipe, Thomas Dohmen, and Andreas Pondorfer (2021). "Religion and prosociality across the globe". In: *Working Paper*.

Vansina, J.M. (1990). *Paths in the Rainforests: Toward a History of Political Tradition in Equatorial Africa*. University of Wisconsin Press.

Wichmann, Soren, Eric W. Holman, and Cecil H. Brown (2016). "The ASJP Database (version 17)". In:

Yu, Xuejin and Wei Huangfu (Nov. 2019). "A Machine Learning Model for the Dating of Ancient Chinese Texts". In: *2019 International Conference on Asian Language Processing (IALP)*. Shanghai, China: IEEE, pp. 115–120. ISBN: 978-1-72815-014-7. DOI: 10.1109/IALP48816.2019.9037653. URL: https://ieeexplore.ieee.org/document/9037653/ (visited on 08/17/2022).

Table 2: Choice of Religious Seed Words

Heading	Seed-Words	Justification
Religion	*religion*	This is straightforward word to include, as the word religion is commonly used.
Sacred books	*sacred*	Here we drop the word 'book' and keep 'sacred' as we do not want to bias towards identifying the spread of books and scripture.
Natural theology	*god, astrology*	The sub-headings for theology focus primarily on deities, and different types of understanding of deities, so 'god' is a fairly broad representation of this concept that appears in common usage. We also include 'astrology' to capture a broader range of natural theology.
The soul	*spirit*	Here soul is a commonly used word that is broadly applicable across all of our religions of interest. We selected *spirit* as a seed-word for the soul category, as the concept of a soul is less universal than the concept of a *spirit*.
Eschatology	*afterlife*	Eschatology is defined in the Oxford English Dictionary as "The department of theological science concerned with 'the four last things: death, judgement, heaven, and hell'."[12] In order to represent this without specifically referencing Christian or another specific understanding, we chose to include *afterlife* as a broad seed-word capturing concerns with what happens after death or the ending of the world.
Worship. Cultus	*worship*	As the word "cult" may have other non-religious connotations and may be more likely used in the study of a certain religion rather than by its practitioners, for this category we chose the word *worship*, which occurs in common usage and is fairly universal.
Religious life	*pray*	For religious life, we chose to include the seed-word *pray*, as the concept and act of prayer appears to be relatively universal across most religions, and without including words such as non-religious concepts "contemplation" or 'meditation".
Religious organization (people)	*priest*	We include the word priest, as well as similar words *monk* and *preacher* to capture a broad range of the people involved in religious organization.
Religious organization (places of worship)	*church, temple, mosque*	We include these seed words for different forms of religious institutions, including other similar words like *synagogue*, *shrine* and *sanctuary* to broadly cover the concept of places of worship.

Note: This table describes how we go from the final list of relevant headings from the Library of Congress Classification in Table B2 to the actual seed words we use for our semantic similarity routine to identify related words across languages.

Table 3: Summary Statistics

Variable	Mean	Std. Dev.	Min.	Max.	N
Any religious language adoption	0.016	0.124	0	1	4,839,955
Share adopted (conditional on any adoption)	0.03	0.056	0	1	12,910
Distance between lender and borrower centroids (km)	8.206	4.556	0	20.029	4,839,955
Centrality of Lender in Religious Language Network	0.005	0.021	0	0.309	4,839,955
Centrality of Borrower in Religious Language Network	0.001	0.011	0	0.309	4,839,955
Number of Religious Words Identified	69.543	106.578	0	3438	4,839,955
Latitude of centroid of lender	18.306	13.569	0.078	59.941	4,839,955
Longitude of centroid of lender	50.912	34.525	10.017	109.985	4,839,955
Latitude of centroid of borrower	6.641	18.324	-51.635	73.135	4,839,955
Longitude of centroid of borrower	52.666	83.585	-173.925	177.657	4,839,955

Note: In this table, we present summary stats of the pair-wise religious language adoption used to reconstruct our estimates of religious origins. This includes summary statistics for the level of pairwise religious adoption, as well as on the network centrality measures of borrower and lender nodes and their coordinates of group centroids. We also share summary statistics of the number of words identified as being religious by the semantic similarity routine, with a histogram of the distribution presented in C3.

Table 4: Calibration: Linguistic network influence identifies geographic origins of spread

Dependent Variable:	log(Distance to Lumbini)		log(Distance to Mecca)	
	Influencer	Adopter	Influencer	Adopter
	(1)	(2)	(3)	(4)
Network influence - religious words	34.43***	- 3.59**	-16.79***	-16.54***
	(2.10)	(1.47)	(1.87)	(2.57)
(Network influence - religious words)^2	-371.98**	-19.88**	148.37***	70.38***
	(39.85)	(8.75)	(25.58)	(16.11)
(Network influence - religious words)^3	1007.07***	25.66**	-81.25 ***	-647.3***
	(158.08)	(11.58)	(86.04)	(22.22)
Number of Words	✓	✓	✓	✓
Distance between partners (cubic)	✓	✓	✓	✓
N	4,839,955	4,839,955	4,839,955	4,839,955
R^2	0.158	0.184	0.107	0.4793

Note: This table examines the relationship between network influence for religious words, and distance to the origins of religious spread. The unit of observation is a language-group pair. Standard errors are two-way clustered by each language group in the pair. *, **, *** denote statistical significance at the 10%, 5%, 1% levels respectively.

Table 5: Validation: Is historically relevant information embedded within languages?

Religious origin:	Buddhism		Islam	
Calibration using:	Buddhism (1)	Islam (2)	Buddhism (3)	Islam (4)
Mean distance (km) using religious loanwords-based estimates	393.2	405.8	392.2	287.9
Mean distance (km) using random estimates	1,438.4	1,387.2	867.5	1,491.8
Difference (km): random - loanwords	1,045.2	981.4	475.3	1,203.9
t-statistics – H_0**: random - loanwords** = 0				
regular t-statistic	106.5***	232.9***	109.6***	304.8***
N	9,533	18,001	9,456	23,243

Note: In this table we present the distances between the centroids of the true origins of Islam and Buddhism and the estimated ones. We do this for both the estimates derived from the calibration exercise (based on table 3) as well as based on random information in place of the calibration. In columns 1 and 2 we show the estimates for Buddhism, calibrated based on the distance to Buddhism (column 1) and the distance to Islam (column 2). In columns 3 and 4 we show the estimates for Islam, calibrated based on the distance to Buddhism (column 3) and Islam (column 4). Towards the bottom of the table we compute the difference between the differences based on the linguistic network calibration and the random information estimates, and present t-tests for the null-hypothesis that the estimates based on random information are the same as those based on linguistic network information. In all cases we can reject the null, on the basis that the distances based on language information are always smaller than those based on random information. The number of observations change from column to column based on the number of estimates assigned to each respective cluster.

Table 6: What does language capture? Origin of Scripture or Religion

Religious origin:	Judaism		Christianity		Hinduism	
Calibration using:	Buddhism (1)	Islam (2)	Buddhism (3)	Islam (4)	Buddhism (5)	Islam (6)
Mean distance (km) to religion origin	1000.2	1,085.8	822.1	726.9	2,757.0	2,937.2
Mean distance (km) to scripture origin	221.9	433.2	448.2	283.9	470.5	1,132.2
Difference (km): religion - scripture	530.1	778.2	373.9	443.0	1,613.8	1,805.0
t-statistics – H_0**: religion - scripture** = 0						
t-statistic	34.8***	90.9***	27.4***	57.3***	188.2***	228.8***
N	6,626	20,944	8,224	25,712	10,410	24,684

Note: In this table we present the estimated latitudes and longitudes for origins of religious spreads alongside the actual origins of religious spread (see appendix A for an explanation of how each actual origin was selected with reference to the historical literature). We also show the p-values for a test that the estimated and actual coordinates are the same, and show that none of the differences are statistically significant.

Appendix A. Historical Accounts of Early Religious Spread

A.1. Buddhism

While Buddhism is less focused on the life of a single sacred individual than other major world religions (Pyysiäinen (2003)), the teachings of the Buddha (meaning 'Enlightened One') form the core tenets of Buddhism (Humphreys (2013)). The Buddha, Siddhattha Gotama (or Siddhartha Gautama), was born in Lumbini (Weise (2013)) less than 20km west of the modern city of Bhairahawa or Siddharthanagar (location a on map figure A1), along what is now the Nepal-India border. Prior to Buddhism, the dominant religion in this region had been Brahmanism, which is closely related to Hinduism Harvey 2012, p.8. The exact dates of the birth and death of the Buddha are not known with certainty, but recent work estimates him to have died within decades of 400 BCE (Prebish (2008), p.2).

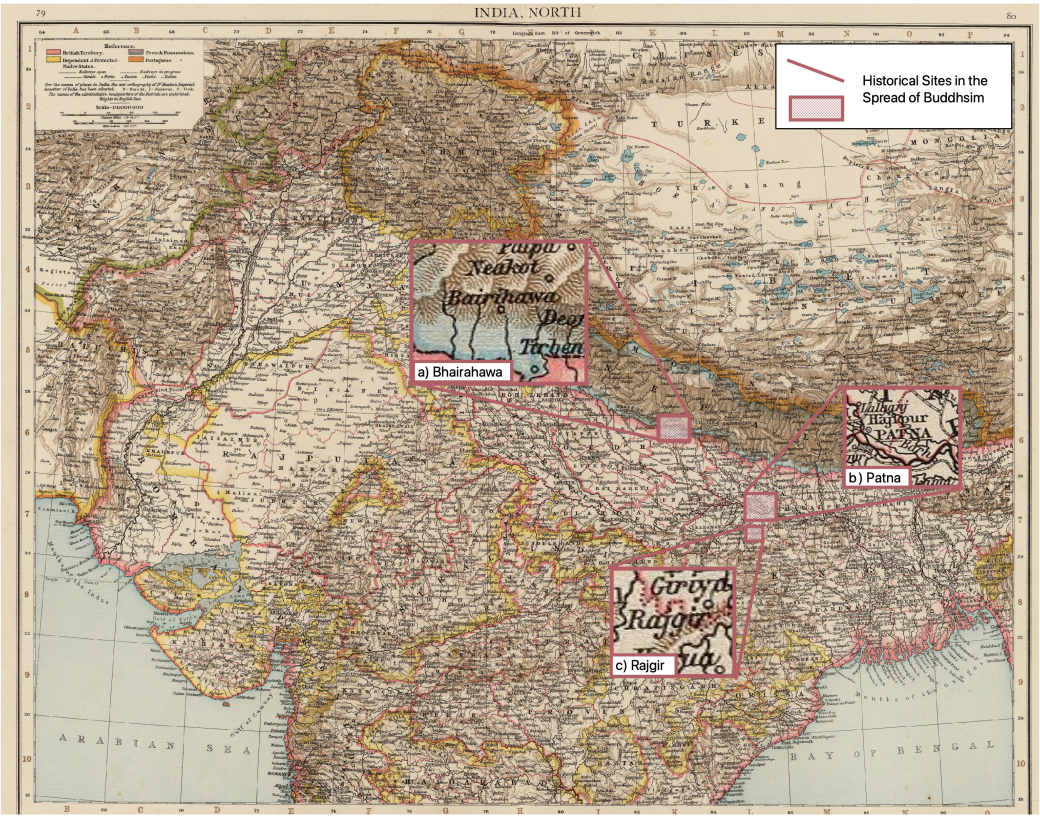


Figure A1: Historical Sites in Spread of Buddhism

Note: The map shows historical sites in the origin and spread of Buddhism. The origins of Buddhism are uncontroversial, with the Buddha's life and the compilation of his teachings into a standard scripture all occurring within the Ganges river basin.
Map Source: (Andree (1895)), accessed through the David Rumsey Map Collection, David Rumsey Map Center, Stanford Libraries. Authors's own highlighting of key sites.

All accounts of the origins of Buddhism recount that Siddhattha was the son of the rulers of the Sakka, living a sheltered life of luxury, before adopting a monastic existence and eventually attaining enlightenment while seated underneath the *Bodhi*, or 'Awaken-

ing' tree (Harvey 2012, p.14-23 and Strickland and Coningham 2020, p.13). The teachings that would form the religion of Buddhism spread as the Buddha wandered, teaching monks, throughout the basin of the Ganges river in north-eastern India (Skilton (1997)). These disciples were then sent to spread Buddhist teachings more broadly Harvey 2012, p.24.

After the Buddha's death, a council of five hundred *Arahats*, or enlightened disciples, was held at Rājagaha in northeast India (location b on map figure A1, south-east of Lumbini), to decide on the *Dhamma* and *Vinaya*, which were the core scriptures that recorded his teachings (K. W. Morgan (1986)). At later councils in nearby Vesālī and Pāṭaliputta, differences over monastic rules among schools began to emerge and by 100 C.E. there were true schisms of doctrine (Harvey (2012), p.88-89). With the introduction of new texts, or *sutras*, the schisms within Buddhism between Mahāyāna school, who accepted the new teachings, and the conservative Theravāda school, who rejected the new teachings, became more deeply entrenched (Harvey (2012), p.95). Another later version of tantric Buddhism emerged from Mahāyāna Buddhism, though with a greater emphasis on secret meanings and rites (Harvey (2012), p.190).

Buddhism began to spread rapidly during the reign of the Emperor Asoka (268–239 BCE) of the Magadhan empire, which included most of contemporary India (Przyluski (1934)). Following his invasion of Kaliṅga in approximately 260 BCE, Asoka began to regret the devastation he had caused, and began to rule according to Buddhist principles in order to improve the lives of his subjects (Premasiri (2022)). During this time, Buddhism occupied a central role in his empire and spread rapidly (Harvey (2012), p.101). Buddhism, however, began to fade in the Indian sub-continent in the face of later hostility from Hindu and Islamic rulers, and a large share of Indian Buddhists were either absorbed into Hinduism, or converted to Islam (Harvey (2012), p.195).

Buddhism spread north into China during a period of competition among the various schools of Buddhism discussed above (Ch'en and Ch'en (1972)). In addition to its adaptation to a different cultural environment, this led to a very distinctive form of Chinese Buddhism (Zürcher (2007), p.2). Many of the Buddhist concepts that followed its introduction into China were innovations consistent with a broadly Indian worldview that were often incompatible with existing Chinese thought (Coleman (2002)). This led to a large literature by devotees attempting to reconcile Buddhist thought with existing worldview (Zürcher (2007), p.12). The exact history of the arrival of Buddhism in China is unknown, but most traditional stories recount the arrival of Buddhist missionaries bringing the new religion (Chen (2004)). While the details are unknown, it is accepted that Buddhism spread into China from the North-West along trade routes (Strauch (2019)). By the beginning of the third century there were a large number of Indian texts of Buddhist scriptures circulating in China (Boucher (1996)). Buddhism continued to spread further into China, most likely following the eastern branches of the continental silk-road

trade routes (Zürcher (2007), p.19 - 26). Scripture in imperial circles further confirms this path:

> There is in the first place the significant fact that [...] the words *upāsaka* and *śramaṇna* figure in the text of an imperial edict. This can only mean that these Indian (or Central Asian) Buddhist terms were known and understood in court circles, and that they meant something to the emperor Zürcher 2007, p.29

It is therefore widely accepted that Buddhism spread broadly, accompanied by Indian-language scripture and conceptual innovations, into China from its origin in north-eastern India. From its origin in Indic language texts, Chinese Buddhism then spread into Vietnam in the third century and by the tenth century was flourishing, including state patronage and an elite cultural position (Huy (1998)). Other forms of Buddhism (including the Theravāda and Mahāyāna schools) had also spread to south Vietnam, though an invasion in the fifteenth century lead to Chinese Buddhism dominating all of Vietnam (Tien and Shih (2016)). The same is broadly true of Korea, where Buddhism spread from monks studying in China, and led to Buddhism become an influential religion among the aristocracy and the populace more broadly (Buswell Jr (2013)). Missionaries sent from Korean kings reached Japan in the mid sixth century, bringing with them the Chinese Buddhism they had adopted (Harvey (2012), p.210-224).

The exact date at which Buddhism arrived in Tibet is unclear, though it is known that the emperor Tri Songdetsen was Buddhism's strongest advocate (Hirshberg (2016)). He founded the first monastery in Tibet in the late eighth century, inviting the Indian Buddhist monk Śāntarakṣita to teach the Indian monastic code he practiced (Blumenthal and Apple (2008)). This initiative was heavily supported by the imperial state, who commissioned committees to translate the original Indian scripture into Tibetan, creating a new standardized religious vocabulary in Tibet (Bray (1991)). Tri Songdetsen's dynasty lasted until the mid-ninth century and during this time laid the roots of Tibetan Buddhism built on the Indian Buddhist teachings above (Kapstein (2013), p.12-16). From Tibet, Buddhism spread further north into Mongolia (Beckwith (1987)). After a missionary monk came to the court of the Khan and Queen Jönggen, they later, in 1578, invited religious leaders from Tibet to initiate them in Buddhist teachings (Elverskog (2015), p.6). Similarly to other Central Asian Buddhist states, the Oirat rulers of Western Mongolia sponsored the acquisition and translation of Buddhist texts (Rawski (2005)). In Mongolia, this was accompanied by the creation of a new script (the 'Clear Script') by scholars and young nobles sent to Tibet for education in Tibetan monastic traditions (Taupier (2015), p.24-32).

It is therefore uncontroversial to locate the origin of Buddhism in the Ganges river basin near the Nepal-India border. From there it then spread south through the Indian

subcontinent, as well as north and east into Tibet and China, from where it next spread into Mongolia, Vietnam, Korea, and Japan. [49] We take Lumbini, the birthplace of the Buddha, as the precise origin location.

A.2. Hinduism

Hinduism is the oldest religion of the five that we consider, and accordingly it is the one with origins that are most uncertain (Narayanan (2009)). While there are differences of opinions among historians, most agree that Hinduism can be traced back to either the Indus civilization (c. 3000-1500 BCE) or the Indo-European civilization during the vedic period (c. 1500-500BCE). Contrasting that with the earliest known writing out of India coming in the 4th century C.E., and it should not be surprising that there exist differences of opinions regarding the precise geographic and temporal origins of Hinduism (G. D. Flood and G. D. F. Flood (1996)).

Perhaps the most regularly referenced origin of Hinduism goes back to about 3000 BCE, near the Indus Valley (Saeed (2019)). The Indus Valley comprises the fertile region around the Indus River (see location *a)* in Figure A2) which runs north-south through East-Central Pakistan. The dating of the Indus civilization to this time period is based on carbon-14 dating of artifacts found during archaeological digs (Alessio et al. (1969)). We know from this type of evidence - with some certainty - that there were two large cities that were important to the Indus Civilization (Raikes (1964)), one was near modern day Moenjo-dāro (see location b in Figure A2), and the other near what is now Hrappā (see location c in Figure A2). While many scholars attribute the origins of Hinduism to these regions, that claim is contested (Prakash (1994)). Importantly, reconstructing a history of the Indus civilization has been difficult since we have not yet made progress on deciphering the script used in the artifacts uncovered from the region (Kenoyer (2006)).

Importantly, there remains considerable debate on the precise origins and timing of this process. The attribution of Hinduism's origins to the Indus valley stems from "cultural clues [that] lead archaeologists to interpret their finds as precursors to later well-known imagery and beliefs" (Hitchcock and Esposito (2004), p. 77). The main alternative to this approach is that Hinduism began with the Indo-European civilization, who either conquered Moenjo-dāro and Harappā, or moved in following natural disaster at some point around 1500 BCE (Parpola (2015)). The Indo-European homeland itself is heavily contested, with significant disagreements between Linguists and Archaeologists. For instance, "On the surface of it, the chasm dividing the respective disciplines could not be wider" (Erdosy (1995)).

Covering the entirety of the debate surrounding the origins of Indo-Europeans is well beyond the scope of this study, as it requires a summary of the spread of Proto-Indo-

[49] Interestingly, and consistent with our overall approach in this project, each step in this path was marked by linguistic adoption.

