Extracting hierarchical features of cultural variation using network-based clustering

Xiran Liu,[1] Noah A. Rosenberg,[2,]*† Gili Greenbaum[3,]*†

1. Institute for Computational and Mathematical Engineering, Stanford University, Stanford, California, USA
2. Department of Biology, Stanford University, Stanford, California, USA
3. Department of Ecology, Evolution and Behavior, Hebrew University of Jerusalem, Jerusalem, Israel

† Joint senior authors
* Corresponding author. Email: noahr@stanford.edu; gil.g@mail.huji.ac.il

Key words: cultural evolution, hierarchical clustering, network

Abstract

High-dimensional data sets on cultural characters contribute to uncovering insights about factors that influence cultural evolution. Because cultural variation in part reflects descent processes with a hierarchical structure—including descent of populations and vertical transmission of cultural traits—methods designed for hierarchically structured data have potential to find applications in the analysis of cultural variation. We adapt a network-based hierarchical clustering method for use in analyzing cultural variation. Given a set of entities, the method constructs a similarity network, hierarchically depicting community structure among them. We illustrate the approach using four data sets: pronunciation variation in the United States mid-Atlantic region, folklore variation in worldwide cultures, phonemic variation across worldwide languages, and temporal variation in first names in the United States. In these examples, the method provides insights into processes that affect cultural variation, uncovering geographic and other influences on observed patterns and cultural characters important in contributing to them.

Social media summary: Network-based clustering reveals structure in cultural variation in pronunciation, folklore, phonemes, and first names

Introduction

In recent years, increasingly available large-scale data sets on aspects of variation across human cultures and within cultures over time have provided rich information about fine-scale details of human cultural variation and the factors that influence its dynamics (Mesoudi, 2016; Kolodny *et al.*, 2018). For example, investigations of variation in folktales among cultures have identified interactions of cultural diffusion and demic diffusion in the spread of folklore and mythology (Bortolini *et al.*, 2017; Thuillard *et al.*, 2018). A study of design features of traditional canoes across Polynesian societies has suggested a faster rate of cultural change in canoe traits that were less significant to functional performance of the watercraft, in line with the faster evolution that occurs for non-functional rather than functional genetic variants (Rogers & Ehrlich, 2008). Studies of variation in the presence and absence of linguistic characters across languages have uncovered influences of ancient migrations on patterns of language variation observed today (Atkinson, 2011; Creanza *et al.*, 2015).

The analysis of complex data to reveal features of cultural variation makes use of a variety of statistical methods designed for high-dimensional data analysis more generally. Such methods include analyses of distance matrices based on cultural traits of interest (Rogers & Ehrlich, 2008; Creanza *et al.*, 2015; Bortolini *et al.*, 2017; Thuillard *et al.*, 2018), multivariate analysis techniques such as principal components analysis (Creanza *et al.*, 2015), correlations involving spatial statistics and geographic maps (Atkinson, 2011; Creanza *et al.*, 2015; Bortolini *et al.*, 2017), and hierarchical tree-based clustering (Creanza *et al.*, 2015; Thuillard *et al.*, 2018).

Viewed in relation to their underlying generative processes, different forms of cultural variation often possess shared features (Cavalli-Sforza & Feldman, 1981; Boyd & Richerson, 1985). Different cultural entities might possess a shared variant, as a result of processes such as the independent origin of functionally significant variants, random recurrence of nonfunctional variants, or cultural exchange. Salient among the forces contributing to patterns of cultural variation is shared descent, so that even if independent origins and cultural exchange are important in specific settings, hierarchical or geographic structure can often contribute to features of cultural variation.

Owing to the importance of shared descent in influencing cultural variation, tools for analyzing cultural variation data can employ methods suited to analysis of genetic data, which also possess signatures of shared descent; thus, many statistical methods used in cultural data analysis are similar to those used for genetic data (Bromham, 2017; Gray *et al.*, 2010; Pagel, 2009). Recently, we have introduced a method, *NetStruct*, for use in understanding genetic variation data that result from hierarchical genetic structure (Greenbaum *et al.*, 2016, 2019). The method, employing ideas from network analysis, produces a distinctive form of visualization of hierarchical population relationships. It has been seen to detect subtle patterns that have been overlooked using earlier forms of data analysis.

The *NetStruct* method consists of three main steps: construction of similarity matrices between entities, community detection in similarity matrices, and hierarchical visualization of communities. The method is general beyond genetic data, as the form of the data contributes only to the choice of similarity function. It can thus be modified for use with other types of data that result from distinct but related generative processes, including data on cultural variation.

Here, we adapt the *NetStruct* method for use in the study of cultural evolution. We examine a variety of data sets on different forms of cultural variation, considering geographic variation in English pronunciation, variation across cultures in folklore, phonemic variation across languages, and temporal variation in frequencies of first names. Using each of the four forms of cultural data, we illustrate the potential of the method as an exploratory tool to reveal features of geographic and temporal structure in cultural phenomena and to extract patterns that can inspire hypotheses about underlying mechanisms. Each example additionally highlights a different aspect of the hierarchical analysis: analyses at different levels of detail in the hierarchy, identification of characters that are important in driving the partitioning, analysis of outliers, and the relationship of the hierarchy to features of entities beyond those used in its construction.

Results

Generalizing the *NetStruct* pipeline

In the first step of the *NetStruct* method, for a set of entities, each having a value for each of a series of characters, we construct an $n \times m$ data matrix A with n rows corresponding to entities and m columns corresponding to characters. Entry A_{ij} gives the value of character j for entity i; this value can be either categorical or quantitative, depending on the type of character.

The similarity between two entities i_1 and i_2, denoted s_{i_1,i_2}, is computed by a function applied to rows i_1 and i_2. We normalize pairwise similarities so that they take on values in [0,1]. The resulting $n \times n$ similarity matrix S then becomes the adjacency matrix of a similarity network. The similarity function is chosen based on a particular application of interest.

In a network, community structure exists when high concentrations of edges occur within certain groups of nodes in the network and low concentrations occur between these groups (Girvan & Newman, 2002). In the second step of *NetStruct*, we iteratively remove edges with lower weights from the network to reveal the finer-scale structure within coarser communities. *NetStruct* uses a community-detection Louvain algorithm (Blondel *et al.*, 2008) together with an iterative edge-pruning method (Greenbaum *et al.*, 2019). The Louvain algorithm maximizes a "modularity score" for each community, quantifying the difference between the actual density of edges within the community and the expected density if all edges in the network were distributed at random while preserving the degree distribution of the network. The Louvain algorithm starts by assigning each node to its own community, sequentially merging nodes into communities in a manner that produces the greatest modularity increase—until no further increase occurs. *NetStruct* iteratively removes edges below a weight threshold of increasing value and applies community-detection in each subdivided community at each iteration, generating hierarchical structure at multiple levels.

Finally, in the last step, the communities detected at each iteration are assembled to form the output hierarchy, which can be visualized as a hierarchical tree coded by a coloring scheme. Because clustering is hierarchical, each entity can belong to multiple communities, or clusters, at different hierarchical levels; that is, each cluster can have finer-scale "child" clusters.

NetStruct visualizes community structure using a diagram that depicts hierarchical relationships among clusters. Each cluster is assigned an interval representation of a color gradient; the root node is assigned the unit interval. Child clusters are assigned equal portions of the interval associated with their parental node. Clusters are colored by the midpoint of the associated interval, such that at each hierarchical level, child clusters of the same parent have colors that are more similar than are those of different parents. The color scheme facilitates interpretation, as the original entities can be labeled by the finest-scale cluster to which they are assigned in the diagram.

To generalize use of *NetStruct* beyond genetic data, we require a function that describes similarity between pairs of entities of interest. Many similarity measures are possible, and *NetStruct* is applied to the similarity matrix after it has been constructed. For a given data set of interest, the similarity function is chosen in a manner suited to the application. We follow Greenbaum *et al.* (2016, 2019) in choosing frequency-weighted similarity measures that emphasize shared rare values of a character.

Variation in pronunciation across locations

For our first example, we examined data on individual variation in pronunciation. Local variation in communication variants has potential to provide insight into cultural transmission and spatial patterns of distinctiveness and interaction in a population (Nerbonne & Kleiweg, 2003; Rendell & Whitehead, 2005; Aplin, 2019). To understand the relationship between geography and individual-level pronunciation of a shared human language, we applied *NetStruct* to data on English pronunciation variation in the middle and south Atlantic region of the United States.

LAMSAS pronunciation data

We obtained pronunciation variation data from the Linguistic Atlas of the Middle and South Atlantic States (LAMSAS) (Kretzschmar Jr. *et al.*, 1993). These data consist of dialect records on pronunciations of everyday words collected in 1933-1942 from eleven states: Delaware, Maryland, New Jersey, New York, North Carolina, Pennsylvania, South Carolina, Virginia, and West Virginia, with some records from eastern Georgia and northeastern Florida included as well.

We restricted our analysis to n=839 informants interviewed by the major field worker (Nerbonne & Kretzschmar, 2003) and m=69 words recorded for most informants. We constructed the $n \times n$ similarity matrix based on phonetic transcriptions of pronunciations of the m words. The similarity is greater when informants share many pronunciations, and when they share rare pronunciations (*Methods*). We then applied *NetStruct* to infer hierarchical structure.

Hierarchical structure of pronunciation variation: levels of detail

Fig. 1 presents the hierarchical structure of pronunciation variation in the LAMSAS data. In Fig. 1A, we color informants on the map by their finest-scale clusters in the tree diagram. In the *NetStruct* color scheme, informants with more similar colors appear closer in the tree diagram, and those with distinct colors are placed in different branches at relatively high levels in the hierarchy. For example, in Fig. 1A, distant colors purple and yellow belong to different major branches of the hierarchy; informants colored purple are mostly in the northern part of the Atlantic region, and those colored yellow are mostly in the southern part.