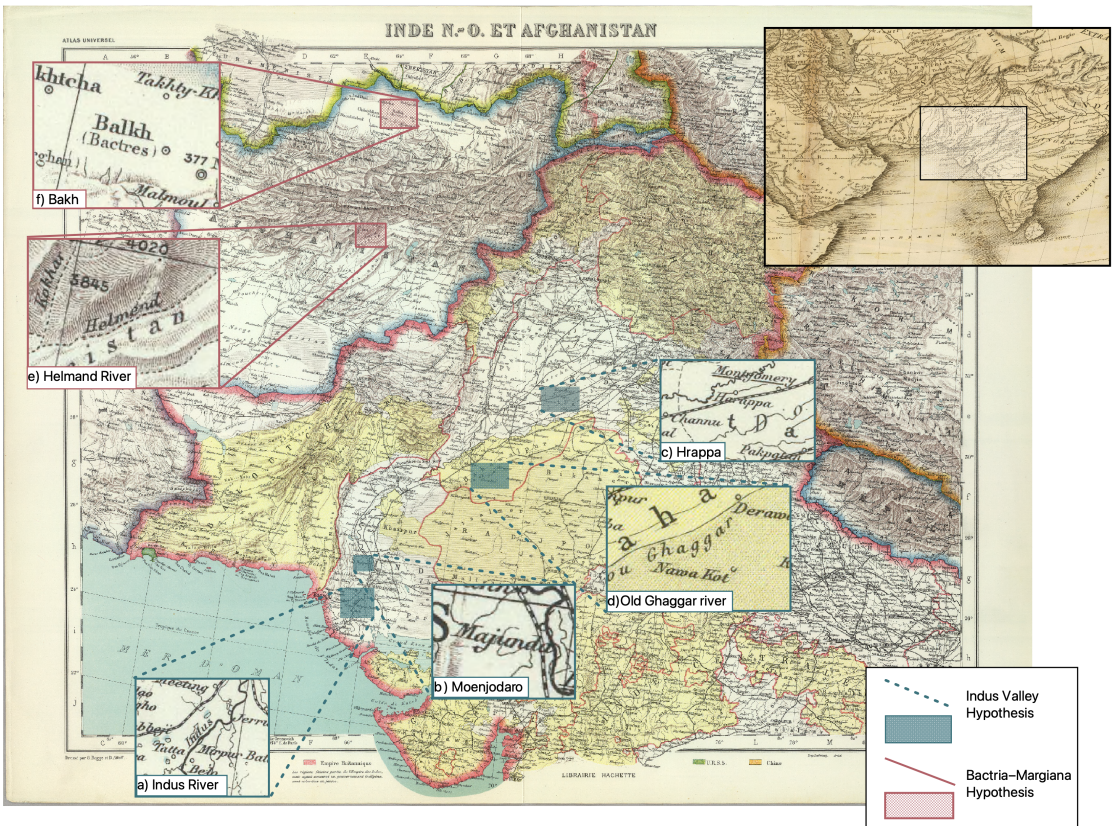


Figure A2: Hypothesised Sites in Spread of Hinduism

Note: The map shows historical sites from the different views on the origins of Hinduism's spread. Locations a) - d) lie in the Indus Valley and correspond to the theory that Indo-Europeans adopted religious practices from Indus Valley civilizations. Locations e) and f) indicate locations near the Bactria–Margiana Archaeological Complex in modern-day Afghanistan, which correspond to the hypothesis that Aryan civilizations initiated the spread of Hinduism prior to their migration to the Indus Valley.
Map Source: Schrader and Martin (1936), accessed through the David Rumsey Map Collection, David Rumsey Map Center, Stanford Libraries. Author's own highlighting of key sites.

European (PIE) languages themselves. This is a long, heavily contested debate, combining perspectives from linguistics, archaeology and genetics. A very brief and inadequate summary is that there are two main theories. First, is the "Steppe Hypothesis" that places early PIE speakers in the Pontic Steppe[50] in the 5th millenium BCE, from which migrants spread IE language south to Turkey, north-west to Europe, and east to east-Asia (Renfrew (1990)). Under this hypothesis, offshoots of the eastern migrating branch settled in the Bactria–Margiana Archaeological Complex (henceforth BMAC), who then migrated south into north-western India, perhaps bringing the original Hindu ideas with them.

The second hypothesis is that the Yamnaya of the Pontic Steppe were ancestors of the Anatolia (from modern-day Turkey), given the lack of direct genetic impact of the Yamnaya in Asia (Barros Damgaard et al. (2018)). This hypothesis is consistent with two western waves of migration into South Asia. A first wave from roughly 3300 BCE

[50]The Pontic Steppe comprises from a predominantly east-west corridor that ranges roughly from Ukraine to Mongolia

originating in central Asia, that did not come along with IE languages, and a second wave between 2300-1200BCE south from the steppe, through the Indus Valley, and into what is now north-west India (Barros Damgaard et al. (2018)). The relevance of these debates to the origins of Hinduism stems from the attribution of the earliest known Hindu scripture to the Indo-European civilization (Srinivasan (1983)), whose origins may differ depending on beliefs about the route PIE speakers took from the west, into south Asia. Without delving into the timing, or precise routes of travel, it seems relatively safe to conclude that:

> The ancestors of the Indian Indo-Europeans had remained for a long time on the borders of the subcontinent, in what are now Afghanistan and Soviet Central Asia. Some time after the Indo-European migration into India, another branch, the ancestors of the Medes and the Persians, left their homeland for what is now Iran and gave their name to that land (the name *Iran* comes from *Airyānām vaējō*, "Realm of the Indo-Europeans") Basham 1991, p. 8.

So to summarize, the oldest Hindu scripture, the *Rg veda*, was probably written by the Indo-European civilization who, at some point, likely inhabited the region near what is now Afghanistan. Around the same time, religious practices by Harappā in the Indus valley may have exhibited themes very similar to modern Hinduism, and which of these societies forms the true origin of Hinduism is not known with certainty. Opinions appear divided based on whether one trusts the evidence based on ancient texts (which would lead to an Indo-European origin) or archaeological evidence (which would lead to an Indus Valley origin). While the archaeologial evidence may seem more concrete than interpretations of linguistic texts, the archaeological evidence is certainly not overwhelming. For example:

> ...[I]t is often confidently stated that the religion of the Harappā culture was an early form of Hinduism. The evidence, in our view is inadequate. The identification of two profile faces is very uncertain (they may be parts of the god's headdresses); the full face of the god is closer to that of a tiger than that of a man; it is not completely certain that the god is ithyphallic, since the marks taken to indicate this may be merely the loose part of a girdle; most of the animals associated with the god are not those specially connected with Siśva. In fact, the evidence for any kind of continuity between this prehistoric god and Siśva is rather weak. Evidence for other features of Hinduism in the civilization of the Indus is even more dubious. It has been suggested that this culture practiced ritual prostitution, in the manner of the temple prostitution of later India, or that the inducement of trance or calm states, later called yoga was practiced. But the evidence for these practices is so tenuous that

> the suggestions are quite unacceptable except as faint possibilities Basham 1991, p. 8.

Even among those who concede that the *Rg veda* is convincing evidence of Indo-European Hindu origins, there has been significant archaeological evidence, and accompanying discussion surrounding whether the BMAC is the appropriate origin of the Indo-Europeans (V. Srivastava and Shrivastva (1981)). This evidence places the Indo-Europeans in southern Afghanistan rather than northern Afghanistan. A large component of this argument is based on geographic references in the *Rg veda*. The Sarasvatī river is one of the only geographical features mentioned consistently throughout the *Rg veda*, in addition to the Indus River, however the Sarasvatī is described in somewhat more detail. It has been assumed to reference what is now known as the Old Ghaggar river (see location d) in Figure A2), also in the Indus valley (Bhadra, Gupta, and Sharma (2009)), but has also been claimed to reference the Helmand river (see location *e)* in Figure A2) (Kochhar (2000); Jain, Agarwal, and Singh (2007); G. Srivastava (2018)). This is significant because it would place the Indo-Europeans near Khandahar, in southern rather than northern Afghanistan (Kochhar (2012)).

While linguistic and textual interpretation tend to place Hindu origins in Afghanistan, and archaeology sometimes favours the Indus Valley, it is not true that there is an absence of archeological evidence linking central and south Asia, from the time that the Indo-European civilization was in central Asia. For instance:

> There remain, nevertheless, impressive parallels in material culture between Central Asia and South Asia in the Late Bronze Age. Shared traits include: specific vessel shapes (bottles, footed goblets, dishes-on-stand, spouted bowls and vessels with applique animals on the rim); kidney-shaped vases of steatite; alabaster columns, discs and statues; shaft-hole axe/adzes; bronze mirrors with anthropomorphic handles; circular stamp seals with snake-motifs; and so on (Sarianidi 1990: 86- 87, figures 14-15). What is more, these traits are all Central Asian in origin... – (Erdosy (1995))

We place the origin of Hinduism with the Indo-European civilization, and follow the mainstream view that they migrated south from the BMAC region. Accordingly, we place our estimate of the historical account at Bakh.

A.3. Islam

Muhammad was born in Mecca in the late 6th century, around the time that the Byzantine Empire had reached its greatest geographic extent, ruling along the coast of the Levant and into Egypt Brown 2011, p.3-9. After taking power in 527, Justinian had won back much of the western part of the empire from Gothic invasion, retaking Rome itself in

536 Moorhead 2013, p.74-79. The Sasanian Empire (also known as the Neo-Persian Empire), which comprised primarily of what is now Iran, had been the main rival to Byzantium since about the third century Dignas and Winter 2007, p.18-44. After the Roman conquest of Palmyra, the frontier between the two empires had moved near the Arabian peninsula Mackintosh-Smith 2019, p.66.

Paganism had been outlawed for over 100 years, but remained quite strong, alongside the rising tide of monotheistic Christianity and Judaism Mackintosh-Smith 2019, p.128. Christianity had become the state religion of the Byzantine Empire, after Constantine's conversion in 312 after a victorious battle, and was a core part of the empire's ideology Sarris 2015, p.3-14. Judaism remained quite strong in Egypt, with a sizeable diaspora community in Alexandria in particular Cohen 2018, p.52–59. The same was especially Mesopotamia and Babylonia, which formed one of the most cohesive and largest of the Jewish Diasporas A. Oppenheimer and N. Oppenheimer 2005, p.8, p.337-340. Zoroastrianism was the other important factor in the religious landscape at the time, as the dominant religion of the Sasanian Empire Curtis 2020, p.200-201. In the Arabic regions the pre-Islamic norm was polytheism, centered in particular on Mecca Mackintosh-Smith 2019, p.117-122, which became an important site of pilgrimage:

> Cultic life focused on a number of practices which survived, in revalorized form, in Islam, including pilgrimage to shrines. It is often assumed that the most important of these shrines was that centered on the Kaaba at Mecca, and that it was the object of a widely-shared pilgrimage cult among the pre-Islamic Arabs. This cult, so the traditional story goes, was tended to by the Quraysh, the tribe to which Muhammad belonged. (Berkey (2004), p. 42)

The accounts of the ascension of the Islamic Empire in the seventh centuries typically highlight its unexpectedness and suddenness; a "breach in cultural continuity unparalleled among the great civilizations" (Hodgson (2009), p.103). How and why this happened remains debated among historians, with a recognition that most if not all of the historical sources come several centuries after Muhammad's death, and reflect a desired, rather than experienced history Mackintosh-Smith 2019, p.124-125,Brown 2011, p.2.

That said, it seems fairly uncontroversial that Muhammad was raised in Mecca (location a in figure A3) and began preaching about a single God, which upset the majority polytheistic pagans that controlled the region Brown 2011, p.16-21. This prompted Muhammad and his followers to flee north to Medina (location b in figure A3), where they established the first Muslim community Brown 2011, p.26-28. Muhammad continued to preach in Medina until his death in 632 CE, after which his teachings were collected into the Qur'an Mackintosh-Smith 2019, p.134. Muhammad is thought to have been illiterate, but the dominant narrative is that his followers recorded what was revealed to him by God, and the Qur'an was spread by manuscripts that were copied by calligraphers

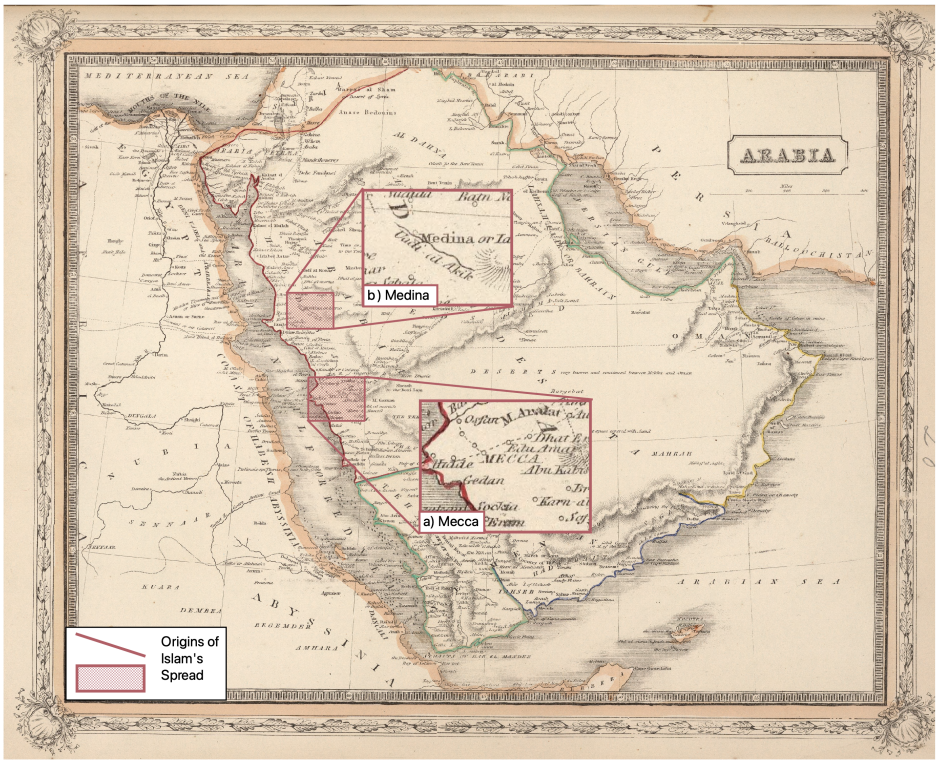


Figure A3: Important Sites in Spread of Islam

Note: The map shows historical sites in the spread of Islam, primarily Mecca and Medina.
Map Source: (Lothian (1848)), accessed through the David Rumsey Map Collection, David Rumsey Map Center, Stanford Libraries. Author's own highlighting of key sites.

Mackintosh-Smith 2019, p.134. The early manuscripts of the Qur'an were compiled on parchment sewn into a codex in a script derived from the Nabatean alphabet. This script allowed substantial variation in recitation which was soon perceived to be a threat to the unity of Islam, and early caliphs made concerted efforts to enforce a common version and prevent variation Dèroche 2022, p.111-139.

Following Muhammad's death in 632 with no clearly specified succession arrangement, various tribes in the region sought to separate themselves from Islam, which had become the dominant religion in the region Robinson 2011, p.193-194. Muhammad's successor, Abu Bakr defeated these tribes in the *ridda* wars, which is considered an important factor in the establishment of an Islamic state that was far more consolidated than any before it in the Arabian peninsula Esposito 2000, p.10-14. Abu Bakr died shortly thereafter, succeeded by Umar, who was assassinated in 644, but not before specifying a committee would select his successor Mackintosh-Smith 2019, p.213. Uthman ibn Affan was selected but was widely despised - largely for nepotism and favourable treatment to his own tribe Robinson 2011, p.204. He too was assassinated, and succeeded by Ali, a cousin of Muhammad. Ali's rule was marked by internal strife and conflict, and he was eventually killed by rebels Esposito 2000, p.15-16. Following his death, Mu'awiya I became caliph in

661 Robinson 2011, p.202-204. Mu'awiya I ruled for approximately two decades, marking the beginning of the Umayyad Caliphate Esposito 2000, p.16. Despite this relative calm, deep-seated and long-lasting divisions replaced the early unity of Islam Mackintosh-Smith 2019, p.221-222.

The reign of these four caliphs prior to Mu'awaiya make-up the Rashidun era, which marked a 24 year geographic expansion of Islam, into both the Byzantine and Sasarian Empires. In fact, the Sasanian Empire almost immediately collapsed following early Islamic conquest, with the last Sasanian emperor, Yazdgerd III killed in 651 in infighting after defeat to Islamic armies and his children going into exile in China Daryaee 2013, p.37-38. Byzantium too suffered sizable territorial losses, with the loss of Syria Donner 1981, p.111-112, as well as Egypt and the Mediterranean islands of Crete, Cyprus and Rhodes Robinson 2011, p.197-197.

Within the former Sasanian Empire, accounts of Zoroastrian temples being destroyed and replaced with Islamic mosques suggest that conquest led to forced religious conversion, though these records may reflect the later desire of Islamic chroniclers to emphasise the victory of Islam over Zoroastrianism (Peacock (2017)). Similarly, within formerly Byzantine Syria, Islamic administration had initially been built upon existing structures, but as Islamic institutional capacity grew, administrators began to encourage conversion, in addition to forced conversion of slaves taken by Muslims Cobb 2010, p.243-251. In Egypt, the mass conversion of Christians began late in the Umayyad Caliphate, when converts were exempt from poll taxes Mikhail 2014, p.64-65. Further away from Mecca and Medina, conversion to Islam took place largely through trade. By the 8th century the region conquered by Islamic forces controlled essentially all of the western portion of the Silk Road, at which point merchants became as important as commanders in Islam's further expansion (D. O. Morgan and Reid (2000)). Islamic expansion in this region emphasised the aspects of Islamic practice and values that supported trade Elverskog 2010, p.24. This was also true in India, where the primary agent of Islamic spread were independent Muslims, and where the resulting 'Monsoon Islam' was adapted to the needs of merchants engaged in exchange in unfamiliar and foreign lands Prange 2018, p.1-7. Muslim merchants were among the first to engage in 'direct trade' which has been attributed as a key force in the propagation of Islam outside of the directly conquered regions.

> On the receiving end, the new religion appealed to the local merchants because it legitimised their economic base more than most belief systems present at that time. Merchants converting to Islam had clear advantages including:
>
> (i) cooperation within the Muslim trading network;
>
> (ii) valuable contacts to expand their trade; and
>
> (iii) rules governing commercial activities naturally favouring Muslims

(Michalopoulos, Naghavi, and Prarolo (2018))

In the end, the origin of the spread of Islam is among the easiest to pinpoint. The origins in Mecca and then Medina were clearly where Islam started, and where the Qur'an was written down for the first time. The Rashidun caliphate, post-Muhammad was led entirely by people who had bee in Muhammad's inner circle while he was alive. Islam conquered the Sasarian empire during this brief period, and then took large swaths of territory from the Byzantine empire, representing the major form of spread, all emanating from Mecca / Medina, and all controlled by people from that region. Largely due to the fact that this expansion via conquest is relatively straightforward and well documented, we can unambiguously attribute the geographic origin of the spread of Islam in the Hejaz region, where Muhammad spent the majority of his life. We therefore take as the region of origin the area of the Arabian peninsula around Mecca and Medina under the rule of Muhammad during his lifetime. We take as the region of origin of the written compiled scripture as the Arabian peninsula under the rule of Abu Bakr. Both of these are represented in figure C7 based on the original maps from Armstrong 2001.

A.4. Judaism

The Israelites established a kingdom in Canaan - just east of the Mediterranean - around 1000 BCE (Hess (2007)). Canaan was to the north-east of Egypt, which had been an established kingdom since at least 3200 BCE, and to the south-west of the Assyrian, and then Babylonian empires (Shaw (2000)). The Israelites were likely Aramean migrants, who had previously lived in parts of what is now western Iraq and eastern Syria (Sparks (2007)). Early in their history the Israelites lived "mostly in the central hill country and the southern region, raising cattle, sheep, and goats and occasionally farming; sometimes they came into conflict with the sedentary population, but for the most part they kept to themselves" (Scheindlin (1998)).

The biblical story of the Israelites begins with Abraham, who was Mesopotamian, but had moved to Canaan with his son Isaac. Isaac's son Jacob moved to Egypt - pulled away from Canaan by famine. Jacob's family (i.e. his 12 sons) had economic success in Egypt, but their lineage became enslaved by the Egyptians until the Exodus, roughly around 1200 BCE. Historical consensus however, appears to be that "[t]here is, in fact, remarkably little of proven or provable historical worth or reliability in the biblical Exodus narrative, and no reliable independent witnesses attest to the historicity or date of the Exodus events" (Redmount and Coogan (1998), p. 89).

What does seem historically verifiable is that "From the end of the thirteenth century to the end of the eleventh century BCE, the Israelites were organized in Canaan as a confederation of twelve tribes" (Scheindlin (1998)). This territory spanned from the plain of Jezreel (now northern Israel) and Sidon (location a on map figure A4) in the

north, on the west by the Jordan river (location b on map figure A4, i.e. around the West Bank today) and in the south by Kadesh (location c on map figure A4).

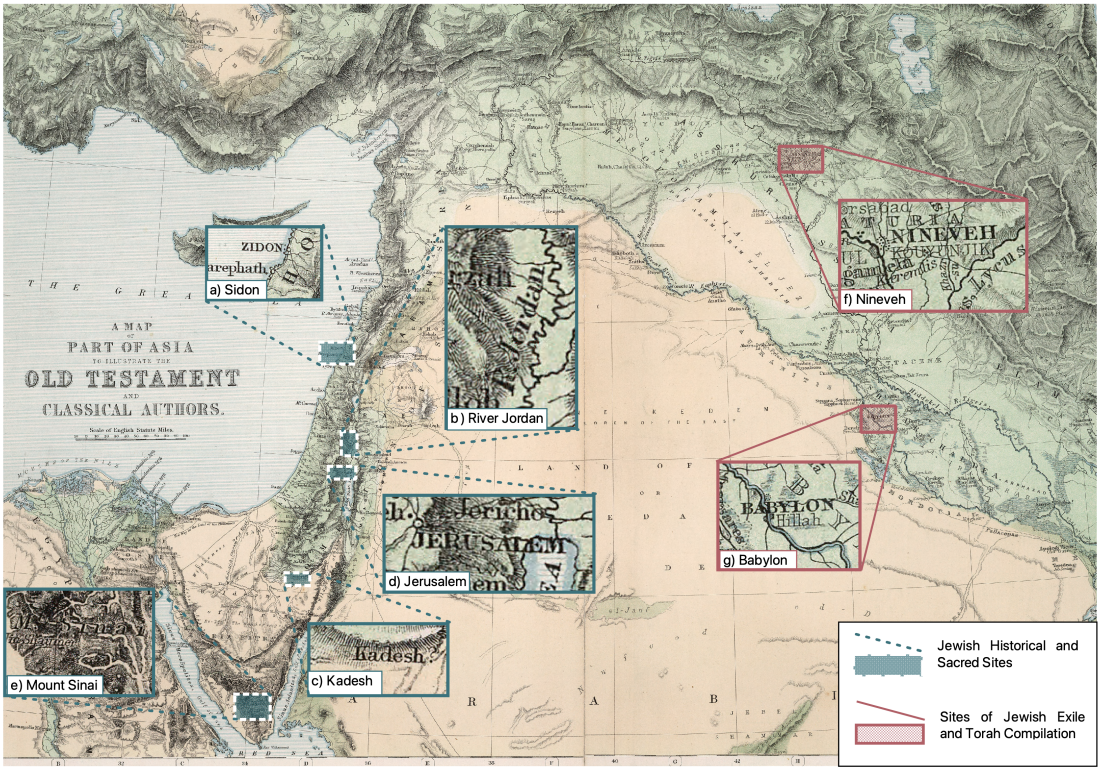


Figure A4: Important Sites in Spread of Judaism

Note: The map shows historical sites in the spread of Judaism, including sites of sacred figures and events in Israel, as well as the locations of Torah compilation in exile.
Map Source: Smith and Muller (1874), accessed through the David Rumsey Map Collection, David Rumsey Map Center, Stanford Libraries.. Author's own highlighting of key sites.

The Israelites economically developed in Canaan, from semi-nomadic to agriculturalists, until roughly 1000 BCE, when the Philistines (a militarily aggressive group from what is now Syria) took control of Canaan by force, including modern Gaza (Drews (1998)). In response the Israelites created a monarchy, with Saul as the first king (Ehrlich, M. C. White, and M. White (2006)). Saul eventually died during the conflict with the Philistines, and was succeeded by David, who eventually took control of the region, and made Jerusalem (location d on map figure A4) the capital of the newly consolidated region that became Israel (McFall (2010)). David was able to expand territory as far north as what is now Syria, and as far south as Mount Sinai (location e on map figure A4) (Scheindlin (1998)).

Solomon, David's son, continued the Davidic dynasty, and built the region into a economic power, though eventually lost territory, and faced difficulties at home with the introduction of taxation, which was unpopular (Scheindlin (1998)). After Solomon, Israel entered a period of relative instability, with many rulers, until Omri (Thiele (1983)). Omri made diplomatic agreements with neighbouring groups, and a period of relative peace

ensued (Pienaar (1994)). This lasted until Assyria (also called Asshur, named for their chief god) under Tiglath-pileser III, embarked on a mission of conquest, captured Israel, and deported much of the population to upper Mesopotamia (i.e. northwestern Iraq, northeastern Syria and southeastern Turkey) including their capital of Nineveh (location f on map figure A4) (Dubovsky (2006)). This is known as the exile of the ten northern tribes, and is the first large scale spread of the Jewish people post-statehood.

Josiah recovered some territories that Assyria had taken after Assyria was weakened in a conflict with the Babylonians - one that they would eventually lose (Rowton (1951)). Once Babylonia controlled Mesopotamia, their king, Nebuchadnezzar, moved south towards Egypt, and took control of the entire region that had previously been help by the Israelites (Pearce and Wunsch (2014)). Jerusalem was burned to the ground, and the people were exiled to Babylonia (Henze (1999)). There they established important religious institutions, and Babylon (location g on map figure A4) became the centre of Jewish life for the next 1100 years (DellaPergola (1997)). This second exile to Mesopotamia is known as the first Jewish Diaspora.

The Babylonian Empire was conquered by Cyrus the Persian in 539 BCE, who established the province of Judea as part of the Persian empire, allowed the practice of Judaism, and also allowed exiles to return to the region that was formerly Judah and Israel (Kuhrt (2007)). Meanwhile, "The Judeans of Babylonia continued to feel connected to the people of Judea by history, family ties, culture and religion, and they remained organized as a distinctive ethnic and religious group" (Scheindlin (1998), p. 28). Babylonian Jews could no longer be considered exiles as they were free to return to Jerusalem if they wanted (of course though, it had been burned to the ground) however, they remained as the longest-lasting diaspora in Jewish history. Indeed, the Jewish community in Iraq remained a presence continually up until 1951. "It is from this period that it becomes appropriate to begin speaking of the Jewish people, meaning all of those who, throughout history and around the globe, have regarded themselves as linked to one another and to the people of the ancient Israelite kingdom, either by ethnicity, culture, intellectual heritage, or religion" Scheindlin 1998, p. 28.

One of the ways that Jewish identity remained despite the lack of common political institutions, geography, language, etc. was the Torah (Newman (2020)). The Torah is said to have been given to Moses at Mount Sinai, but had not been a prominent part of Jewish life until it was promoted heavily by Judean elders in Babylonia (Scheindlin (1998)). Historians believe that, in fact, this was the time at which the Torah was compiled, in order to develop a Jewish identity, codifying religious practices in the absence of any national institutions (Kalmin (2006)). Even more certain is that Babylonian exile sparked the writing of the Talmud - codifying Jewish laws - and perhaps the core canonical Jewish text (Ilan (2009); Neusner (1970)).

The promotion of the Torah / Talmud marks the beginning of an attempt to geo-

graphically diffuse the core elements of Judaism. For instance, during the last half of the fifth century BCE, the Torah was instituted as the official law of Judea. On the basis of this influence from the Babylonian Jewish Diaspora, Judea became theocratic until Alexander the Great conquered the Persian empire.

There are two potentially relevant locations that may geo-locate the origins of the spread of Judaism. Which is more relevant depends on whether we think of the Jewish homeland, or the origin of scripture as being the dominant mode of spread. Judaism would have spread by word of mouth, emanating from the homeland, in Jerusalem. However, scripture was written in Babylon, so if scripture is the dominant mode of global spread, then this should be the appropriate origin location. Our estimates favour scripture as a dominant mode of global spread, for both Judaism and Christianity, so we place the origins of the global spread of Judaism in Babylon, the capital of ancient Babylonia.

A.5. Christianity

The earliest followers of Jesus formed a Jewish sect living in Judea, the land that had previously been the Kingdom of Judah, prior to Greek and then Roman control Freeman 2009, p.19. Jesus and his followers lived in Jerusalem, the historical centre of Judah under David, and at the time under the control of the Roman client king Herod Hill 2020, p.30-31. Following Jesus' death, the first Christian community was established in Jerusalem (map location a in figure A5).Ludlow 2009, p.14 At the time, followers were a combination of Hebrews, speaking Aramaic and Greek, and Hellenists, who spoke only Greek Freeman 2009, p.44.

Aramaic was the *lingua franca* of the coastal Levantine region, and Jesus' own original language for his teachings Freeman 2009, p.21. That said, Greek was "the *lingua franca* of the Middle East in the times of Jesus, and it was the language in which, in a rather vulgar marketplace form, most Christians spoke in everyday life during the Church's first two centuries. By the sixth and seventh centuries, Greek was ousting Latin as the official surviving language of the Eastern Roman Empire, with the strong encouragement of the Christian Church (MacCulloch (2010), p. 43). It was likely this interaction of the original disciples, who primarily spoke Aramaic, with Greek-speaking followers in Jerusalem that led to teachings being translated from Aramaic to Greek Freeman 2009, p.22. Indeed, most core Christian words stem from Greek.

> To the very ordinary Jewish name of this man, Joshua/Yeshua (which also ended up in a Greek form, ('Jesus'), his followers added '*Christos*' as a second name, after he had been executed on a cross. It is notable that they felt it necessary to make this Greek translation of a Hebrew word 'Messiah,' or 'Annointed One,' when they sought to describe the special, foreordained character of their Joshua. (MacCulloch (2010), p. 19)

Similarly,

> When Christians first described their own collective identity, with its customs, structures and officer-bearers, they used the Greek word *ekklēsia*, which has passed hardly modified into Latin and its successor languages...*Ekklēsia* is already common in the Greek New Testament: there it means 'Church,' but it is borrowed from Greek political vocabulary, where it signified the assembly of citizens of the *polis* who met to make decisions. (MacCulloch (2010), p. 26)

The early community of Jesus' followers maintained strong continuity with Jewish thought, and early commemoration of Jesus as divine (rather than as another prophet) led to tensions with the larger Jewish community in Jerusalem Freeman 2009, p.39-43. These tensions became especially clear over the acceptance of Gentiles (non-Jews) into the Christian community in major centres of the Eastern Roman Empire, such as Antioch Hill 2020, p.39-40. Missionary activity was prevalent in this period (33-100 CE), with early centres of Christianity being established primarily in the Greek speaking eastern-half of the Roman empire Ludlow 2009, p.24. While some early gospel was in Aramaic (including Jesus' teachings) Freeman 2009, p.21 by far the majority of Christian writings was in Greek (Jaeger (1985)).

Christianity continued to grow throughout the first and second centuries, but remained quite fractionalized in beliefs as well as geographically Ludlow 2009, p.34. One source of unity for the early church was the common belief in the Gospel of Matthew, Mark, Luke and John Hill 2020, p.63.[51] Beyond the Four Evangelists, the early spread of Christianity is largely attributed to Paul of Tarsus (on the south coast of modern Turkey), who (unlike Jesus) did not speak Aramaic, but was rather a native Greek speaker Harrill 2012, p.24. He travelled throughout the Mediterranean, and had a profound influence on beliefs, mostly among gentiles in urban Greek centres Hill 2020, p.40 and later in Rome MacCulloch 2010, p.225. His importance to the spread of Christianity is unquestioned: "Paul dominates any history of early Christianity.He is the loner who made Christianity universal,the authoritarian who wrote in terms of the equality of all before God." Freeman 2009, p.47 despite not having met Jesus himself Freeman 2009, p.50. The gospels are thought to have been written about 50 years after Jesus' death.Hill 2020, p.42 They began being quoted in Christian texts around 200 CE at the latest MacCulloch 2010, p.197. They form the beginning of the New Testament, which is largely followed by letters written by Paul Senior 2022, p.11-12.

The earliest known complete compilation of the New Testament was written by Athanasius of Alexandria (location b in map figure A5) in 367 CE (Lindberg (2009)). By

[51] Incidentally, the etymology of the word Gospel comes from the Old English word Godspell, for 'good news,' which in turn was a literal translation of the Greek word for good news *evaggelion* MacCulloch 2010, p.182.

the end of the fourth century, the New Testament had become Christian canon (following a number of councils in North Africa), and this development is crucial in the spread of Christianity Senior 2022, p.76-91. In fact, "[i]f we seek one explanation of why 'Catholic' Christianity so successfully elbowed aside both the gnostic alternatives and the tidy-mindedness of Maricon, it is to its sacred literature that we should point" (MacCulloch (2010), p. 128).

When we look towards the origins of the spread of Christianity, the influence of Paul in the Mediterranean and the establishment of the New Testament are the key factors. Since Paul spoke Greek, was influential in urban Greek centres, and wrote much of the New Testament, which is in Greek, it seems sensible to locate the origins of the spread of Christianity in the Greek-speaking Eastern Roman Empire. As such, North Africa, and especially Alexandria was a crucial centre of Christianity since it was where the New Testament was first compiled. Other possible origins include Constantinople, where Christianity was institutionalized following the move of the capital from Rome, and of course, the origin of Christianity itself was clearly in Judea. Spread via word of mouth would have emanated from here. Overall, anywhere in the eastern portion of the Byzantine empire seems like a sensible choice, but following the theme of favouring scripture and standardized body of thought as a vehicle of spread, we place the origin of Christian religious spread in Alexandria.

Figure A5: Important Sites in Spread of Christianity

Note: The map shows historical sites in the origin and spread of Christianity. In particular, we highlight Alexandria, where the Old Testament was first compiled as religious scripture. We also highlight Jerusalem, which was the principal setting for the life and teaching of Jesus and where most disciples lived.
Map Source: Tanner (1826), accessed through the David Rumsey Map Collection, David Rumsey Map Center, Stanford Libraries.. Author's own highlighting of key sites.

Appendix B. Supplementary Information: Materials and Methods

B.1. Language Data

This section describes the construction of the language data. A separate paper builds on this data set as well (Blouin and Dyer (2022)). Accordingly, that paper borrows heavily from the data description below, and there is overlap in the descriptions of methods so that both articles can be self-contained.

i) Semantic Similarity Routine To capture religious loanwords, we use a seed word approach (starting from a small number of words in English). From these seed words, we algorithmically find all related words using a routine based on semantic models from nearly three hundred languages, where we first translate the seed word into native words in each of the three hundred languages then identify closely associated words. This process is outlined in figure B2 and shows how we go beyond translations from the English seed words. This means we do not only use associations from English, but instead include associations of meanings representing a broad geographic and cultural range. From this expanded set of words we can then construct the share that were borrowed.

To implement the seed words routine, we started with a list of seed words.[52] These words represent the concepts, people and places of worship in the major religions we are aiming to represent. These words were deliberately selected to cover religious concepts, without prioritising the means of religious spread or specifically including religious texts.

One priority was to choose seed words in as hands-off a way as possible. Accordingly, to guide the seed words we should select, we started from the Library of Congress Classifications system, focusing on Subclass BL (Codes BL1-2790, Religions, Mythology, and Rationalism) as in table B1 below.[53] While there may be another list of seed words that is more appropriate for this exact context, our primary concern was transparency. One option would have been to compile our own list of seed words tailored to the context, but this would leave a large degree of methodological freedom to search over plausible lists until the desired result is obtained. We therefore use the Library of Congress Classifications as our source of seed words, as it is a reasonably objective and widely-known classification system. Again, as described in section B.1.1 and figure B2, we expand this list using semantic associations from nearly three hundred languages, so we are not only including the words selected by librarians in a specific context. This restricts the potential for finding spurious results through iterative search over seed word lists.

We then remove headings related to Mythology, Rationalism and the study or classi-

[52] These seed words are: religion, god, priest, afterlife, spirit, pray, worship, sacred, church, temple, mosque. We translate these into nearly three hundred languages and use semantic models trained for each of these languages (coverage a broad range of the world) to identify associated words so that this is not overly biased towards English words or word-associations. See appendix B.1.1 for further details.

[53] Original classification schema sourced from https://www.loc.gov/aba/cataloging/classification/lcco/lcco_b.pdf.

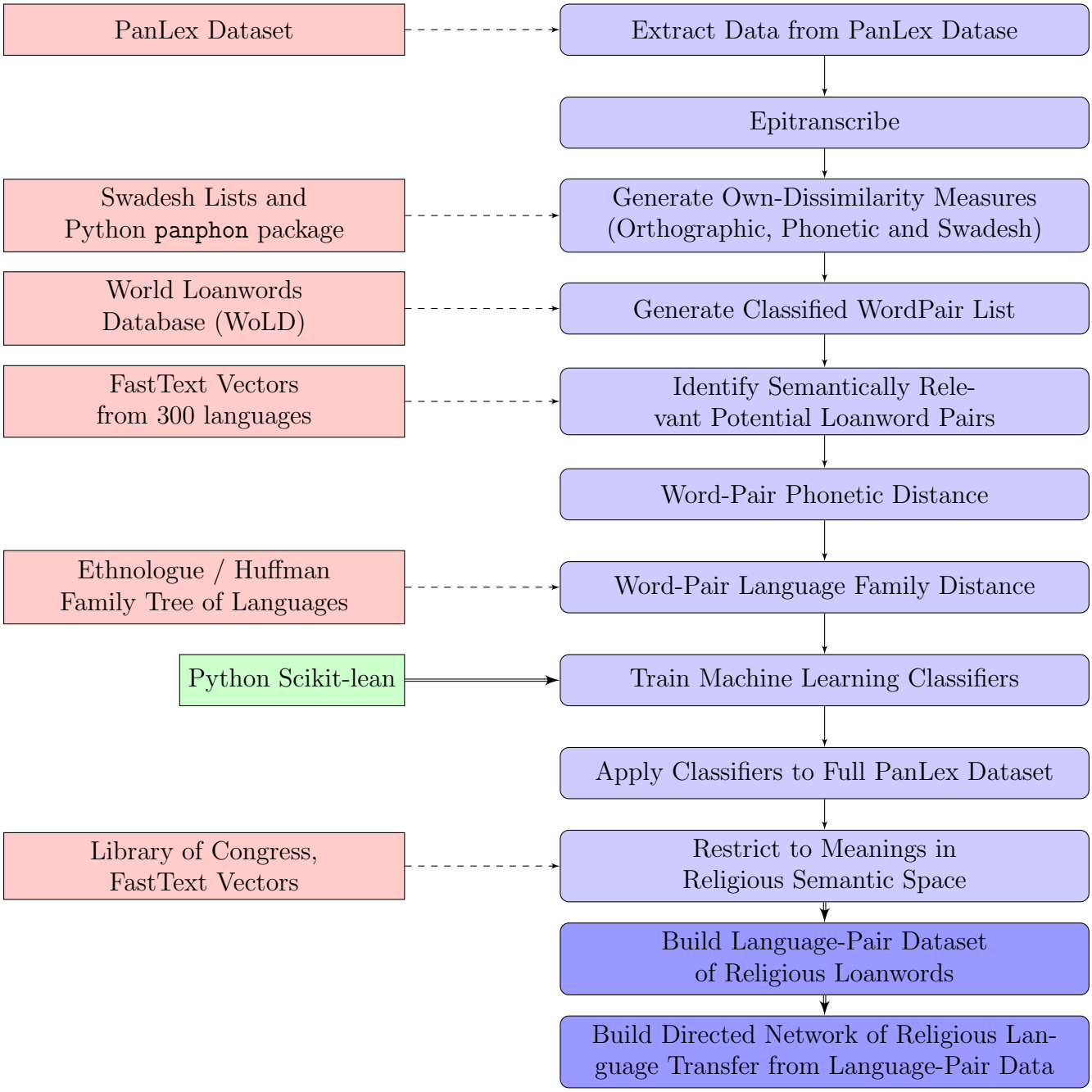


Figure B1: Machine Learning Dataset & Processing flow-chart

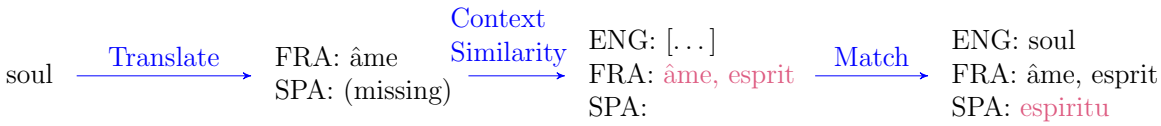


Figure B2: Semantic analysis illustration: capturing semantically similar words

Note: This figure illustrates the way that we go from seed words in English to words in other languages, using word-associations in many languages. We show an illustrative example. Consider the (hypothetical) case where the English word *soul* does not have a translation into Spanish. We translate the word *soul* into French, then look at words in the French semantic model that are similar to *âme*, and identify the word *esprit*. We then translate the word *esprit* into the Spanish *espiritu*. This example illustrates how we are able to match to words that cannot be directly translated from English, and going beyond word-associations in English.

fication of religions. We also remove headings related to the history of specific religions, and specific religious doctrines. We also drop any references to technical classification words such as *General*, as shown in table B2.

From the remaining words we then derive our list of seed words, replacing the esoteric terms with those more likely to be found in used language, and those that are least likely to have non-religious connotations, as described with justifications for each choice in table 2.

With these seed-words in hand, we implement a well-established semantic analysis routine based on word vectors trained on Wikipedia (see Bojanowski et al. (2017)) for two hundred and ninety-four languages. The intuition is to represent words as vector values in a 300-dimensional vector space, where each of these dimensions is intuitively related to a 'feature' that captures the relationship between two words.[54] Representing (or *embedding*) words as vectors in this space (as first developed by Mikolov, Sutskever, et al. (2013)) leads to words that occur in similar contexts being closer together. The advantage of the approach in Bojanowski et al. (2017) is to consider sub-word letter combinations as well as entire words, making it feasible for languages with many words, and for words that do not occur often in the training data. This routine first finds direct translations of the seed-words (i.e. with identical meaning identifiers in PanLex) among the covered languages from the meaning identifiers in PanLex, the main database of words we used. To consider a broad range of associations, we consider two meanings to be similar if their word-vector representations are similar to a seed word or its direct translation in any of the covered languages. We take these 'similar meanings' and again translate the expanded word-set using the PanLex meaning IDs, to get a large list of words in each language that are related to our various topics.