To examine the two major clusters at a finer level of detail, we reapply the coloring, for each cluster assigning the root node the color corresponding to the midpoint in the unit interval (Fig. 1, B and C). Within each of the two clusters, finer levels of the hierarchy group together informants who are geographically closer. In the cluster that contains most of the individuals from the more northerly regions (Fig. 1B), pronunciation distinctions can be observed in groups corresponding largely to New York and to West Virginia. In the cluster that contains most of the more southerly individuals (Fig. 1B), a distinction is noticeable between finer clusters corresponding to North Carolina and to Virginia, with some individuals in both states placed in small clusters.

We repeat the process to examine Fig. 1C in even finer detail. This analysis, in Fig. 1D and Fig. 1E, illustrates that at lower levels of the hierarchy, clusters are not always associated with geographical features. However, we observe that year of birth is strongly associated with cluster assignment at this local geographic scale (Fig. 1, D and E). In other words, in some tree branches, clusters within a branch correspond to age structure, rather than to geography.

This analysis highlights that our clustering extracts one set of features from pronunciation variation at high hierarchical levels—geographical variation in informants—and at lower hierarchical levels, it captures other features, such as age structure. The analysis of multiple hierarchical levels assists in the interpretation of the patterns both at the broadest scale as well as at fine-scale levels.

Variation in folklore motifs across cultures

In the study of folklore and mythology, recurring plot patterns, or "motifs," occur across cultures. Motif variation can provide insight into cross-cultural patterns, including migrations and cultural transmission in relation to ethnolinguistic barriers (Berezkin, 2010; Bortolini *et al.*, 2017; Korotayev *et al.*, 2017; Thuillard *et al.*, 2018). Here we used folklore motifs to analyze cultural variation, identifying motifs important in constructing the proposed hierarchical relationships.

Database of folklore

We examined data on presence and absence of folklore motifs in individual cultures. Using folklore data from around the world, Berezkin *et al.* (Berezkin *et al.*, 2009; Korotayev *et al.*, 2017) tabulated recurring motifs prominent in links between folklore traditions, defining a motif to be "any image, compositional structure, episode or chain of episodes found in more than one text." Berezkin *et al.* reported a list of cultures for each motif.

We focused our analysis on the n=65 regions in the Berezkin *et al.* database and the m=2459 motifs appearing in at least two of these regions. We computed similarities between pairs of regions based on numbers of shared motifs, negatively weighted by motif frequency (*Methods*).

Hierarchical structure of folklore variation: important characters

Fig. 2A presents the hierarchical structure of motif variation extracted using the pairwise similarities calculated based on all motifs. The geographic regions are mostly clustered into three large areas: Eurasia and Africa (purple), North America (blue), and South America (orange), with varied placement of populations from Australia and Oceania.

To identify which motifs are most important for extracting the hierarchical features, we adopt the normalized mutual information (NMI) approach to compare hierarchies constructed using different sets of motifs (Greenbaum *et al.*, 2019). For a pair of hierarchical clusterings, the NMI measure ranges from 0 to 1, quantifying the information obtained about one clustering by observing another. The NMI measure is high when two clustering hierarchies describe similar clustering structures (*Methods*). The NMI approach is flexible in that it enables comparisons between subsets of the hierarchical structure, for example by comparing only the leaves of the hierarchy.

We sampled 100 random subsets of 20, 50, 100, and 500 motifs, for each subset applying *NetStruct* to extract a hierarchy from the similarity network based on the sampled motifs. We then computed the NMI between the hierarchy of the sampled motifs and the hierarchy for all motifs, both for the full tree and for only the leaf clusters. In both NMI analyses, as the number of motifs in the subset increases, the mean of the NMI distribution increases (Fig. 2, B and D). The hierarchy produced by a larger subset of motifs is more informative than those generated with fewer motifs.

Different motifs can be more informative or less informative regarding the hierarchical structure of the data. For example, a motif found in all regions, or one not correlated with the main cultural patterns, will not be informative about the clustering. To identify the most informative motifs, we sampled 5,000 subsets of 20 motifs with replacement, counting occurrences of motifs in the 200 subsets possessing the highest NMI with the full tree and those possessing the highest NMI for leaf clusters. With random sampling, the expected number of occurrences of each motif in the top 200 subsets is $20/m\times200\approx1.6$.

The five most informative motifs for the full hierarchical structure appear in Fig. 2C. The motif most frequently found in high-NMI subsets is 'trickster is a feline,' appearing in 16 of 200 subsets ($P=1.5\times10^{-11}$, binomial test). This motif is common in Central and South America. 'To sort grain' has 11 occurrences ($P=9.8\times10^{-7}$), and the next three most informative motifs have eight occurrences each ($P=2.7\times10^{-4}$) and are also associated with large geographic regions (Fig. 2C). Some informative motifs correspond to natural or cultural phenomena restricted by geography, such as the practice of agriculture and the habitat ranges of animals.

The three most informative motifs for the fine-scale cultural structure represented by the leaves of the hierarchy are shown in Fig. 2E. Each appears eight times, above the number expected from random sampling ($P=2.7\times10^{-4}$). Two of these, 'a drop of blood' and 'the packed kingdom' have restricted geographic ranges. This result suggests that motifs of local folklore contribute to fine-scale features of the hierarchy.

In addition to visualizing hierarchical patterns of variation in folklore in relation to geography, this analysis demonstrates the use of *NetStruct* to identify characters—folklore motifs in this case—that play an important role in driving the hierarchical structure. The analysis of many subsets of characters, and the identification of those that appear in subsets that give rise to high NMI with the full-data analysis, uncovers those that contribute most to hierarchical clustering patterns.

Variation in phonemes across languages

A salient feature of linguistic variation is phonemic variation: variation in the sounds present within languages. Phonemic variation can be used to study inter-language relationships and population migrations (Atkinson, 2011; Creanza *et al.*, 2015; Fort & Pérez-Losada, 2016; Pérez-Losada & Fort, 2018), and for our next example, we analyzed hierarchical structure in worldwide phonemic variation.

Ruhlen phoneme database

Creanza *et al.* (2015*)* analyzed two databases that have been assembled on phonemes across large numbers of languages. We applied *NetStruct* on one of these, the Ruhlen database, as studied by Creanza *et al.* (2015) to explore phoneme-based hierarchical structure across languages. This database contains presence/absence information for 728 phonemes, organized by language classification and geography.

In our analysis, we included all n=2082 languages and m=454 phonemes that exist in more than one language. We then constructed the hierarchy based on the pairwise frequency-weighted phoneme-sharing similarities calculated from the $n \times m$ data matrix (*Methods*).

Hierarchical structure of phonemic variation: outlier entities

The hierarchy extracted from phonemic variation clusters languages in accord with geography on a broad scale (Fig. 3A). In Fig. 3B, major clusters tend to be localized within continents, in many places co-occurring with other such clusters.

Fig. 3, panels C-E, highlights patterns in local regions. In northeast Siberia (Fig. 3D), four of five languages of the Chukotko-Kamchatkan language family—Alyutor, Chukchi, Kerek, and Koryak—cluster in one branch (purple to yellow colors), whereas the Kamchadal language is alone in another (green). Indeed, the first four languages and Kamchadal are assigned to different branches in the family, Chukotian and Itelmen, and the unity of the family has been uncertain (Fortescue, 2005).

In East Africa (Fig. 3E), three languages shown in a distinct color from their surrounding languages—Dahalo, Hadza, and Sandawe—are the only three languages in the region that are click languages, a phonemic group of languages for which clicks function as normal consonants (Westphal, 1971). Similarly, in Northeast Asia (Fig. 3C), Korean, a language isolate, is clustered into a branch distinct from other neighboring languages.

This analysis, like the analyses of pronunciation and folklore motifs, illustrates the use of the *NetStruct* framework to identify geographic effects on entities of interest (assemblages of pronunciation variants, folklore motif repertoires, and phoneme inventories). The local patterns additionally illustrate the potential of the method for understanding effects on entities—in this case, languages—whose placements in the hierarchy differ from those of their geographic neighbors.

Variation in first names over time

Frequencies of first names among births in a population represent a rich source of cultural data, enabling tests about mechanisms of cultural change (Hahn & Bentley, 2003; Gureckis & Goldstone, 2009; Berger *et al.*, 2012; Kessler *et al.*, 2012; Acerbi & Bentley, 2014; O'Dwyer & Kandler, 2017). Our final example used *NetStruct* to analyze relationships among names in their patterns of temporal variation.

Social Security data on first names

Data on frequent first names from Social Security card applications for births starting in 1880 are provided publicly by the U.S. Social Security Administration. Separately for male and female names, for each year of birth, frequency data are provided. We analyzed female and male names separately, restricting attention to 1397 female and 1074 male names of total frequency greater than or equal to 10,000 through the end of 2019.

Considering each year 1880-2019 separately, the data set gives two $n \times m$ matrices with m=140 years, and n=1397 for female and n=1074 for male names. The similarity score between two names is computed based on the Pearson correlation between their frequency vectors over the m years of available data (*Methods*). We generated the *NetStruct* hierarchy from these similarities.

To interpret the *NetStruct* hierarchy, we made use of state-specific data, which are available alongside the national data starting from 1910. In the state-level data, each of n names has 53 vectors of counts of length 110, for 53 locations (50 states plus District of Columbia, Puerto Rico, and other territories) and 110 years (1910-2019). After normalizing counts from each year by the total number of individuals for that year, we identified for each name the state with the greatest mean normalized frequency over 110 years. In other words, we labeled each name by the state in which it was most frequent.