In table B3 we present the most frequent English words associated with the meanings

[54] A common example of what these dimensions represent is described in Mikolov, Yih, and Zweig (2013): *we examine the vector-space word representations that are implicitly learned by the input-layer weights. We find that these representations are surprisingly good at capturing syntactic and semantic regularities in language, and that each relationship is characterized by a relation-specific vector offset. This allows vector-oriented reasoning based on the offsets between words. For example, the male/female relationship is automatically learned, and with the induced vector representations, "King - Man + Woman" results in a vector very close to "Queen."*

Table B1: Unfiltered LCCO Schema BL1-2790: Religions, Mythology, and Rationalism

LCCO Code	Heading	Sub-Heading
BL1-50	Religion General	
BL51-65	Philosophy of religion. Psychology of religion.	
BL70-71	Sacred books General	
BL71.5-73	Biography	
BL74-99	Religions of the world	
BL175-265	Natural theology	
BL175-190		General
BL200		Theism
BL205-216		Nature and attributes of Deity
BL217		Polytheism
BL218		Dualism
BL220		Pantheism
BL221		Monotheism
BL224-227		Creation. Theory of the earth
BL239-265		Religion and science
BL270	Unity and plurality	
BL290	The soul	
BL300-325	The myth. Comparative mythology	
BL350-385	Classification of religions	
BL410	Religions in relation to one another	
BL425-490	Religious doctrines General	
BL430		Origins of religion
BL435-457		Nature worship
BL458		Women in comparative religion
BL460		Sex worship. Phallicism
BL465-470		Worship of human beings
BL473-490		Other
BL500-547	Eschatology	
BL550-619	Worship. Cultus	
BL624-629.5	Religious life	
BL630-632.5	Religious organization	
BL660-2680	History and principles of religions	
BL660		Indo-European. Aryan
BL685		Ural-Altaic
BL687		Mediterranean region
BL689-980		European. Occidental
BL1000-2370		Asian. Oriental
BL2390-2490		African
BL2500-2592		American
BL2600-2630		Pacific Ocean islands. Oceania
BL2670		Arctic regions
BL2700-2790	Rationalism	

Note: This table shows the original classification schema from the Library of Congress Classification system (sourced: https://www.loc.gov/aba/cataloging/classification/lcco/lcco_b.pdf) that we used as the basis for our list of religion seed words.

identified as related to our religious seed-words, ordered by frequency. Of these, only one (*September)* appears unrelated to religion, and moves beyond the original wordlist in a way that appears to capture a broad understanding of religion. In addition, in table C4

Table B2: Filtered LCC Schema BL1-2790: Religions, Mythology, and Rationalism

LCCO Code	Heading	Sub-Heading
BL1-50	Religion	
BL70-71	Sacred books	
BL175-265	Natural theology	
BL200		Theism
BL205-216		Nature and attributes of Deity
BL217		Polytheism
BL218		Dualism
BL220		Pantheism
BL221		Monotheism
BL224-227		Creation. Theory of the earth
BL239-265		Religion and science
BL290	The soul	
BL500-547	Eschatology	
BL550-619	Worship. Cultus	
BL624-629.5	Religious life	
BL630-632.5	Religious organization	

Note: This table shows the classification schema after removing headings related to the study of religions, mythology, rationalism, and specific locations or doctrines. We also drop any mentions of the word "general".

we plot histograms of the wordcounts of words identified as related to religion, as well as words of any type in figure C3.

There are a number of important advantages to doing this. The first one is that it allows for broader coverage. Some of the languages in PanLex have more coverage than others, and expanding the set of words that we examine increases the odds that one or more of them is included in the less heavily documented languages.

Second, we think it is important not to narrow-in too closely on the loanwords data. Our intention was to develop a way to examine *global* patterns in language transmission. Rather than getting into the process of defending the loanword status of specific word pairs - which is the focus of linguists[55] - our approach is to acknowledge that any automated approach will come with error, and we should accordingly manage that error to the best of our ability. One way of doing this is by exploring averages of larger sub-samples, whenever possible.

ii) Loanword Classification Details: Classification of likely loanwords was done using Random Forest and Extremely Randomized Forest. To choose hyper-parameters we used a grid search method over the number of features available at each split of the decision tree; the maximum depth of the decision tree; and the minimum number of observations

[55]We view our approach as complementary to the work that linguists do. It is certainly not a substitute, since we cannot claim with anywhere near the same level of certainty that any particular word pair is, or is not, a loanword pair.

Table B3: Semantic Routine - English Words

Expression
acclaimed
instruct
synagogue
spirit
see
priest
audacity
exertion
god
elohim
devi
religion
worship
temple
astrology
stamina
solicit
shrines
church
sacrifice
sacred
afterlife
preoccupied
pray
mosque
mind
demigod
idolise
holy
september

Note: This table shows the English expressions associated with additional meaning identified from the semantic similarity routine beginning with the seed words.

in each final leaf. We select the parameters that performed best on different folds of the training set.

Of primary importance is the features that the machine learning classifier used to predict loanwords. One of those factors is the linguistic distance between the languages, to allow the algorithm to rule out cognates. We use the share of of overlapping nodes in the language trees to measure this. Next, linguists typically consider a variety of factors related to how similar a potential loanword (called a *target word*) is to the the potential source of the loanword (i.e. the *source word*), and how similar the source and target words are to the typical words in each language. This is helps to determine how likely it is that

one word originated from the other, and how likely it is that each word originated within their own language. Accordingly, we include a number of measures that indicate the linguistic similarity of both the target word and the source word to their own respective languages; and features that measure the similarity of the potential target and source words to each other. Lastly, we include the *difference* between these two measures - i.e. between the own-language similarity of the source and target words.

Each of the similarity measures that we include are based on either orthographic similarity, or phonetic similarity. We use a number of orthographic similarity measures, all based on work by Jaro (1989) and Winkler (1990). These measures take values ranging from 0 to 1, with 1 indicating identical spellings. Accounting for phonetic similarity is more complicated because not all phonetic differences are considered equal signals of loanword status. For example, in English moving from the IPA phoneme ɓ to p is a more natural and common slip than moving from ɓ to ŋ. To address this we follow Mortensen et al. (2016). The intuition is to construct a weighted-distance between sounds, where larger weights are assigned to differences in sounds that are less likely to evolve over time. This accounts for the possibility that a word was borrowed but has evolved over time. For example, a sound's sonority is unlikely to drift and is given a high weight while the length of a vowel is more likely to drift and is given a low weight.

Further complicating matters is that some phonemes are more common in some languages than others. To address this we constructed all 2- and 3-gram phonemes that exist in a language, based on the IPA transcriptions. From here it is straightforward to compute the likelihood that a particular word - represented by an n-gram phoneme - is native to a particular language. This provides a measure for the likelihood that the sound of any given word is native to any given language.

Using these classifiers and features, we implemented an ensemble Voting Classifier that predicts the likelihood that any word-pair represents a loanword in a particular direction. We then applied this to all potential word-pair combinations that were in the same semantic space, and are therefore reasonable candidates. To allow for drift in meaning and usage, we do not only use direct translations and also consider word-pairs where the meanings are above a threshold similarity. To do this, we use all two hundred and ninety-four languages with vector models trained by Bojanowski et al. (2017), which maps words into a vector space where words appearing in the same contexts are closer together.[56] We then identify all pairs of meanings associated with word-vectors whose similarity meets the threshold of 0.70 in at least one of the two hundred and ninety-four languages. This allows the algorithm to consider potential matches of similar meanings, accounting for variation in usage and meaning of words across languages. We use this set

[56]The approach in Bojanowski et al. (2017) is especially appropriate here, as their model consider sub-word letter combinations as well as modelling entire words, which expands coverage of languages with many words that are rare in available training data.

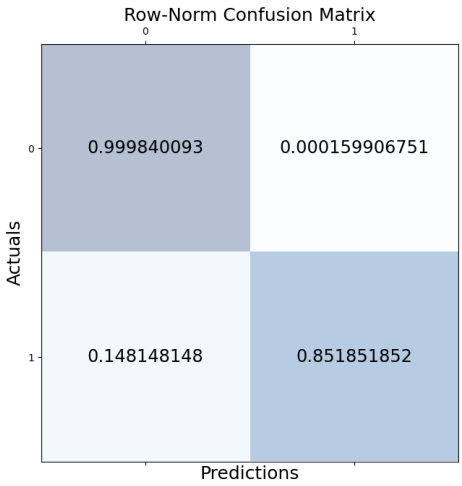


Figure B3: Confusion Matrix

Note: Here we present a confusion matrix normalized by rows (i.e. normalizing over frequency of true loanword classes) to provide a more detailed assessment of the prediction process. Here 0 refers to non-loanword pairs and 1 refers to loanword-pairs. The rows refer to the actual loanword classes, while the columns refer to the predicted classes.

of similar meanings to build our set of potential loanword pairs to which we then apply our classifier algorithm. Whenever two source words were identified for the same loanword, we kept the source word with the highest probability from the second stage classifier. We also drop loanwords where the source word was itself identified as a loanword, so our final measure of language exchange only includes unambiguously identified loanword pairs. We present a confusion matrix by loanword class in figure B3 to illustrate the algorithm's performance on religious words.

B.2. Constructing Eigenvector Centrality

To compute the role of a group within the religious exchange network we compute the directed eigenvector centrality. There are numerous ways of calculating the centrality of a node in a network. Some focus on how many connections a given node has, how far it is from other nodes in the network, a node's importance in connecting other nodes, or more intricate *prestige* measures that factor in the importance of nodes connected to the node of interest. Eigenvector centrality is one example of the latter *prestige* measures of centrality, first proposed by Bonacich (1972).[57] This measure is constructed as follows for a graph $G := (V, E)$ with $|V|$ vertices. First, define:

[57]For further discussion of the different types of centrality measures, see Jackson (2010).

$$\mathbf{A} = \begin{pmatrix} \mathcal{L}_{1,1} & \cdots & \mathcal{L}_{1,j} \\ \mathcal{L}_{2,1} & \cdots & \mathcal{L}_{2,j} \\ \vdots & \ddots & \vdots \\ \mathcal{L}_{i,1} & \cdots & \mathcal{L}_{i,j} \end{pmatrix} \tag{6}$$

As an adjacency matrix where adjacency is defined here by the degree of religious linguistic influence in the pair. This religious influence between a pair is the aggregate religious influence over all religions.[58] The centrality c_j is dependant on the centrality of its neighbours, capturing the idea of *prestige* where a node's importance is based on how much it influences other important nodes.

$$c_j = \frac{1}{\lambda} \sum_{i \in M(j)} c_i \tag{7}$$

Here M is the set of neighbours of j, but in our case, we consider all pairwise connections, each with weight $\mathcal{L}_{ij}$ indicating how much each other node is influenced by node j, which gives

$$c_j = \frac{1}{\lambda} \sum_{i \in V} \mathcal{L}_{ij} c_i \tag{8}$$

where λ is a constant. This generates the familiar eigenvector equation:

$$\lambda \mathbf{x} = \mathbf{A}' \mathbf{c} \tag{9}$$

The vector $\mathbf{c}$ is the eigenvector of adjacency matrix $\mathbf{A}$ with transpose $\mathbf{A}'$. λ is the corresponding eigenvalue. The adjacency matrix, because we have a directed graph of language influence is comprised of the loanwords adopted by i, from j. Figure **??** shows examples of some of the prominent religious language networks.

B.3. K-Means Clustering

Groups are often influenced by multiple sources, and while we are able to estimate the centroid of this influence on an observation-by-observation basis, this estimate is not

[58] Consider a hypothetical example, where group A influenced group B by spreading both Hinduism and Buddhism. In this case, the aggregate religious influence of A on B would include both of these religious influences.

exactly what we set out to do, which was to identify the geographic origins of the spread. We are interested in finding the centroids of the estimates produced in 4.B, which we interpret as approximately the origins of religious influence.[59]

To do this we use k-means clustering, which will pull out of the data a number of clusters of influence. In our case, we know that there are 5 major clusters to be found, which makes the problem much simpler.

The procedure partitions n observations into k sets (S) in a way that minimizes the sum of squares within each set. Formally,

$$argmin_S \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2 \tag{10}$$

In our case μ is the mean of the estimated coordinates assigned to each cluster, and $k = 5$. To arrive at a solution, the algorithm starts from an initial set of starting values for cluster centroids, and assigns all observations to the nearest centroid. After doing this assignment, the mean of all observations assigned to an initial centroid is used to generate an updated centroid. All observations are then reassigned to be matched with the updated centroid that they are closest to. A newer updated set of centroids is generated from the mean of observations under the new assignment. This process of reassignment and reestimation of centroids repeats iteratively until the assignment of observations to centroids does not change.

There are three tuning parameters often used with k-means clustering. One is the similarity or dissimilarity measure. Options include the Euclidian distance; the Canberra distance; measures based on correlation coefficients; angular separation similarity; etc. We use the simplest possible measure, the absolute value distance.

The second choice is the starting values of the cluster centroids. When there are multiple partitions that provide stability this can make a big difference. We specified that 5 origin estimates from the pairwise location estimates be chosen at random. The main alternative to doing this would be to choose the actual origin locations as the starting point. However, since an aim of the empirical exercise is to see how close we can estimate origins of spread in contexts without a wealth of information about the origins, we preferred to avoid using 'actual' information wherever possible. One reason for this is that the method is far less precise than could be achieved using primary sources, making it most useful in cases where there is little to no information about the origins being estimated.

Finally, and least importantly is the number of iterations we allow the algorithm to undergo if it cannot find stability. We allowed a maximum of 10,000, and this was never

[59]The partitioning of centroids into clusters is represented graphically in figure **??**.

a binding constraint, so any value above 10,000 (and likely many values below it as well) would achieve the same result.

Cluster locations are assigned to religions going from east to west. So the cluster with the centroid farthest east is assigned to Buddhism, and so on for Hinduism, Islam, Judaism and Christianity. This same procedure is used to assign clusters of random coordinates to religions.

B.4. Alternative Clustering Algorithms

In addition to the k-means clustering routine described above, we also show robustness to a number of other clustering routines. We first present results using agglomerative hierarchical clustering using the the Ward variance minimization algorithm and a standard Euclidean distance metric. The intuition of this method is to begin with each observation in its own cluster, and iteratively combine them depending on the distance between the existing clusters to combine the most similar clusters in such a way as to minimize within-cluster variance. We also use the Median WPGMC (Weighted Pair Group Method with Centroids) method, which simply computes distance between clusters as the mean Euclidean distance between their centroids. The Median version of this application assigns equal weight to subclusters that were merged in a previous round. when computing a cluster's centroid. The third alternate algorithm we use is the Weighted or WPGMA (Weighted Pair Group Method with Arithmetic Mean) method with Chebyshev distance, which is similar but takes the arithmetic mean of all distances between members of two clusters, rather than the distance between centroids.For the distance between points, the Chebyshev distance takes the maximum of the distance between points for each of their elements:

$$d(u, v) = \max_i |u_i - v_i|$$

We use this range of clustering methods and distance metrics to ensure that our results are robust to various different methods in addition to our primary k-means approach.

B.5. Clustering Evaluation: Miscategorization and Silhouette Score

The silhouette score is an assessment tool that is often used to evaluate k-means clustering algorithms (originally attributable to Rousseeuw (1987)). It measures how 'tight' the clusters are, or how similar observations within a cluster are to each other relative to observations in a different cluster.

Intuitively, it considers the mean distance between observations within a cluster (call this a_i), and then takes the distance between those observations in the next nearest cluster

(call this b_i). For each observation, we can then compute the silhouette score as:

$$s_i = \frac{b_i - a_i}{max(a, b)} \tag{11}$$

The interpretation of this is that a 0 implies that the clusters are as good as randomly assigned - they could just as easily have been assigned to a different cluster. In this case $a_i = b_i$, so observations are no closer to observations in their own cluster than observations in a different cluster. If, on the other hand $s_i = 1$, this means that all observations within a cluster are identical to each other, i.e. $a_i = 0$.

The silhouette score is useful for evaluating our clustering algorithm by allowing us to identify observations that are mis-categorized. Typically $0 < s_i < 1$ for well performing algorithms, but it is possible that the clustering algorithm mis-categorizes observations into the wrong cluster. In this case we could face the situation where an observation is closer to the observations in a neighbouring cluster than it is to observations in its own cluster. For these cases we would observe that $s_1 < 0$, since $b_i < a_i$. To identify the number of clusters with the best performance, we compute the share of mis-categorized observations.

B.6. Possible biases in source datasets

The approach we use is designed to include data from a broad range of language groups in a way that is not biased towards the characteristics of languages to prevent algorithmic bias. Nevertheless, it is possible that features of the original datasets may be biased, and that these biases may carry through to the results if correlated with other features (Kleinberg et al. (2018)), though we take precautions to limit this. In particular, we do not include features that explicitly include information about the geographic characteristics, or relative prestige of languages, etc. This is an important concern for the discussion of language influence, which can often be influenced by other confounding factors. For example, in the case of Swahili:

> *Emphasis on loanwords can be an issue loaded with conflicting emotions in Swahili studies. Some object to it on the grounds that giving undue weight to foreign influences detracts from and belittles the Africanness of Swahili language and culture. Others appear to feel proud about the presence of many loanwords which makes Swahili a peer of such prestigious languages and cultures as Persian, Arabic and Hindi* Schadeberg 2009, p.93.

For this reason, we do not include any features that directly convey any information about specific language families that might be subject to bias in source datasets. We instead measure family distance between languages as a share of the depth of the tree of

language families that is common, so it does not bias against parts of the world where the studies of languages has been systematically different, and hence have different average depth of family tree splits.

While we use the largest possible set of models of semantic similarity to identify possible loanwords, there are possible biases introduced by any algorithmic decision on which meanings are related enough to be potential loanwords. This is, however, a necessary step as semantic similarity is one of the primary factors linguists consider. We also feel that - by considering a broad range of data-driven word associations from hundreds of languages - that our methodology is more flexible and introduced fewer biases than any approach based on translations.

In addition to there being biases potentially introduced by the datasets we combine, there are also potential biases introduced by selective inclusion into those datasets. To control for different levels of language resources available, we control for the size of the PanLex lexicon for a given language ($LexiconSize_i$) in our regressions. This is because the number of words included - and the type of words included - may systematically vary in a way that could potentially impact our results. By directly controlling for this, we address most of these concerns. Furthermore, any language-specific biases in source datasets that mean we are more or less likely to identify loanwords are most likely to impact the share of loanwords overall, rather than making it more or less likely that the pattern of loanword borrowing is shifted in one direction or another.

Appendix C. Additional Evidence

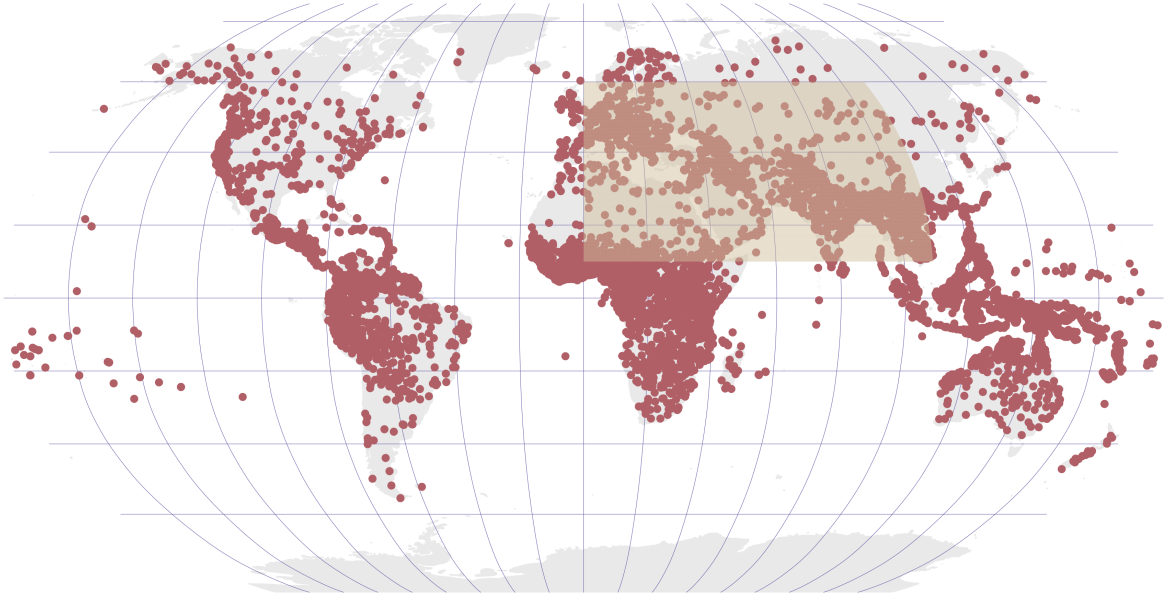


Figure C1: Map of PanLex Language Groups

Note: This map shows each of the borrower and lender languages in the PanLex dataset. The shaded area indicates the area we consider in this paper.