Hierarchical structure of variation in temporal patterns among names: features of entities

We present the hierarchical structure extracted from time series data on the frequencies of female names, as well as the temporal trends of the corresponding names, in Fig. 4A, with seven major branches of the hierarchy colored differently. The same visualization for male names appears in Fig. 4E, with five major branches assigned different colors. Both for female and for male names, names in branches of different color have distinct frequency trends over time, with those on the left indicating names that had greatest frequency at the beginning of the time series.

Fig. 4B and Fig. 4F relabel the hierarchies in Fig. 4A and Fig. 4E by the state in which a name has occurred most frequently over the full data set. The calculation of the state with highest frequency for a name is described in *Methods*. Initially, the most populous states were New York and Pennsylvania; California and Texas have been most populous more recently. Thus, the leftmost names, frequent early in the period, tend to be associated with New York, and to some extent, Pennsylvania and Texas. Names in the center are more closely associated with California. Names in the rightmost clusters are associated with California or Texas, whose recent population growth has reduced the difference from California in the number of annual births.

Because the patterns in Fig. 4B and 4F are driven in large part by population sizes of states, we next relabel the hierarchies using a frequency that is normalized by population size. In particular, we group states into four regions—Midwest, South, Northeast, and West—normalizing the region-wise count of each name by the total number of individuals in the region. The calculation of the region with highest normalized frequency for a name is described in *Methods*. Fig. 4C and Fig. 4G relabel the hierarchies in Fig. 4A and Fig. 4E by the region in which the normalized frequency is greatest. In this relabeling, the South is the region that has the largest number of names associated with it, both for females and for males. This pattern is particularly pronounced at the beginning of the time series, during which the South was the region of greatest frequency for large numbers of names.

Fig. 4D and Fig. 4H examine the hierarchies in relation to a second variable: the number of syllables in names. As was seen when considering names by the state with highest frequency, much structure is observable with this variable. Female names in Clusters 1 and 2 of Fig. 4A, which share a common predecessor node as the parent cluster, have similar temporal trends, with a high frequency in the early 20th century. In a fine-scale analysis, however, they separate into a branch whose names have fewer syllables (Cluster 1; e.g., Mary, Helen), and a branch whose names have more syllables (Cluster 2; e.g., Dorothy, Virginia). For male names, later names tend to have more syllables than earlier names (Fig. 4H).

In summary, the *NetStruct* analysis reveals relationships in co-occurrences of names, identifying names with similar temporal trends. The recoding of clustering hierarchies by additional variables—the state with highest frequency, and the number of syllables—illustrates the use of *NetStruct* in understanding attributes that correlate with, and potentially contribute to, relationships among entities. The visualization can potentially suggest analyses of other factors that influence the dynamics, including immigration, regional correlations, and differences in naming practices by state over time.

Discussion

Inspired by the potential of hierarchical clustering analyses to illuminate features of population-genetic variation, we have adapted the network-based clustering framework *NetStruct* for use in the analysis of cultural variation. In four examples, we have illustrated several aspects of the framework in applications to data matrices representing a set of entities, each associated with values of a set of characters. These applications demonstrate the potential of *NetStruct* to extract broad- and fine-scale relationships among entities. They illustrate the use of *NetStruct* to analyze relationships of geography with clustering patterns, to uncover the characters that drive relationships, and to understand effects on cultural data points of interest in specific scenarios. The algorithmic perspective incorporates flexibility in the design of similarity measures and in visualization schemes to aid the analysis.

Interpretations of data on cultural variation

The four examples illustrate the potential of *NetStruct* for producing novel visualizations to deepen the understanding of cultural entities—pronunciation repertoires of individuals, folklore repertoires of cultures, phoneme inventories of languages, and time series of name frequencies. The technique can uncover hierarchical features underlying the variation in cultural traits at different scales, and it enables the examination of different hierarchical levels. For example, the LAMSAS data, the data set among the four that has been studied for longest, has given rise to numerous analyses of dialect variation, often seeking to partition the Atlantic region into dialects (Lee & Jr, 1993; Nerbonne & Kretzschmar, 2003; Nerbonne 2015); our approach contributes to observing hierarchical divisions at multiple levels, to detecting spatially continuous variation beyond the level of dialects, and to identifying birth date as a variable that contributes to deviations from spatial patterns.