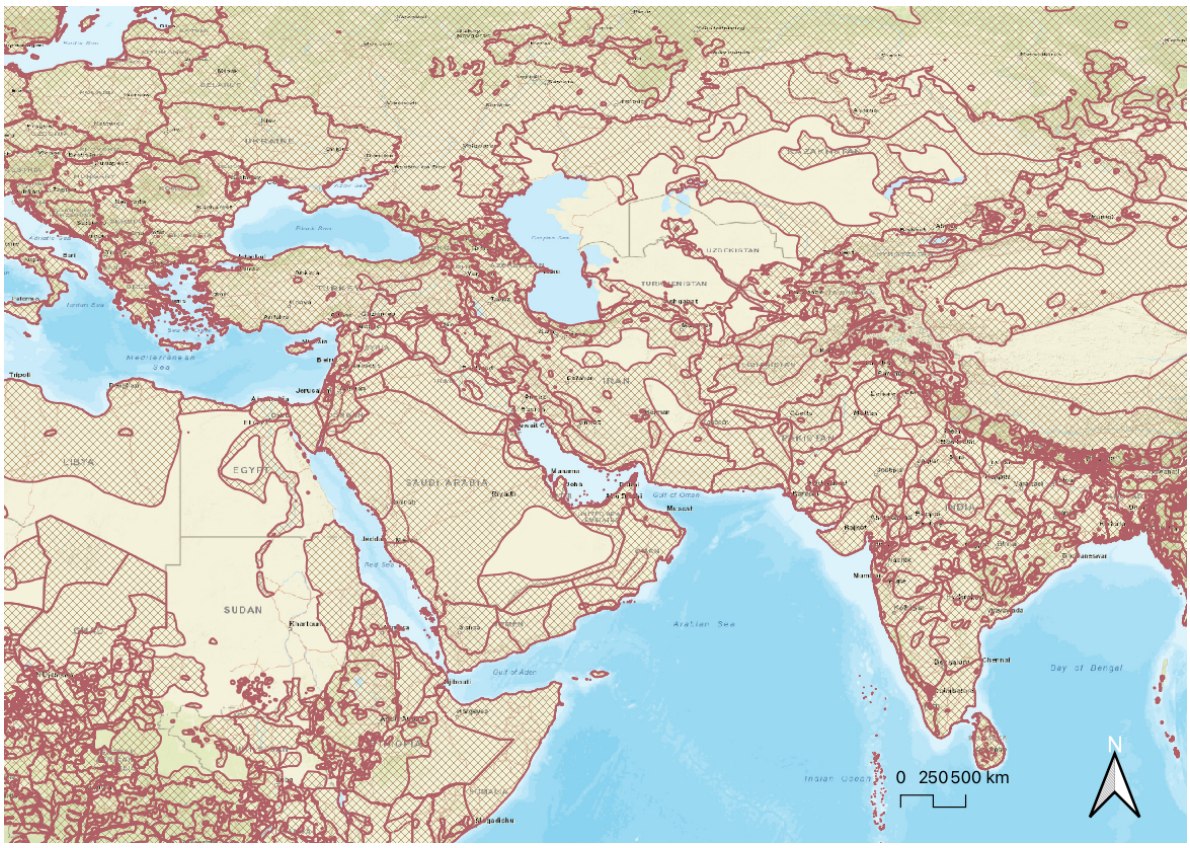


Figure C2: Language Groups in the Ethnologue

Note: The map shows the original boundaries in the Ethnologue. In our data, a language location is a centroid of these boundaries. Notably the center of the boundaries are different from the population hubs, and could even reflect locations where there is little or no population.

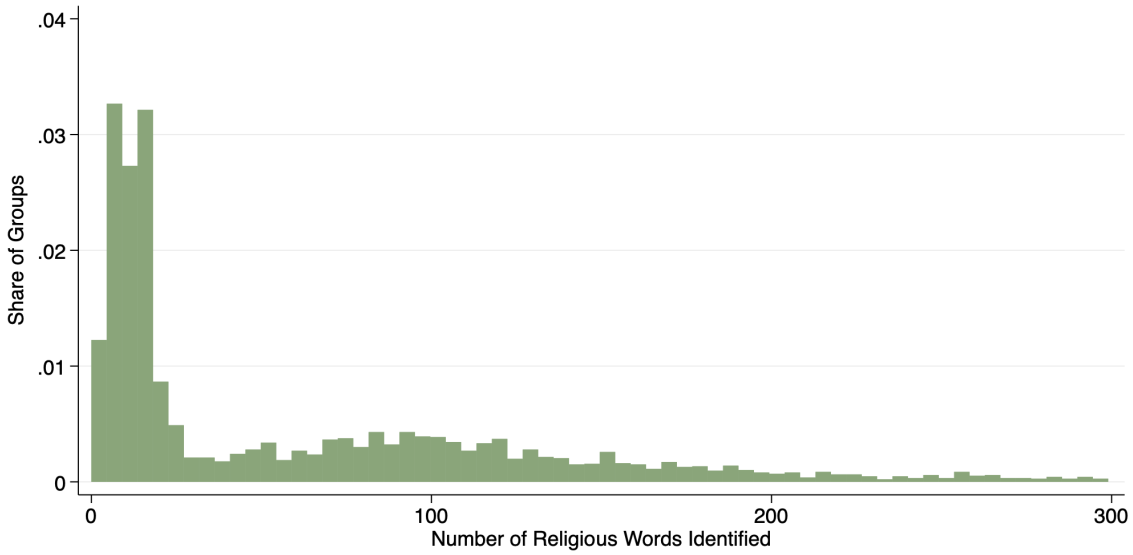


Figure C3: Histogram of Religious Words

Note: The figure shows a histogram of religious words in the data. On the y-axis we plot the share of language groups in each x-axis bin, while the x-axis plots bins of the number of religious words. We see that only about 1% of languages have near-zero words, supporting the claim that there is near complete coverage of language with religious words. The x-axis is censored at 300 to avoid scale distortion. The censoring cuts off 0.0061% of the sample.

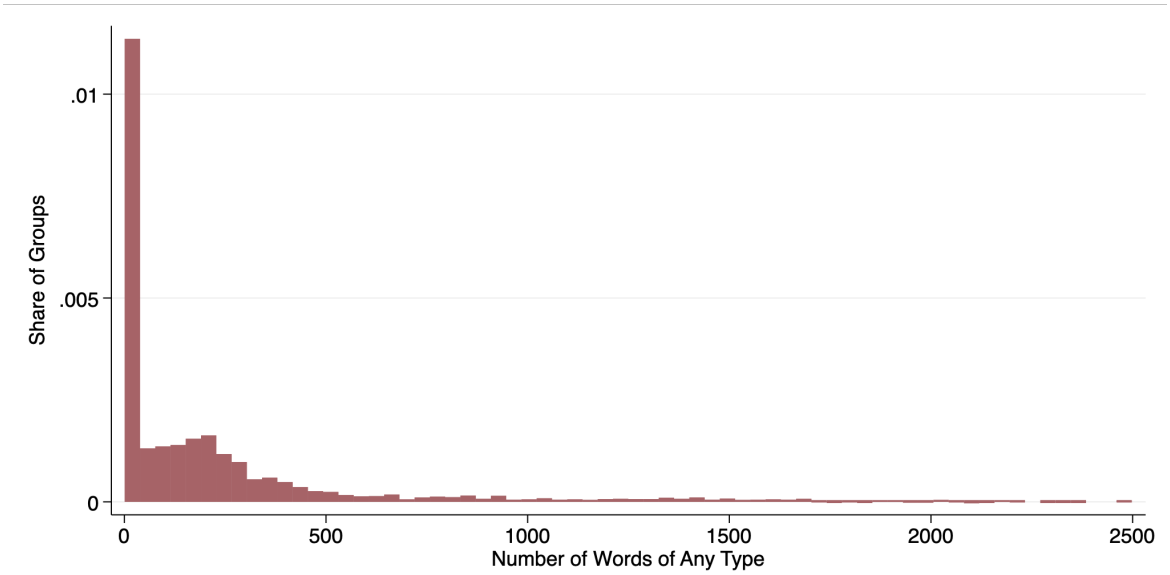
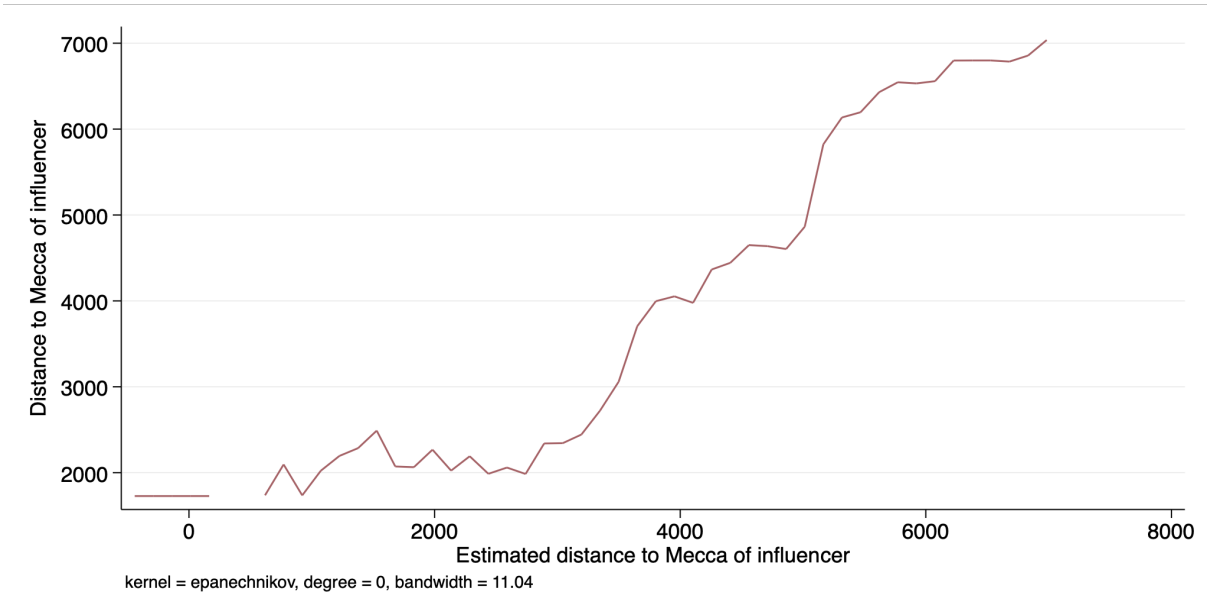
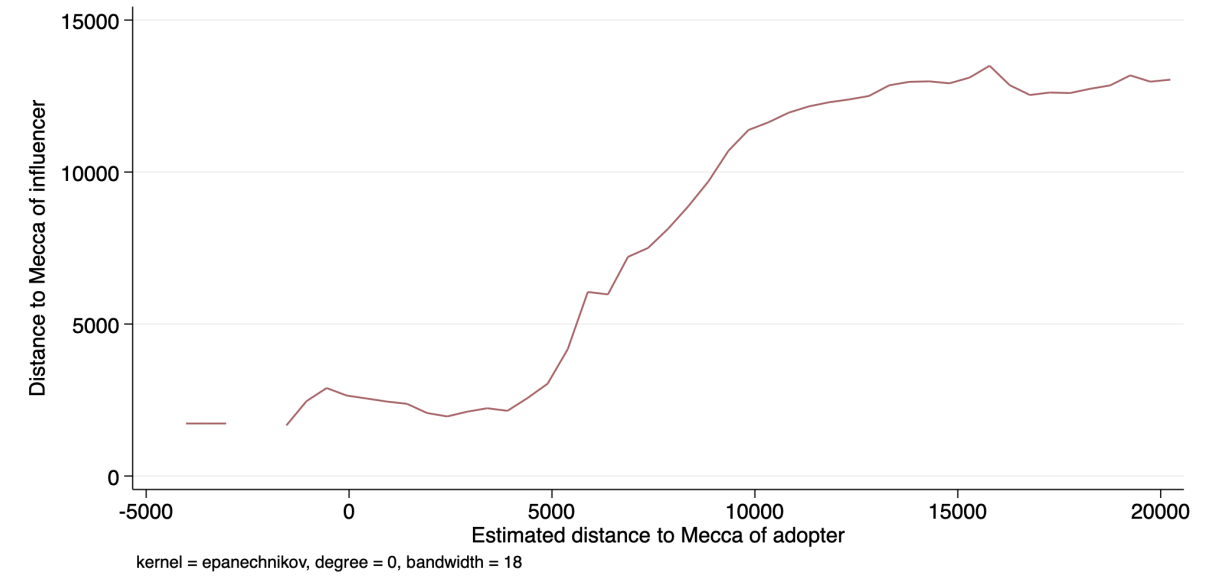


Figure C4: Histogram of All Words

Note: The figure shows a histogram of all words in the data. On the y-axis we plot the share of language groups in each x-axis bin, while the x-axis plots bins of the number of words. We see that only about 1% of languages have near-zero words, reinforcing that there is near complete global coverage in the language data. The x-axis is censored at 2,500 to avoid scale distortion. The censoring cuts off 0.0011% of the sample.



(a) influencer



(b) adopter

Figure C5: Calibration: Actual versus Estimated Distance to Mecca

Note: The figures show the relationship between the estimated distance to Mecca and the actual distance to Mecca for both influencers (panel A) and adopters (panel B). In each case we see a positive relationship, as expected.

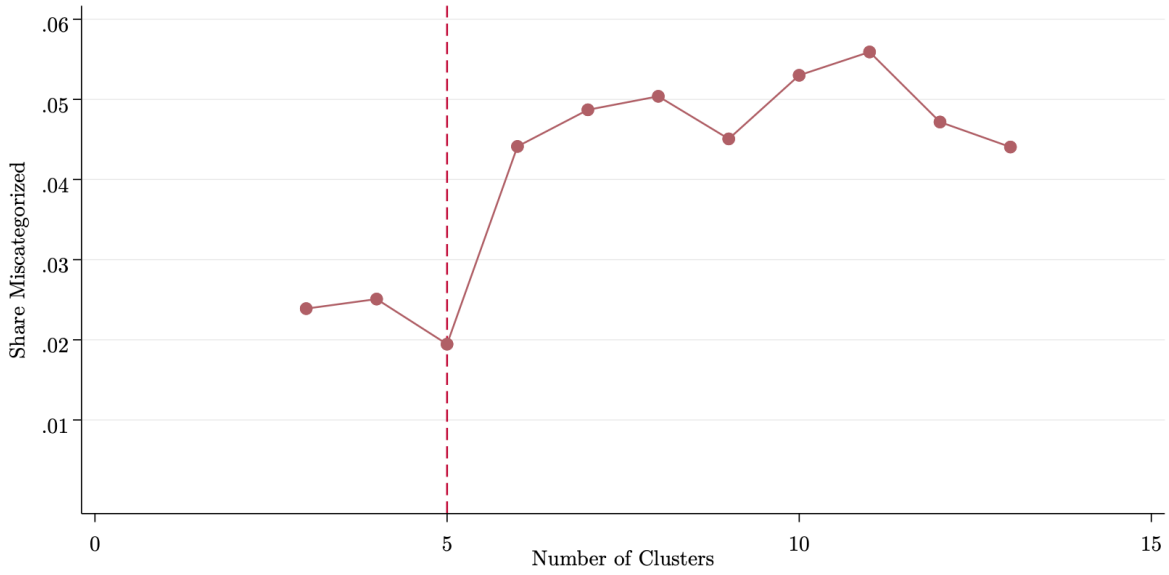
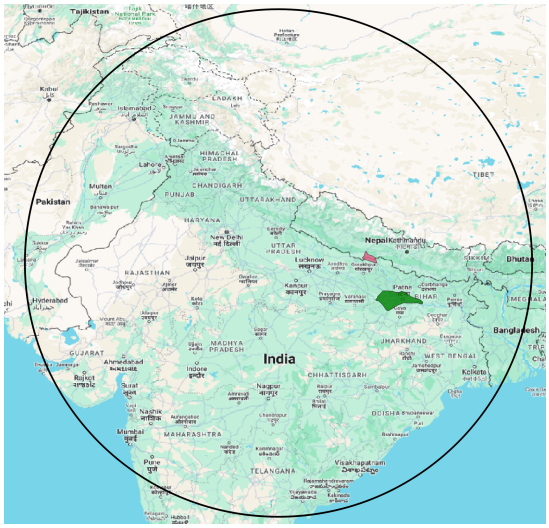
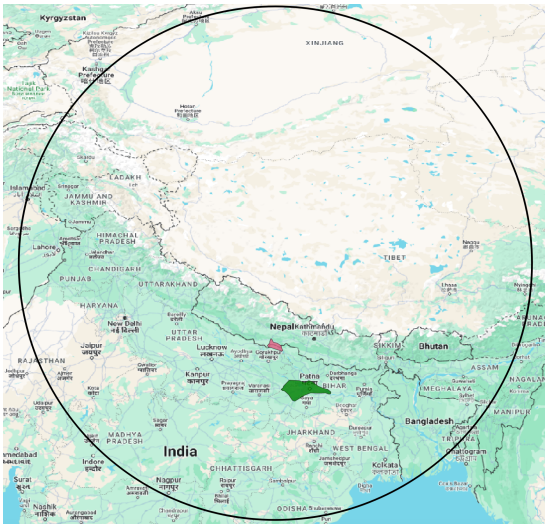


Figure C6: Miscategorized Observations by Num. Clusters

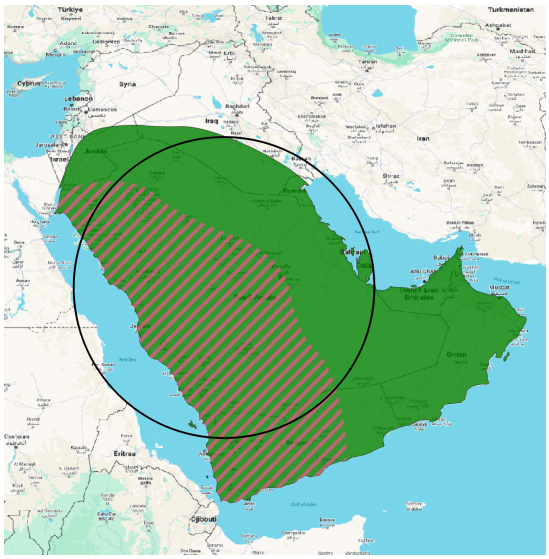
Note: This figure presents the share of mis-categorized observations for each number of clusters, where a mis-categorized observation is one where the silhouette score is less than zero, explained in more detail in B.5. This figure demonstrates that the share of mis-categorized observations is minimized with three clusters.