Fewer studies have examined the folklore data set that we have considered. With the use of normalized mutual information (NMI), we have shown that *NetStruct* can help to identify informative motifs for describing broad- and fine-scale structures of folklore variation. The recurrence of a shared motif in widely separated cultural groups has been useful for reconstructing cross-cultural contact and examining cultural diffusion. In this context, past studies have considered the diffusion of specific motifs, sometimes chosen as those that are widespread or that have particular cultural salience (Korotaev *et al.*, 2006; Berezkin, 2010; Ross *et al.*, 2013; Tehrani, 2013). Rather than choosing motifs based on prior significance, the NMI approach identifies motifs that are most informative about cultural groupings from patterns of motif occurrence alone. The identification using NMI of motifs of particular informativeness can further focus the choice of specific motifs for use in detailed analysis of diffusion patterns of folklore across worldwide cultural groups; studies such as those examining ‘The Tale of the Kind and the Unkind Girls’ (Ross *et al.*, 2013) and ‘Little Red Riding Hood’ (Tehrani, 2013) can be informative for interpreting patterns in well-known motifs, but studies of other motifs might be more informative for understanding cultural diffusion.

NetStruct requires little prior knowledge of data sets of interest. For the phonemes, as in the principal components analysis of Creanza *et al.* (2015), *NetStruct* identifies broad-scale geographic differentiation by a method that supposes no prior relationships among entities. Our analysis illustrates the potential to highlight distinctions of certain languages from their neighbors, finding that phonemic distinctiveness can reflect distinctiveness of one language in relation to others.

For first names, previous studies of the data have examined many aspects, including spatial correlations (Barucca *et al.*, 2015; Pomorski *et al.*, 2016) and phonemic influences (Berger *et al.*, 2012); our analyses of the state of greatest popularity and of patterns in syllables contribute further to understanding patterns in name frequencies. Some studies of naming patterns are model-based, assuming factors that drive the variation and incorporating these factors as variables in the models to compare to observed trends (Hahn & Bentley, 2003; Gureckis & Goldstone, 2009; Berger *et al.*, 2012; Kessler *et al.*, 2012; Acerbi & Bentley, 2014; O'Dwyer & Kandler, 2017); our approach can augment such studies by suggesting hypotheses that can be used in evaluating different generative models.

In our choices of examples for application of *NetStruct*, the four data sets had several features in common. First, in each case, entities corresponding to rows of the initial data matrix had a natural set of relationships reflected in the *NetStruct* hierarchy—geographic proximity of informants for the pronunciation data, geographic proximity of cultures for the folklore data, geographic proximity of languages for the phonemic data, and proximity in time of the period of greatest popularity for the data on names. Second, additional salient attributes of the entities were possible to consider—birth dates for pronunciation informants, locally specific components of folklore such as geographically restricted cultural practices and animal ranges, family memberships for languages, and states of greatest popularity and numbers of syllables for names. Additional data sets with spatial structure, temporal structure, or both, such as data on attributes of ceramics or other artifacts of material culture, or data on individual variation in word choices or other idiolectal variation, potentially provide natural examples as well. For future data sets, the existence of geographic and temporal structure and the availability of other meaningful attributes on entities of interest can be used to support use of *NetStruct* and to guide interpretation of the results that it produces.

Limitations and extensions

We have chosen to focus on similarity measures borrowed from genetics in which the sharing of a rare genetic variant between two individuals or populations suggests recent common ancestry. Similarly, for cultural data, in which shared descent is also a salient phenomenon, our use of a frequency-weighted trait-sharing similarity measures presupposes the potential importance of shared rare variants in characterizing relationships between entities. However, the choice of similarity measure occurs prior to application of *NetStruct*; the emphasis of similarity measures on shared rare variants can therefore be tuned as appropriate to a specific type of data. A possible systematic difference from the genetics context is that fast-evolving cultural data could generate more homoplasy than is seen for genetic markers (Tehrani & Collard, 2002; Haasl & Payseur, 2011), so that a shared rare variant could be less meaningful in cultural data than in genetic data. Distance-based hierarchical clustering studies in genetics have generally identified many shared features in population relationships irrespective of the similarity measure considered, even for fast-evolving genetic markers with significant homoplasy (Takezaki & Nei, 1996). In a preliminary analysis of the choice of similarity measure, considering the LAMSAS data, we see that generally similar patterns are obtained with two additional similarity measures: a measure that is not frequency-weighted, and a measure designed specifically for linguistic data (Figs. S1 and S2). With a specific scientific question and dataset, measures that encode aspects of similarity of greatest interest can be considered, and researchers can employ multiple similarity statistics to identify patterns that are robust and patterns that are distinctive to particular measures.

As in genetic studies that use tree-like models of population relationships, we have assumed that a hierarchical relationship between clusters exists, focusing on transferring the application of a hierarchical clustering method from population-genetic data to data on cultural variation. In cultural data, as is often seen in population-genetic data, the appropriate generative model that underlies the data need not be fully tree-like. Studies in population genetics have introduced methods for testing the suitability of evolutionary trees for explaining patterns of genetic variation, a key concept being the “treeness” of the data (Cavalli-Sforza & Piazza 1975; Patterson et al. 2012; Pickrell & Pritchard 2012). It would be of interest to develop comparable approaches for testing the extent to which a hierarchical structure from *NetStruct* explains cultural variation data; a permutation test of Greenbaum *et al.* (2016) for significance of clustering in a two-level *NetStruct* hierarchy containing a root and offspring nodes, devised in the population-genetic context, can potentially be adapted for arbitrary hierarchies and applied to data on cultural variation.