(a) Buddhism calibrated w. Islam



(b) Buddhism calibrated w. Buddhism

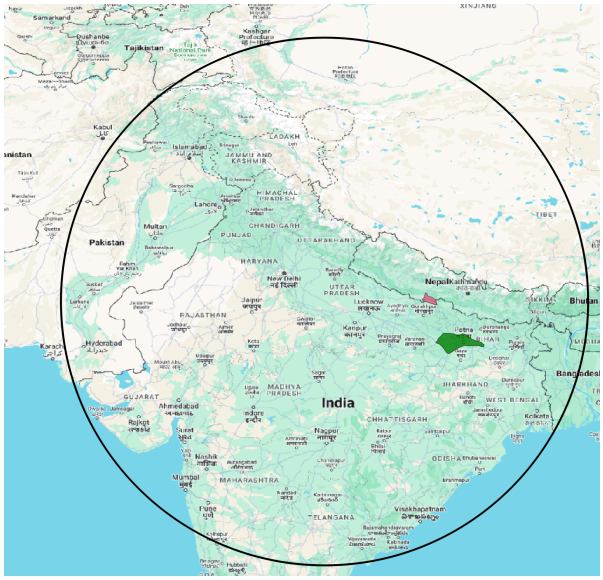


(c) Islam calibrated w. Islam

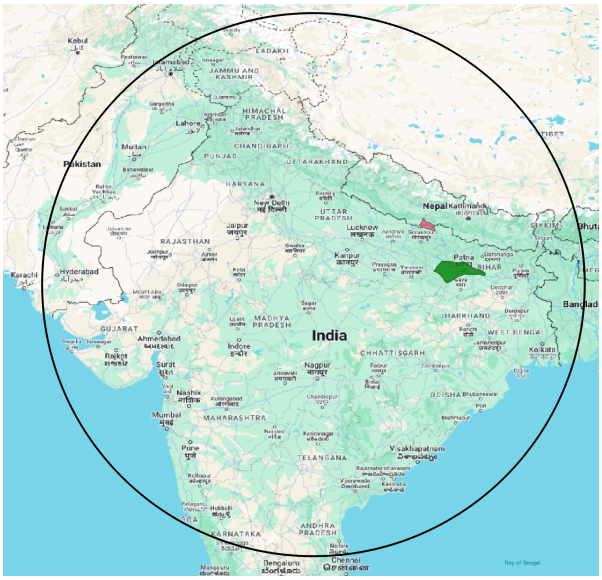
(d) Islam calibrated w. Buddhism

Figure C7: Estimated versus actual origins of Islam and Buddhism

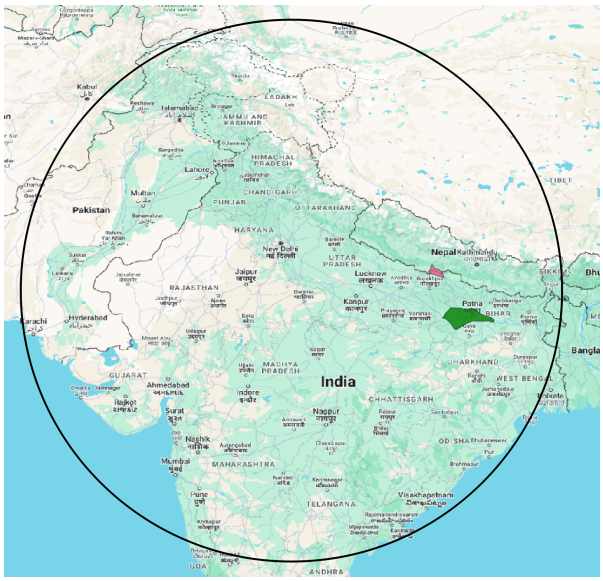
Note: The map shows 75% confidence areas of estimates for both Islam and Buddhism, calibrated with each of Islam and Buddhism. In each subfigure, we present the historical account of the true origins of Buddhism and Islam in pink, with the origins of scripture in green. In the case of Islam and Buddhism these are very similar, and in subfigures (c) and (d) they overlap substantially, but we use this convention throughout the article. In subfigure (a) we show the Buddhism estimate calibrated using the distances to Islam, estimated in table 4. In subfigure (b) we show the Buddhism estimate calibrated using the distances to Buddhism, estimated in table 4. In subfigure (c) we show the Islam estimate calibrated using the distances to Islam, estimated in table 4. In subfigure (d) we show the Islam estimate calibrated using the distances to Buddhism, estimated in table 4.



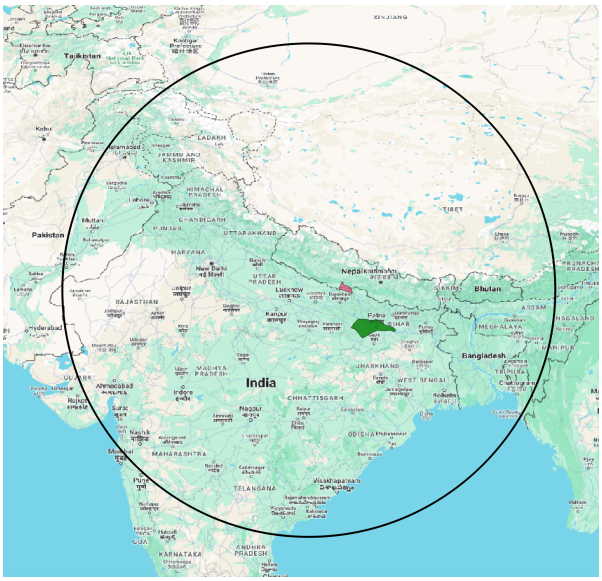
(a) Linear specification



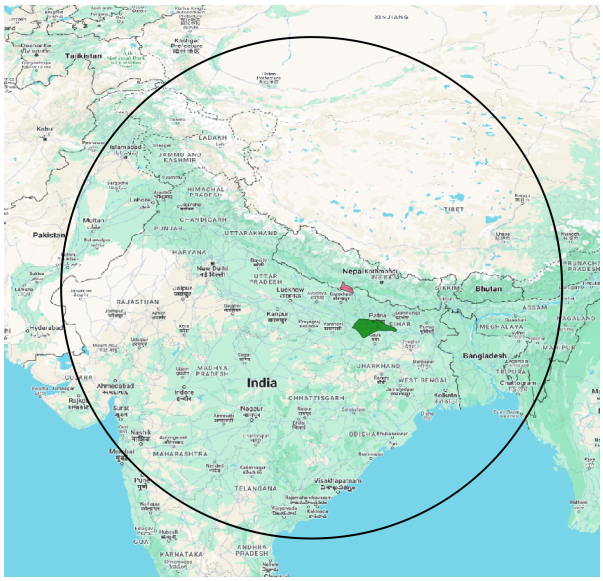
(b) Quadratic specification



(c) linear dependent variable (instead of log)



(d) Betweenness Centrality



(e) Degree Centrality

Figure C8: Buddhism Estimates: Robustness to Alternate Calibration Specifications

Note: This figure shows 5 robustness exercises for the calibration of Buddhism estimates. In each case, the estimates are similar to the main estimates.

(a) Linear specification

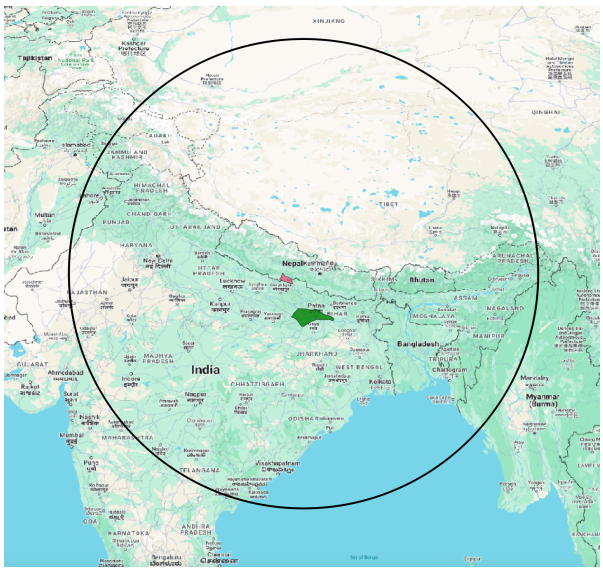
(b) Quadratic specification

(c) linear dependent variable (instead of log)

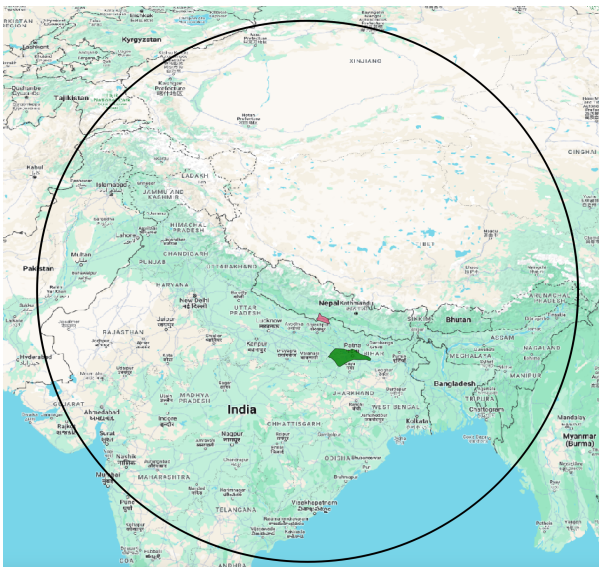
(d) Betweenness Centrality

(e) Degree Centrality

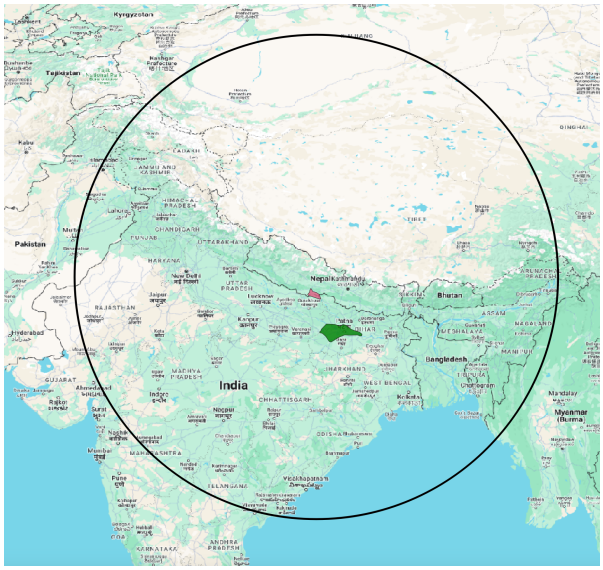
Figure C9: Islam Estimates: Robutsness to Alternate Calibration Specifications



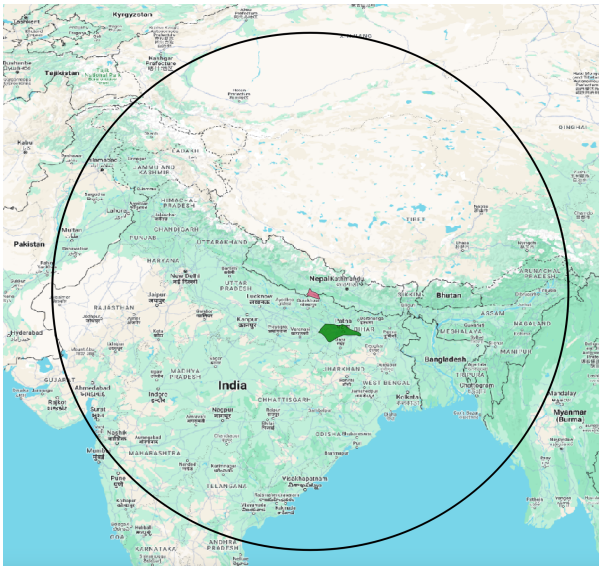
(a) Weighted Euclidian clustering



(b) Ward clustering



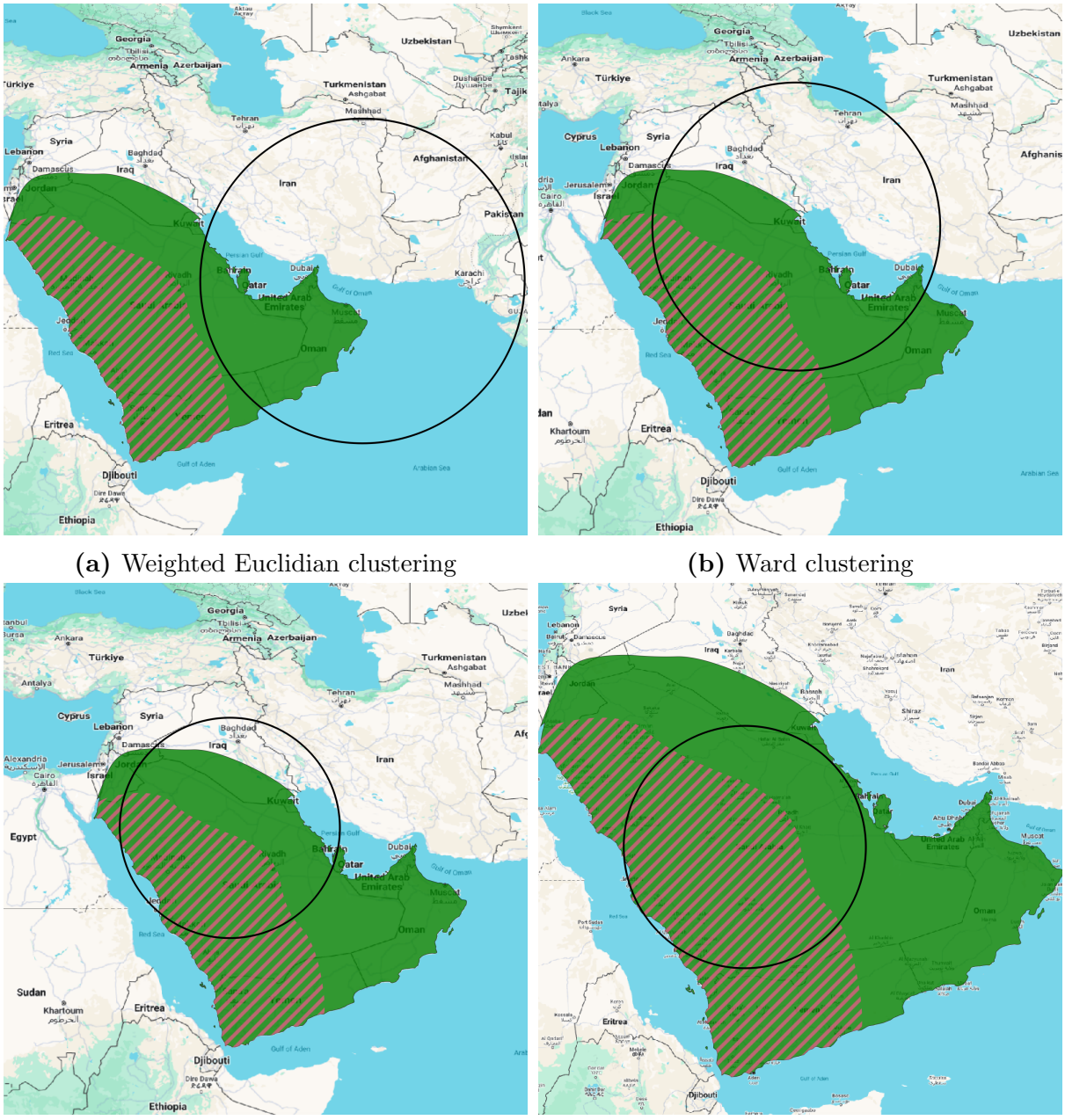
(c) Chebyshev clustering



(d) Median Euclidian Clustering

Figure C10: Buddhism Estimates: Robustness to Alternate Clustering Routines

Note: This figure shows four robustness exercises for the clustering of Buddhism estimates. In each case, the estimates are similar to the main estimates.



(a) Weighted Euclidian clustering

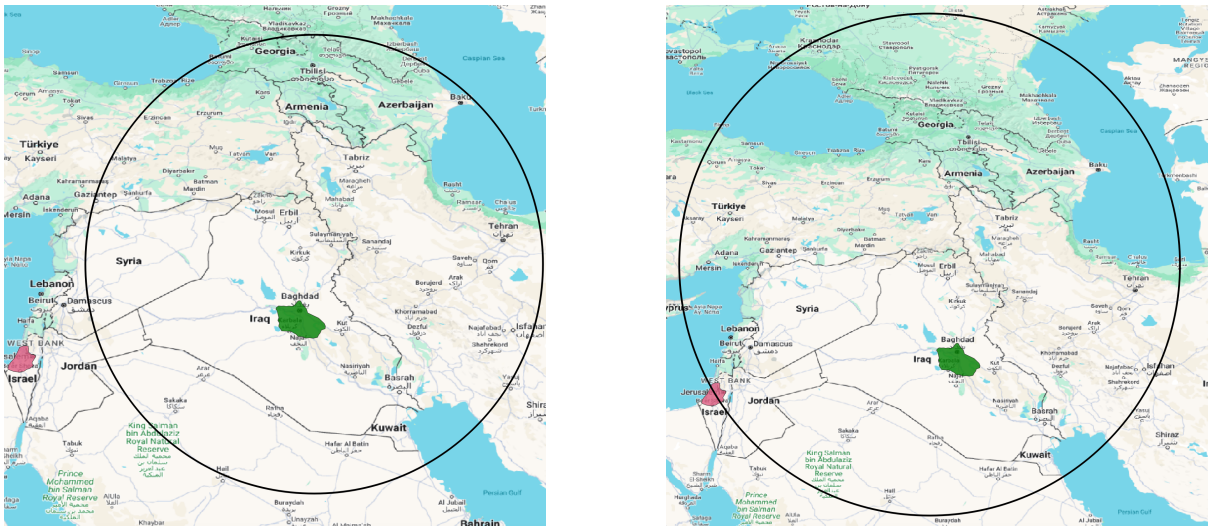
(b) Ward clustering

(c) Chebyshev clustering

(d) Median Euclidian clustering

Figure C11: Islam Estimates: Robustness to Alternate Clustering Routines

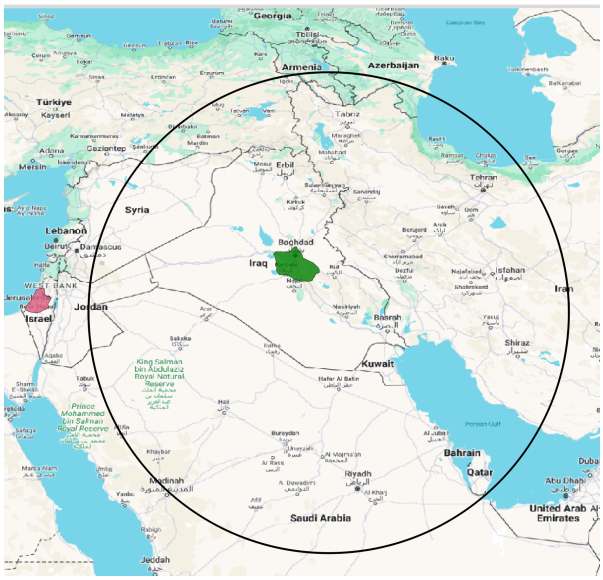
Note: This figure shows four robustness exercises for the clustering of Islam estimates. In each case, the estimates are similar to the main estimates.



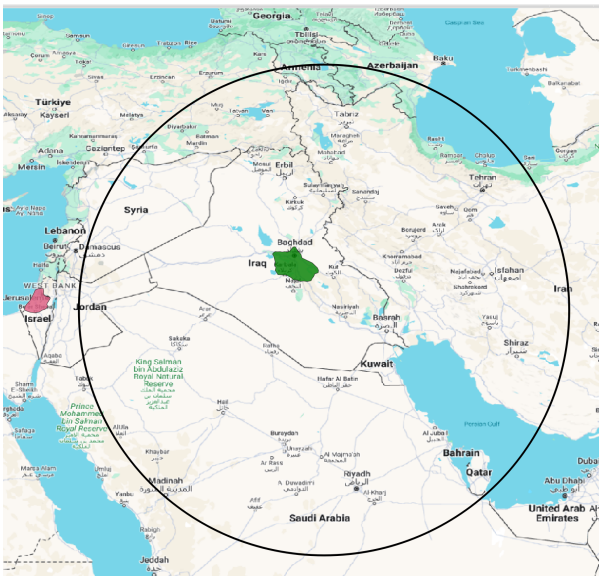
(a) Juddaism calibrated w. Islam **(b)** Juddaism calibrated w. Buddhism

Figure C12: Judaism estimates calibrated using Islam and Buddhism

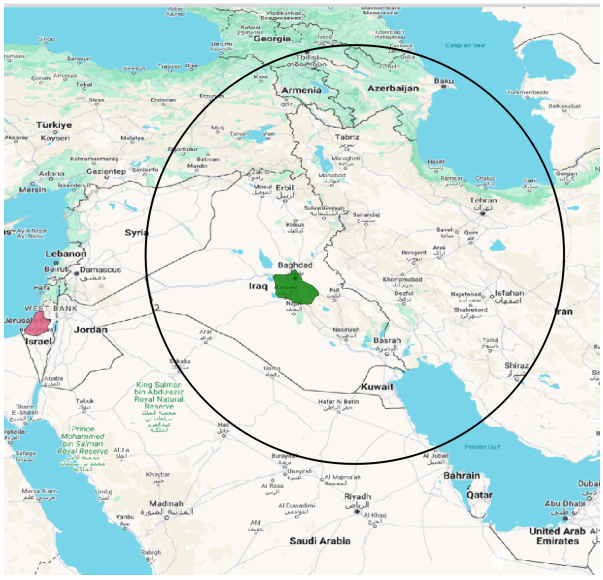
Note: The map shows 75% confidence areas of estimates for Judaism, calibrated with each of Islam and Buddhism. In each subfigure, we present the historical account of the true origins of Buddhism and Islam in pink, with the origins of scripture in green. In subfigure (a) we show the Judaism estimate calibrated using the distances to Islam, estimated in table 4. In subfigure (b) we show the Judaism estimate calibrated using the distances to Buddhism, estimated in table 4.