Conclusions: uses and applications of *NetStruct*

In population genetics, the interplay of evolutionary processes contributes to producing hierarchical patterns in genetic composition among populations. Similarly, in the study of cultural data, many forces interact to shape hierarchical trait variation. Interpreting the clustering results requires consideration of multiple interacting processes and phenomena, including global and local selection pressures on specific cultural variants (e.g. positive, negative, or balancing), linkage of multiple variants in “cultural complexes” (similar to genetic linkage), and random drift. As in the study of genetic data, geographic patterns need not uniquely identify the underlying processes; for example, similarly to the phenomenon of convergent evolution in genetic data, convergent evolution of cultural variants (e.g. Tehrani & Collard 2002; Mesoudi et al. 2006; Rogers & Ehrlich 2008) can produce a level of similarity that can be conflated with shared descent. For example, in our phonemic analysis in Figure 3, the potential for rapid change in languages can produce similarity in phonemes of otherwise distant languages. The cluster of languages colored in light green in Figure 3, which includes languages from sub-Saharan Africa, the Caucasus, and western North America, may result from convergent evolution combined with linkage of phoneme complexes that have developed independently. Consideration of the mechanistic processes underlying cultural data while incorporating domain-knowledge specific to datasets of interest is important in interpreting the hierarchical structure generated by *NetStruct*. Because cultural data often possess the type of geographic structure, temporal structure, hierarchical categorization, or defining attributes for which *NetStruct* results can be productively interpreted, patterns from *NetStruct* can be informative alongside other statistical methods for assessing specific generative models for cultural data.

As a non-model-based tool, the strength of the method lies in its potential as an exploratory approach for producing informative patterns, patterns that potentially inspire hypotheses about factors that drive the features of cultural variation. We suggest that the use of this exploratory approach should be accompanied by analyses of hypotheses based on additional methods and domain knowledge; analyses of data on variation in artifacts of culture can be productively advanced by adding *NetStruct* to the repertoire of the field of cultural evolution.

Methods

Between-informant similarity for LAMSAS pronunciation variation

We obtained the LAMSAS data from the project website (www.lap.uga.edu/Site/LAMSAS.html). To eliminate systematic effects of different interviewers, we considered only informants interviewed by the main interviewer, G. Lowman, who collected the earliest LAMSAS data (Nerbonne & Kleiweg, 2003). We therefore restricted attention to 839 informants interviewed during 1933-1942. Because words chosen for pronunciation differed across interviews, many words only appear in the records of a subset of informants. We considered only words collected for at least 700 informants, resulting in a list of 69 words.

Consider n informants and m words. Suppose word j has l_j distinct transcriptions, counting diacritics. Entry A_{ij} of the data matrix is a categorical variable that indicates the transcription of word j for informant i: $A_{ij} \in \{1,2,\ldots,l_j\}$ if the information of word j has been collected from informant i, or $A_{ij} = 0$ if word j is unavailable for informant i.

We computed a frequency-weighted transcription-sharing similarity, adapting the allele-sharing similarity for genetic data (Greenbaum *et al.*, 2019). For two informants i_1 and i_2, their frequency-weighted transcription-sharing similarity is calculated as

$$s_{i_1 i_2} = \frac{\sum_{j=1}^{m} (1-p^{j}_{A_{i_1 j}})^{I(A_{i_1 j}=A_{i_2 j})} I(A_{i_1 j}\neq 0 \wedge A_{i_2 j}\neq 0)}{\sum_{j=1}^{m} I(A_{i_1 j}\neq 0 \wedge A_{i_2 j}\neq 0)}, \qquad (1)$$

where $p^{j}_{A_{ij}}$ is the frequency of transcription A_{ij} for word j. The indicator function $I(\cdot)$ is 1 if the condition holds, and it is 0 otherwise. The similarity matrix S is obtained by normalization:

$$S_{i_1 i_2} = \frac{s_{i_1 i_2} - s_{min}}{s_{max} - s_{min}}, \qquad (2)$$

where $s_{\min} = \min_{i_1,i_2}\{s_{i_1 i_2}\}$ and $s_{\max} = \max_{i_1,i_2}\{s_{i_1 i_2}\}$.

In rare instances, two individuals have no shared words with data present. In these cases, we assigned for the similarity score the mean similarity of the remaining pairs.

Between-region similarity for folklore motif variation

We downloaded the folklore data from the Berezkin *et al.* database (http://www.ruthenia.ru/folklore/berezkin/). The database provides (in Russian) for each indexed motif, a list of all numbered regions in which the motif is present. Considering all 2495 motifs, we constructed the matrix of presence/absence entries, associating the region and motif names with matrix rows and columns, respectively.