(a) Linear specification

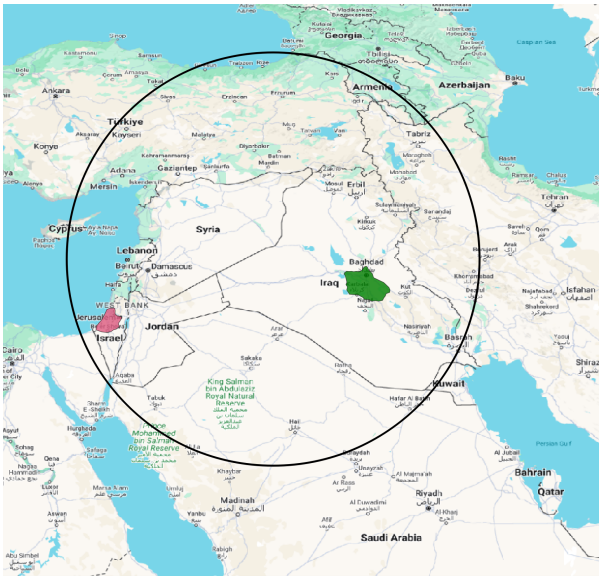


(b) Quadratic specification



(c) linear dependent variable (instead of log)

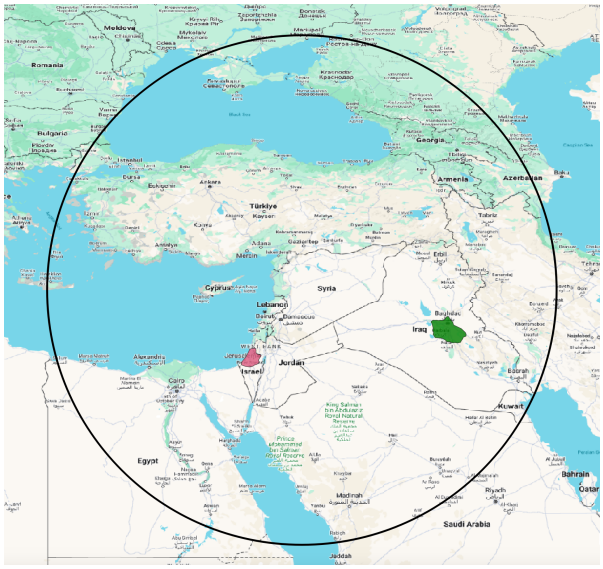
(d) Betweenness Centrality



(e) Degree Centrality

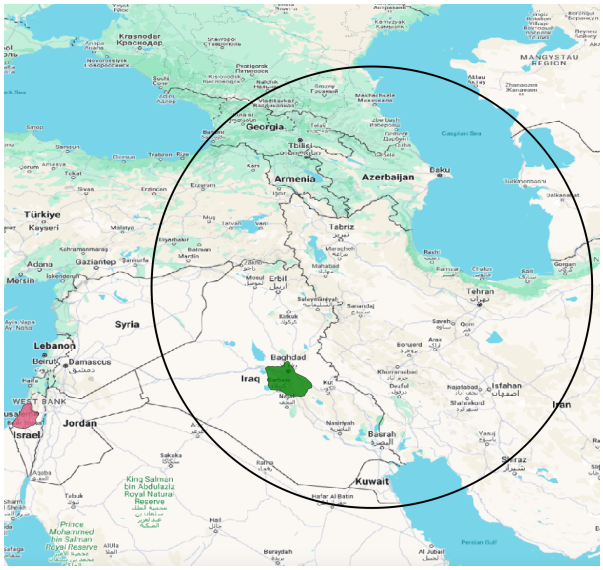
Figure C13: Judaism Estimates: Robustness to Alternate Calibration Specifications

Note: This figure shows 5 robustness exercises for the calibration of Judaism estimates. In each case, the estimates are similar to the main estimates.



(a) Weighted Euclidian clustering

(b) Ward clustering

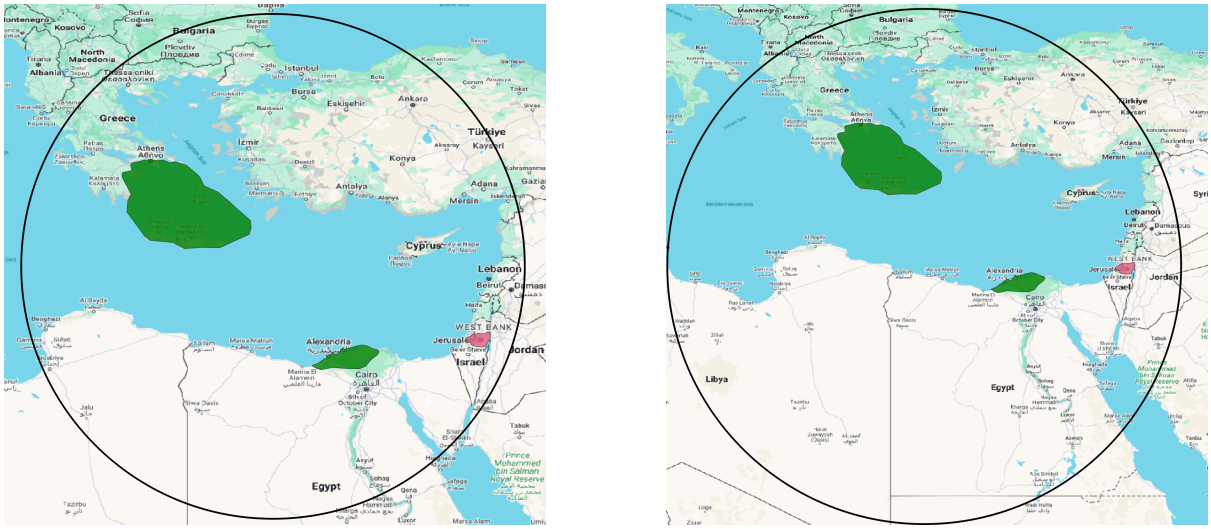


(c) Chebyshev clustering

(d) Median Euclidian Clustering

Figure C14: Judaism Estimates: Robustness to Alternate Clustering Routines

Note: This figure shows four robustness exercises for the clustering of Buddhism estimates. In each case, the estimates are similar to the main estimates.



(a) Christianity calibrated w. Islam

(b) Christianity calibrated w. Buddhism

Figure C15: Christianity estimates calibrated using Islam and Buddhism

Note: The map shows 75% confidence areas of estimates for Christianity, calibrated with each of Islam and Buddhism. In each subfigure, we present the historical account of the true origins of Buddhism and Islam in pink, with the origins of scripture in green. In subfigure (a) we show the Christianity estimate calibrated using the distances to Islam, estimated in table 4. In subfigure (b) we show the Christianity estimate calibrated using the distances to Buddhism, estimated in table 4.

(a) Linear specification

(b) Quadratic specification

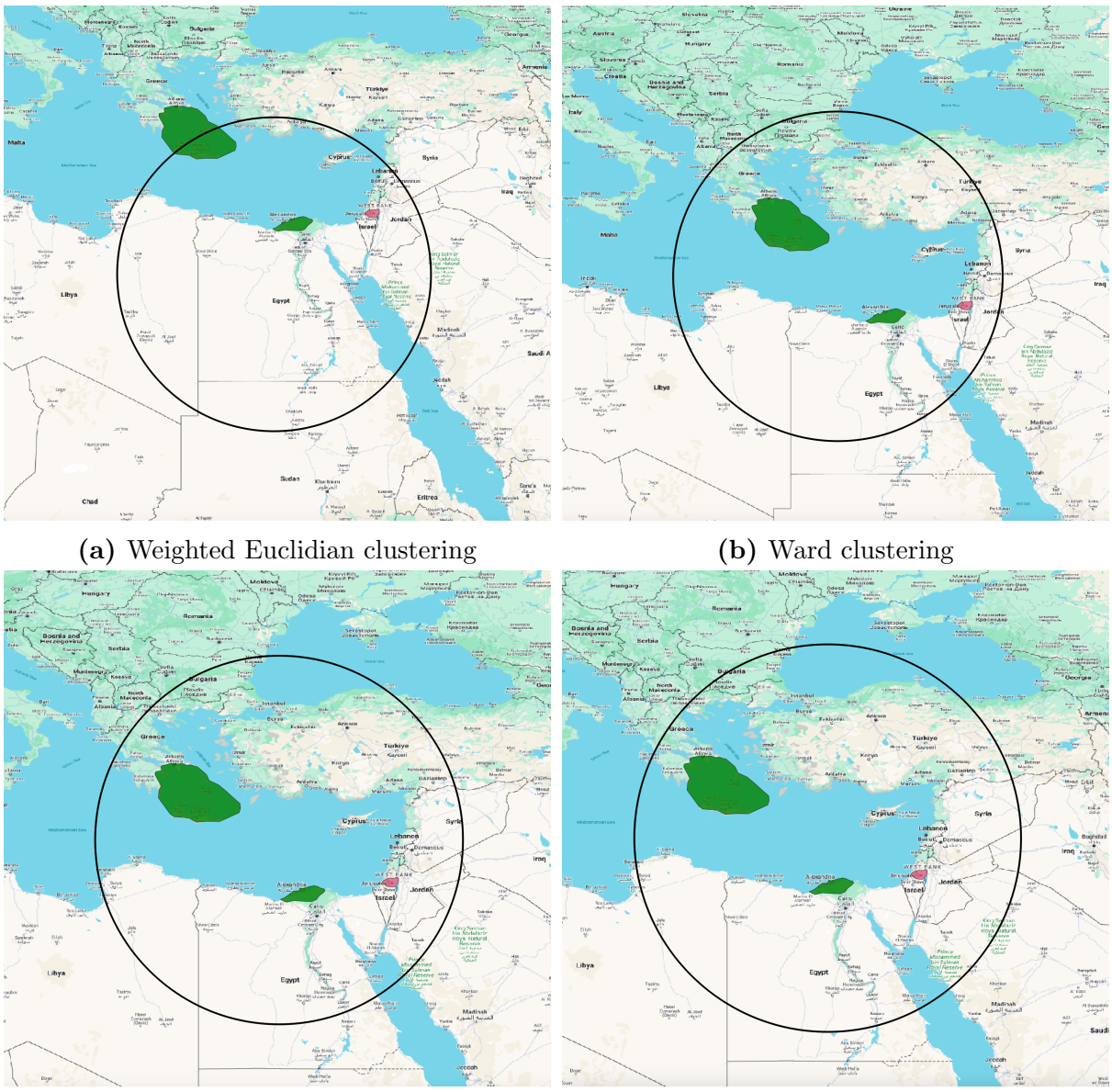
(c) linear dependent variable (instead of log)

(d) Betweenness Centrality

(e) Degree Centrality

Figure C16: Christianity Estimates: Robustness to Alternate Calibration Specifications

Note: This figure shows 5 robustness exercises for the calibration of Christianity estimates. In each case, the estimates are similar to the main estimates.



(a) Weighted Euclidian clustering

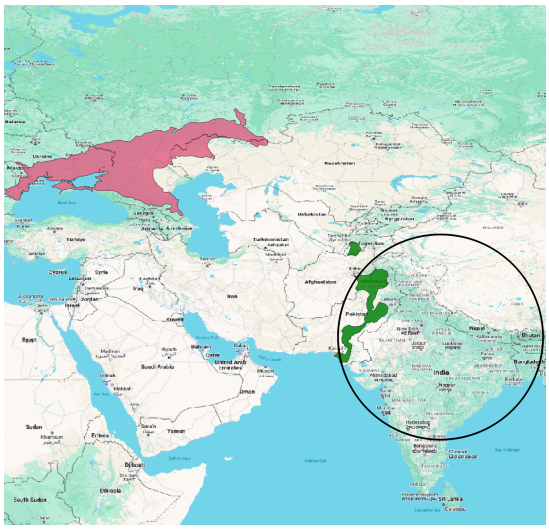
(b) Ward clustering

(c) Chebyshev clustering

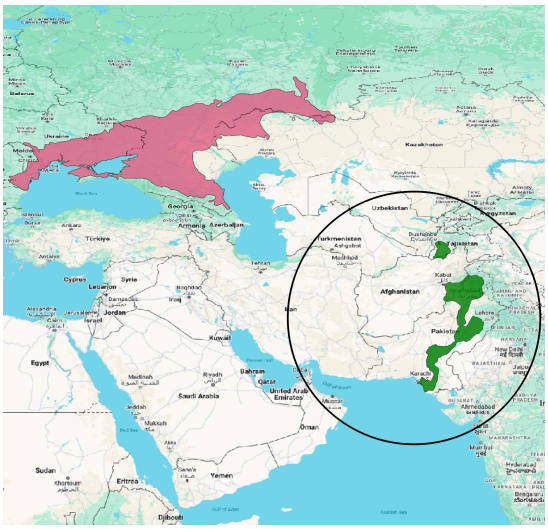
(d) Median Euclidian Clustering

Figure C17: Christianity Estimates: Robustness to Alternate Clustering Routines

Note: This figure shows four robustness exercises for the clustering of Buddhism estimates. In each case, the estimates are similar to the main estimates.



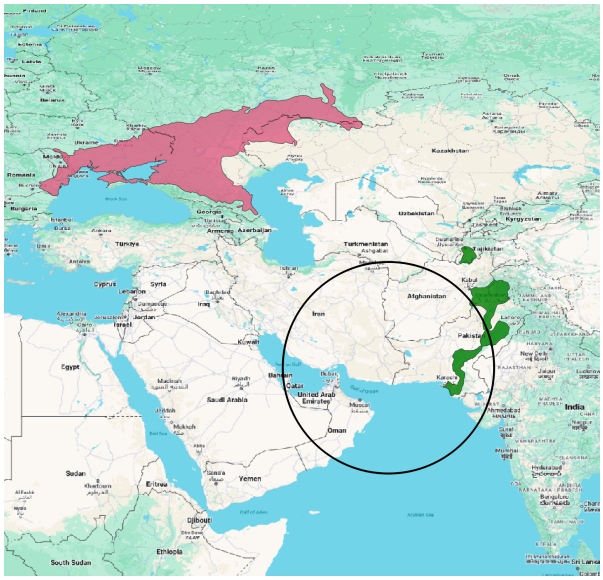
(a) Hinduism calibrated w. Islam



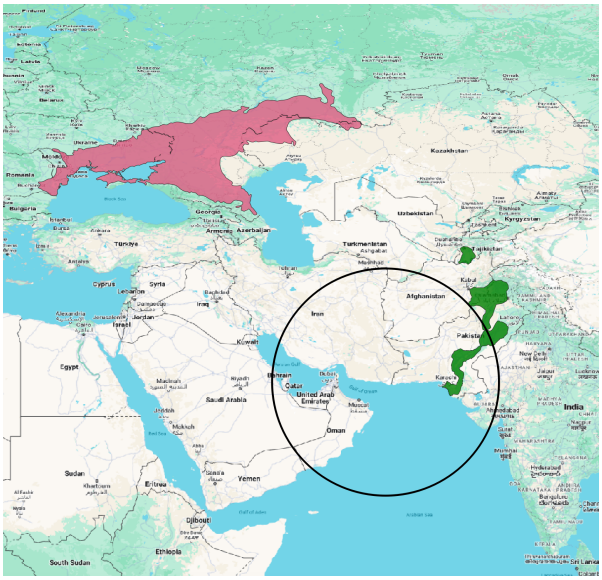
(b) Hinduism calibrated w. Buddhism

Figure C18: Hinduism estimates calibrated using Islam and Buddhism

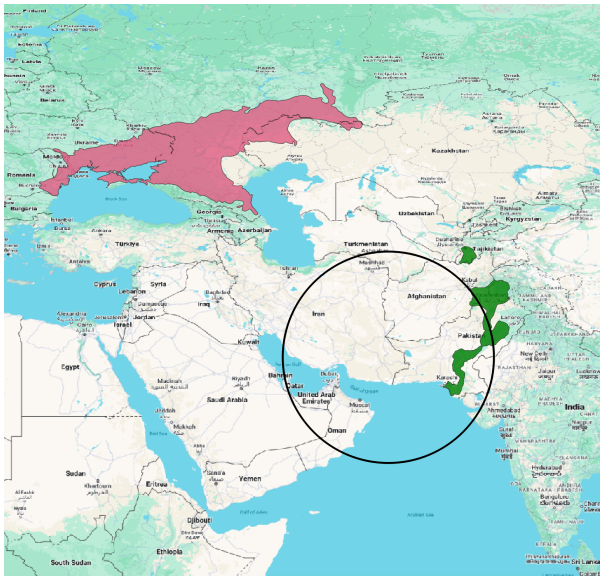
Note:
The map shows 75% confidence areas of estimates for Hinduism, calibrated with each of Islam and Buddhism. In each subfigure, we present the historical account of the true origins of Buddhism and Islam in pink, as before. In this case however, the the origins of scripture are also contested, so we show one hypothesis (Indus Valley) in green and the other (BMAC) in yellow. In subfigure (a) we show the Hinduism estimate calibrated using the distances to Islam, estimated in table 4. In subfigure (b) we show the Hinduism estimate calibrated using the distances to Buddhism, estimated in table 4.



(a) Linear specification



(b) Quadratic specification



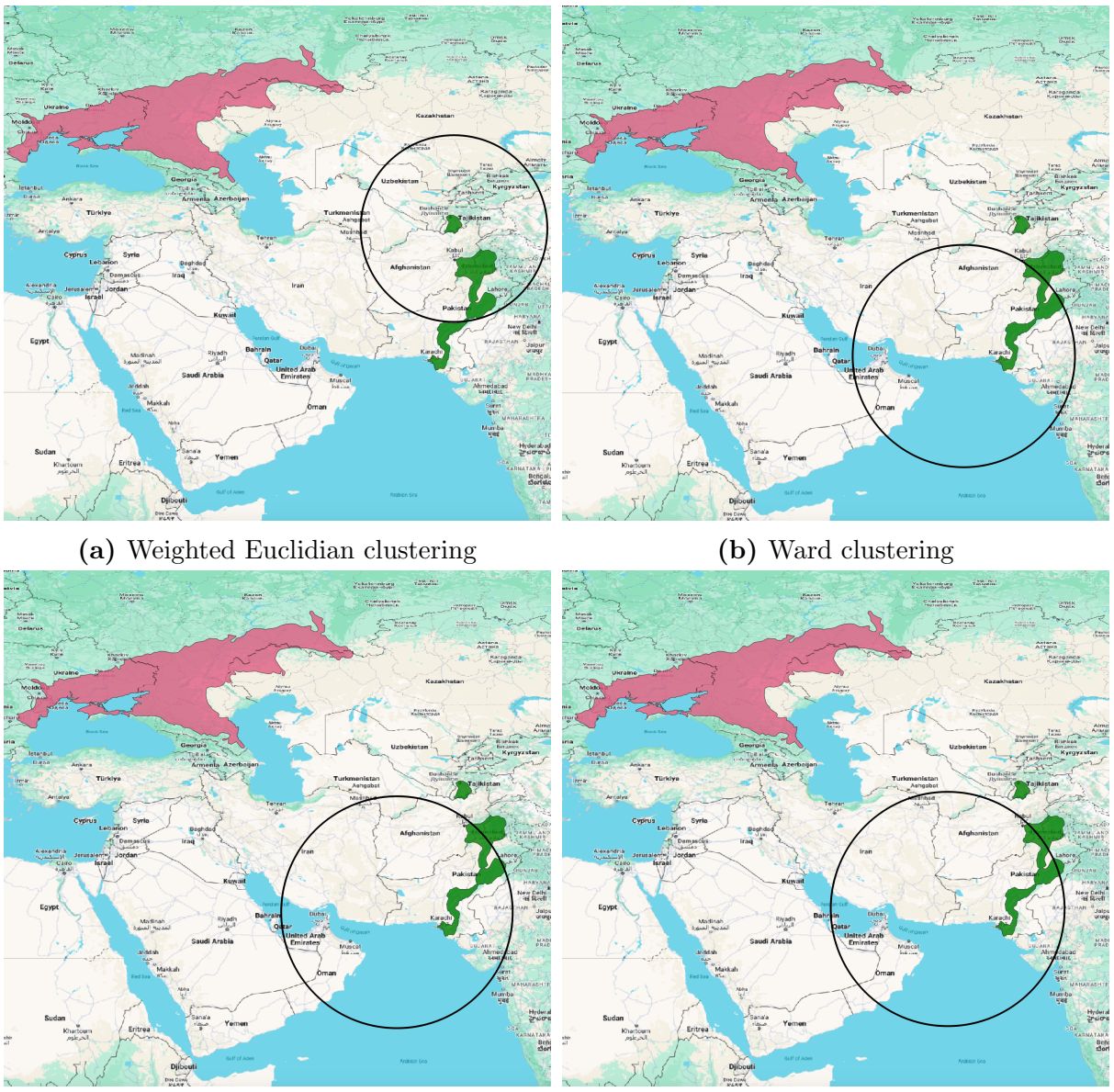
(c) linear dependent variable (instead of log)

(d) Betweenness Centrality

(e) Degree Centrality

Figure C19: Hinduism Estimates: Robustness to Alternate Calibration Specifications

Note: This figure shows 5 robustness exercises for the calibration of Hinduism estimates. In each case, the estimates are similar to the main estimates.



(a) Weighted Euclidian clustering

(b) Ward clustering

(c) Chebyshev clustering

(d) Median Euclidian Clustering

Figure C20: Hinduism Estimates: Robustness to Alternate Clustering Routines

Note: This figure shows four robustness exercises for the clustering of Buddhism estimates. In each case, the estimates are similar to the main estimates.

References

Alessio, M et al. (1969). "University of Rome carbon-14 dates VII". In: *Radiocarbon* 11.2, pp. 482–498.

Andree, Richard (1895). *India North.*

Armstrong, Karen. (2001). *Islam : a short history.* eng. Publication Title: Islam : a short history. London: Phoenix. ISBN: 1-84212-462-5.

Barros Damgaard, Peter de et al. (2018). "The first horse herders and the impact of early Bronze Age steppe expansions into Asia". In: *Science* 360.6396, eaar7711.

Basham, Arthur Llewellyn (1991). *The origins and development of classical Hinduism.* Oxford University Press, USA.

Beckwith, Christopher I (1987). "The Tibetans in the Ordos and North China: considerations on the role of the Tibetan Empire in world history". In:

Berkey, Jonathan (2004). "The formation of Islam". In: *Religion and society in the Near East, 600* 800.

Bhadra, Bidyut K, AK Gupta, and JR Sharma (2009). "Saraswati Nadi in Haryana and its linkage with the Vedic Saraswati River—integrated study based on satellite images and ground based information". In: *Journal of the Geological Society of India* 73.2, pp. 273–288.

Blouin, Arthur and Julian Dyer (2022). "How Cultures Converge: An Empirical Investigation of Linguistic Exchange, Trade, and Power". In: *Working paper.*

Blumenthal, James and James Apple (2008). "Śāntaraksita". In:

Bojanowski, Piotr et al. (2017). "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146.

Bonacich, Phillip (1972). "Factoring and weighting approaches to status scores and clique identification". In: *Journal of mathematical sociology* 2.1, pp. 113–120.

Boucher, Daniel J (1996). *Buddhist translation procedures in third-century China: a study of Dharmaraksa and his translation idiom.* University of Pennsylvania.

Bray, John (1991). "Language, tradition and the Tibetan Bible". In: *The Tibet Journal* 16.4, pp. 28–58.

Brown, Jonathan A.C. (Mar. 2011). *Muhammad: A Very Short Introduction.* Oxford University Press. ISBN: 978-0-19-955928-2. DOI: 10.1093/actrade/9780199559282.001.0001.

Buswell Jr, Robert Evans (2013). "Buddhism in Korea". In: *The Religious Traditions of Asia.* Routledge, pp. 355–362.

Ch'en, Kenneth Kuan Sheng and Kenneth Kuan Shêng Ch'en (1972). *Buddhism in China: A historical survey.* Vol. 1. Princeton University Press.

Chen, Jinhua (2004). "The Indian Buddhist Missionary Dharmaksema (385-433): A New Dating of His Arrival in Guzang and of His Translations". In: *T'oung Pao*, pp. 215–263.

Cobb, Paul M. (Nov. 2010). "The empire in Syria, 705–763". In: *The New Cambridge History of Islam*. Ed. by Chase F. Robinson. 1st ed. Cambridge University Press, pp. 226–268. ISBN: 978-1-139-05593-2. DOI: 10.1017/CHOL9780521838238.008. URL: https://www.cambridge.org/core/product/identifier/CBO9781139055932A012/type/book_part (visited on 06/09/2022).

Cohen, Rodrigo Laham (2018). *The Jews in Late Antiquity*. Arc Humanities Press. ISBN: 978-1-64189-909-3. DOI: 10.5040/9781641899093. URL: https://www.bloomsburymedievalstudies.com/encyclopedia?docid=b-9781641899093 (visited on 06/08/2022).

Coleman, James William (2002). *The new Buddhism: The western transformation of an ancient tradition*. Oxford University Press.

Curtis, John, ed. (Jan. 2020). *Studies in Ancient Persia and the Achaemenid Period*. The Lutterworth Press. ISBN: 978-0-227-90706-1 978-0-227-17705-1. DOI: 10.2307/j.ctv10vm0td. URL: http://www.jstor.org/stable/10.2307/j.ctv10vm0td (visited on 06/08/2022).

Daryaee, Touraj (2013). *Sasanian Persia: the rise and fall of an empire*. New paperback edition. London ; New York: I.B. Tauris & Co. Ltd in association with the Iran Heritage Foundation. ISBN: 978-1-78076-378-1.

DellaPergola, Sergio (1997). "Some fundamentals of Jewish demographic history". In: *Papers in Jewish demography* 13, p. 371.

Dèroche, François (Jan. 2022). *The One and the Many: The Early History of the Qur'an*. Trans. by Malcolm De Bevoise. Yale University Press. ISBN: 978-0-300-26283-4 978-0-300-25132-6. DOI: 10.2307/j.ctv270kvh1. URL: http://www.jstor.org/stable/10.2307/j.ctv270kvh1 (visited on 06/09/2022).

Dignas, Beate and Engelbert Winter (2007). *Rome and Persia in late antiquity: neighbours and rivals*. English. OCLC: 181341246. Cambridge: Cambridge University Press. ISBN: 9780511342486 9780511341427 9780511619182 9781281085061 9786611085063 9780511341953 9780511340840 9780511567988. URL: https://doi.org/10.1017/CBO9780511619182 (visited on 06/07/2022).

Donner, Fred McGraw (1981). *The early Islamic conquests*. Princeton, N.J: Princeton University Press. ISBN: 978-0-691-05327-1.

Drews, Robert (1998). "Canaanites and Philistines". In: *Journal for the Study of the Old Testament* 23.81, pp. 39–61.

Dubovsky, Peter (2006). "Tiglath-pileser III's campaigns in 734-732 BC: Historical background of isa 7; 2 Kgs 15-16 and 2 Chr 27-28". In: *Biblica*, pp. 153–170.

Ehrlich, Carl S, Marsha C White, and Marsha White (2006). *Saul in story and tradition*. Vol. 47. Mohr Siebeck.

Elverskog, Johan (2010). *Buddhism and Islam on the Silk Road*. English. OCLC: 1165551086. ISBN: 978-0-8122-0531-2. URL: https://doi.org/10.9783/9780812205312 (visited on 06/09/2022).

— (Feb. 2015). "Whatever Happened to Queen Jönggen?" In: *Buddhism in Mongolian History, Culture, and Society*. Ed. by Vesna A. Wallace. Oxford University Press, pp. 3–22.

Erdosy, George (1995). "Language, material culture and ethnicity: Theoretical perspectives". In: *The Indo-Aryans of Ancient South Asia: Language, Material Culture and Ethnicity, Berlin*, pp. 1–31.

Esposito, John L (2000). *The Oxford History of Islam*. English. OCLC: 1101047372. Oxford: Oxford University Press, Incorporated. ISBN: 978-0-19-977100-4. URL: https://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=5763596 (visited on 06/09/2022).

Flood, Gavin D and Gavin D Flood Flood (1996). *An introduction to Hinduism*. Cambridge University Press.

Freeman, Charles (2009). *A New History of Early Christianity*. Yale University Press. ISBN: 978-0-300-12581-8. URL: http://www.jstor.org/stable/j.ctt1nq44w (visited on 06/12/2022).

Harrill, J. Albert (2012). *Paul the Apostle: His Life and Legacy in Their Roman Context*. Cambridge: Cambridge University Press. ISBN: 978-1-139-04951-1. DOI: 10.1017/CBO9781139049511. URL: http://ebooks.cambridge.org/ref/id/CBO9781139049511 (visited on 06/10/2022).

Harvey, Peter (2012). *An introduction to Buddhism: Teachings, history and practices*. Cambridge University Press.

Henze, Matthias H (1999). *The madness of King Nebuchadnezzar: the ancient Near Eastern origins and early history of interpretation of Daniel 4*. Vol. 61. Brill.

Hess, Richard S (2007). *Israelite religions: an archaeological and biblical survey*. Baker Academic.

Hill, Jonathan (2020). *The History of Christianity The Early Church to the Reformation*. English. OCLC: 1226585177. Chicago: Lion Hudson LTD. ISBN: 978-1-912552-51-1. URL: http://public.eblib.com/choice/PublicFullRecord.aspx?p=6421316 (visited on 06/12/2022).

Hirshberg, Daniel (2016). *Remembering the Lotus-Born: Padmasambhava in the History of Tibet's Golden Age*. Vol. 19. Simon and Schuster.

Hitchcock, Susan Tyler and John L Esposito (2004). *Geography of religion: where God lives, where pilgrims walk*. National Geographic Society.

Hodgson, Marshall GS (2009). *The Venture of Islam, Volume 1: The Classical Age of Islam*. Vol. 1. University of Chicago press.

Humphreys, Christmas (2013). *The wisdom of Buddhism*. Routledge.

Huy, Nguyen Ngoc (1998). "The Confucian incursion into Vietnam". In: *Confucianism and the Family*, pp. 91–104.

Ilan, Tal (2009). "The Torah of the Jews of Ancient Rome". In: *Jewish Studies Quarterly* 16.4, pp. 363–395.

Jackson, Matthew O. (Nov. 2010). *Social and Economic Networks*. Princeton University Press. ISBN: 978-1-4008-3399-3 978-0-691-13440-6. DOI: 10.2307/j.ctvcm4gh1. URL: http://www.jstor.org/stable/10.2307/j.ctvcm4gh1 (visited on 05/31/2022).

Jaeger, Werner (1985). *Early christianity and greek paideia*. Harvard University Press.

Jain, Sharad K, Pushpendra K Agarwal, and Vijay P Singh (2007). *Hydrology and water resources of India*. Vol. 57. Springer Science & Business Media.

Jaro, Matthew A. (1989). "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida". In: *Journal of the American Statistical Association* 84.406, pp. 414–420.

Kalmin, Richard (2006). *Jewish Babylonia between Persia and Roman Palestine*. OUP USA.

Kapstein, Matthew T. (Oct. 2013). *Tibetan Buddhism: A Very Short Introduction*. Oxford University Press.

Kenoyer, Jonathan Mark (2006). "The origin, context and function of the Indus script: Recent insights from Harappa". In: *Proceedings of the Pre-symposium of RIHN and 7th ESCA Harvard-Kyoto Roundtable*, pp. 9–27.

Kleinberg, Jon et al. (2018). "Human decisions and machine predictions". In: *The quarterly journal of economics* 133.1, pp. 237–293.

Kochhar, Rajesh (2000). *The Vedic people: Their history and geography*. Orient Longman.

— (2012). "10 On the identity and chronology of the Rgvedic river Sarasvati". In: *Archaeology and language III: Artefacts, languages and texts*, p. 257.

Kuhrt, Amélie (2007). "Cyrus the Great of Persia: images and realities". In: *Representations of Political Power: Case Histories from Times of Change and Dissolving Order in the Ancient Near East*, pp. 169–91.

Lindberg, Carter (2009). *A brief history of Christianity*. John Wiley & Sons.

Lothian, John (1848). *Arabia*.

Ludlow, Morwenna (2009). *The early church*. English. OCLC: 893332629. London; New York: I.B. Tauris. ISBN: 978-0-85773-559-1. URL: http://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=1814133 (visited on 06/12/2022).

MacCulloch, Diarmaid (2010). *Christianity: The first three thousand years*. Penguin.

Mackintosh-Smith, T. (2019). *Arabs: A 3,000-Year History of Peoples, Tribes and Empires*. Yale University Press.

McFall, Leslie (2010). "The Chronology of Saul and David". In: *Journal of the Evangelical Theological Society* 53.3, p. 475.

Michalopoulos, Stelios, Alireza Naghavi, and Giovanni Prarolo (2018). "Trade and Geography in the Spread of Islam". In: *The Economic Journal* 128.616, pp. 3210–3241.

Mikhail, Maged S. A. (2014). *From Byzantine to Islamic Egypt: Religion, Identity and Politics after the Arab Conquest*. I.B.Tauris & Co Ltd. ISBN: 978-0-7556-9525-6 978-1-78453-481-3. DOI: 10.5040/9780755695256. URL: https://www.bloomsburymedievalstudies.com/encyclopedia?docid=b-9780755695256 (visited on 06/09/2022).

Mikolov, Tomas, Ilya Sutskever, et al. (2013). "Distributed Representations of Words and Phrases and their Compositionality". In: Publisher: arXiv Version Number: 1. DOI: 10.48550/ARXIV.1310.4546. URL: https://arxiv.org/abs/1310.4546 (visited on 06/13/2022).

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (June 2013). "Linguistic Regularities in Continuous Space Word Representations". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: Association for Computational Linguistics, pp. 746–751. URL: https://aclanthology.org/N13-1090.

Moorhead, John (2013). *Justinian*. English. OCLC: 864743310. Routledge. ISBN: 978-1-306-18350-5 978-1-317-89878-8 978-1-315-84574-6 978-1-317-89879-5 978-1-317-89877-1 978-1-138-83640-2 978-0-582-06303-7. URL: http://site.ebrary.com/id/10814081 (visited on 06/07/2022).

Morgan, David O. and Anthony Reid (Jan. 2000). "Introduction: Islam in a plural Asia". In: *The New Cambridge History of Islam*. Ed. by David O. Morgan and Anthony Reid. 1st ed. Cambridge University Press, pp. 1–18. ISBN: 978-1-139-05613-7. DOI: 10.1017/CHOL9780521850315.002. URL: https://www.cambridge.org/core/product/identifier/CBO9781139056137A005/type/book_part (visited on 06/09/2022).

Morgan, Kenneth W (1986). *The path of the Buddha: Buddhism interpreted by Buddhists*. Motilal Banarsidass Publ.

Mortensen, David R. et al. (2016). "PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors". In: *COLING*.

Narayanan, Vasudha (2009). *Hinduism*. The Rosen Publishing Group, Inc.

Neusner, Jacob (1970). *The Formation of the Babylonian Talmud*. Vol. 17. Brill Archive.

Newman, Judith H (2020). *The Invention of Judaism: Torah and Jewish Identity from Deuteronomy to Paul. By John J. Collins*.

Oppenheimer, A'haron and Nili Oppenheimer (2005). *Between Rome and Babylon: studies in Jewish leadership and society*. eng ger. Texts and studies in ancient Judaism = Texte und Studien zum antiken Judentum 108. Tübingen: Mohr Siebeck. ISBN: 978-3-16-148514-5.

Parpola, Asko (2015). *The roots of Hinduism: the early Aryans and the Indus civilization*. Oxford University Press, USA.

"Zoroastrian Fire Temples and the Islamisation of Sacred Space in Early Islamic Iran" (Apr. 2017). en. In: *Islamisation*. Ed. by A. C. S. Peacock. Edinburgh University Press, pp. 102–117. ISBN: 978-1-4744-1712-9 978-1-4744-3498-0. DOI: 10.3366/edinburgh/9781474417129.003.0006. URL: https://edinburgh.universitypressscholarship.com/view/10.3366/edinburgh/9781474417129.001.0001/upso-9781474417129-chapter-006 (visited on 06/09/2022).

Pearce, Laurie E and Cornelia Wunsch (2014). *Documents of Judean exiles and West Semites in Babylonia in the collection of David Sofer*. cdl Press.

Pienaar, Daniel N (1994). "Aram and Israel during the reigns of Omri and Ahab reconsidered". In: *Journal for Semitics* 6.1, pp. 34–45.

Prakash, Gyan (1994). "Subaltern studies as postcolonial criticism". In: *The American historical review* 99.5, pp. 1475–1490.

Prange, Sebastian R. (May 2018). *Monsoon Islam: Trade and Faith on the Medieval Malabar Coast*. 1st ed. Cambridge University Press. ISBN: 978-1-108-33486-0 978-1-108-42438-7 978-1-108-43814-8. DOI: 10.1017/9781108334860. URL: https://www.cambridge.org/core/product/identifier/9781108334860/type/book (visited on 06/09/2022).

Prebish, Charles S (Charles Stuart) (2008). "Cooking the Buddhist books: the implications of the new dating of the Buddha for the history of early Indian Buddhism". In: *Journal of Buddhist Ethics* 15. ISSN: 1076-9005.

Premasiri, PD (2022). "Implications of Buddhist Political Ethics for the Minimisation of Suffering in Situations of Armed Conflict". In: *Contemporary Buddhism*, pp. 1–15.

Przyluski, Jean (1934). "Origin and Developement of Buddhism". In: *The Journal of Theological Studies* 35.140, pp. 337–351.

Pyysiäinen, Ilkka (2003). "Buddhism, Religion, and the Concept of" God"". In: *Numen* 50.2, pp. 147–171.

Raikes, Robert L (1964). "The end of the ancient cities of the Indus". In: *American Anthropologist* 66.2, pp. 284–299.

Rawski, Evelyn S (2005). "Qing Publishing in Non-Han Languages". In: *Printing and Book Culture in Late Imperial China* 27, p. 304.

Redmount, Carol A and Michael David Coogan (1998). "Bitter Lives. Israel in and out of Egypt". In: *The Oxford history of the Biblical world*, pp. 79–121.

Renfrew, Colin (1990). *Archaeology and language: the puzzle of Indo-European origins*. CUP Archive.

Robinson, Chase F (2011). *The new Cambridge history of Islam. sixth to eleventh centuries Volume I, Volume I,* English. ISBN: 9781139055932 OCLC: 758901785.

Rousseeuw, Peter J (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of computational and applied mathematics* 20, pp. 53–65.

Rowton, Michael B (1951). "Jeremiah and the Death of Josiah". In: *Journal of Near Eastern Studies* 10.2, pp. 128–130.

Saeed, M Yousuf (2019). "South Asia in the Postcolonial Era". In: *The Changing World of Contemporary South Asian Poetry in English: A Collection of Critical Essays*, p. 1.

Sarris, Peter (2015). *Byzantium: a very short introduction*. First edition published in 2015. A very short introduction. OCLC: ocn907657616. Oxford, U.K. ; New York: Oxford University Press. ISBN: 978-0-19-923611-4.

Schadeberg, Thilo C. (2009). "Loanwords in Swahili". In: *Loanwords in the World's Languages: A Comparative Handbook*. Ed. by M. Haspelmath and U. Tadmor. De Gruyter Mouton, pp. 77–102.

Scheindlin, Raymond P. (1998). *A Short History of the Jewish People: From Legendary Times to Modern Statehood*. MacMillan.

Schrader and L. Vivien St Martin (1936). *Inde N-O. et Afganistan*.

Senior, Donald (Jan. 2022). *The New Testament: A Guide*. en. 1st ed. Oxford University Press. ISBN: 978-0-19-753083-2 978-0-19-753087-0. DOI: 10.1093/oso/9780197530832.001.0001. URL: https://oxford.universitypressscholarship.com/view/10.1093/oso/9780197530832.001.0001/oso-9780197530832 (visited on 06/12/2022).

Shaw, Ian (2000). "Egypt and the outside world". In: *The Oxford history of ancient Egypt*, pp. 314–29.

Skilton, Andrew (1997). *A concise history of Buddhism*. Windhorse Publications.

Smith, William and Charles Muller (1874). *A map of part of Asia to illustrate the Old Testament and classical authors*.

Sparks, Kenton L (2007). "Religion, identity and the origins of Ancient Israel". In: *Religion Compass* 1.6, pp. 587–614.

Srinivasan, Doris M (1983). "Vedic Rudra-Śiva". In: *Journal of the American Oriental Society*, pp. 543–556.

Srivastava, GS (2018). "Saraswati River: It's Past and Present". In: *The Indian Rivers*. Springer, pp. 503–521.

Srivastava, VC and VC Shrivastva (1981). "Recent Archaeological Researches in Afghanistan and their bearing on Indian History". In: *Proceedings of the Indian History Congress*. Vol. 42, pp. 631–640.

Strauch, Ingo (2019). "Buddhism in the West? Buddhist Indian Sailors on Socotra (Yemen) and the Role of Trade Contacts in the Spread of Buddhism". In: *Buddhism and the Dynamics of Transculturality: New Approaches* 64, p. 15.

Strickland, Keir and Robin Coningham (Sept. 2020). "Archaeologists Uncover the Roots of Buddhism." In: *Agora* 55.3, pp. 13–18.

Tanner, Henry S. (1826). *The Countries Traveled by the Apostles*.

Taupier, Richard (Feb. 2015). "The Western Mongolian Clear Script and the Making of a Buddhist State". In: *Buddhism in Mongolian History, Culture, and Society*. Ed. by Vesna A. Wallace. Oxford University Press, pp. 23–36.

Thiele, Edwin R (1983). *The mysterious numbers of the Hebrew kings*. Kregel Academic.

Tien, Nguyen Tran and Chih-yu Shih (2016). "Buddhist Influence in Vietnamese Diplomacy Toward China Lessons from the History of Religion". In: *Understanding 21st Century China in Buddhist Asia History, Modernity, and International Relations*, p. 45.

Weise, Kai (2013). *The Sacred Garden of Lumbini: Perceptions of Buddha's Birthplace*. UNESCO.

Winkler, William (Jan. 1, 1990). *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage*.

Zürcher, Erik (2007). *The Buddhist Conquest of China : The Spread and Adaptation of Buddhism in Early Medieval China*. English. Vol. 3rd ed. Sinica Leidensia. Leiden: Brill.