We denote the $n \times m$ matrix by A, where n=65 is the number of regions and m=2459 is the number of motifs appearing in at least two regions. A_{ij}=1 if motif j appears in region i, and A_{ij}=0 otherwise. The pairwise frequency-weighted motif-sharing similarity for two regions i_1 and i_2 is calculated by a weighted Jaccard distance:

$$s_{i_1 i_2} = \frac{\sum_{j=1}^{m} (1-f_j) A_{i_1 j} A_{i_2 j}}{\sum_{j=1}^{m} (1-f_j) I_{(A_{i_1 j} \neq 0 \vee A_{i_2 j} \neq 0)}}. \qquad (3)$$

The quantity $f_j = \frac{1}{n} \sum_{i=1}^{n} A_{ij}$ is the frequency of motif j across all regions. Eq. 3 places greater weight on contributions of less frequent motifs and less weight on common motifs.

We applied the same normalization from Eq. 2 to obtain a normalized similarity matrix that we used in our analysis.

Normalized mutual information (NMI)

Denote two hierarchical clusterings on n entities by C^1 and C^2, Suppose they partition the same set of entities $E=\{e_1, e_2, \ldots, e_n\}$ into k and l clusters C_1^1, C_2^1 ..., C_k^1 and C_1^2, C_2^2 ..., C_l^2, respectively, where $\cup_{j=1}^{k} C_j^1 = \cup_{j=1}^{\ell} C_j^2 = E$. Note that for each clustering—for example, C—the clusters are not necessarily disjoint, so that each e_i can belong to multiple clusters C_j; indeed child clusters $C_{j'}$ are contained in parent clusters C_j, or $C_{j'} \subset C_j$. The NMI between these two hierarchical clusterings is then computed from the C_j^1 and C_j^2 following the procedure of Greenbaum *et al.* (2019). This approach is flexible in the sense that NMI can also be computed for subsets of the clusters in the hierarchy, rather than for the entire set of clusters. To address clustering at the finer scale of the hierarchy, we computed NMI for the set of leaf clusters at the tips of the hierarchy.

Between-language similarity for phoneme inventories

We obtained phoneme data from the supplement of Creanza *et al.* (2015). The similarity calculation follows that of the motif-sharing similarity, except that A now represents an $n \times m$ matrix of n=2082 languages and m=454 phonemes. Eq. 3 gives the similarity between a pair of languages, with f_j denoting the frequency of a phoneme among languages; we normalized the similarity matrix by Eq. 2 for our subsequent analysis.

Between-name similarity for name frequency profiles

We downloaded the name data from https://www.ssa.gov/oact/babynames/limits.html. For the analysis, performed separately for female and male names, a matrix entry A_{ij} tabulates the number of appearances of name i in year j, normalized by the total number of individuals in year j. We write A_{ij}=0 if name i is absent during year j, or if it is rare enough to have been omitted from the database for privacy reasons (fewer than 5 appearances nationally). For each pair of rows i_1, i_2 of A, we computed the Pearson correlation r_{i_1,i_2} between them, and transformed it to a value in [0,1] by $s_{i_1,i_2} = (r_{i_1,i_2} + 1)/2$.

To obtain syllable counts for individual names, two raters separately assigned the counts, discussing cases of disagreement to assign a number of syllables. We computed averages of the number of syllables for names in specific clusters.

States of highest frequency and regions of highest normalized frequency for names

Separately for female and male names, let B_{ijk} denote the number of appearances of name i in state k in year j. The frequency of name i in state k is calculated as $f_{ik} = \left(\sum_j B_{ijk}\right)/\left(\sum_{k'} \sum_j B_{ijk'}\right)$. The state of highest frequency for name i is obtained by $\mathrm{argmax}_k(f_{ik})$, which we denote the majority state in Fig. 4B and 4F.

The states are then grouped into four regions described in Fig. 4. For each region l containing a group of states, let C_{ijl} denote the number of appearances of name i in region l in year j, or $C_{ijl} = \sum_{k \in l} B_{ijk}$. For each year j, this number of appearances is divided by the total number of individuals in a region to obtain the fraction that it represents of all names in the region during year j: $C^*_{ijl} = C_{ijl}/\sum_{i'} C_{i'jl}$. Averaging across years, the normalized frequency of name i in region l is then calculated as $g_{il} = \left(\sum_j C^*_{ijl}\right)/\left(\sum_{l'} \sum_j C^*_{ijl'}\right)$. The region of highest normalized frequency for name i is obtained by $\mathrm{argmax}_l(g_{il})$, which we denote the majority region in Fig. 4C and 4G.

Acknowledgements

We thank E. Alimpiev for assistance with the folklore data and D. Cotter for assistance with the name data.

Authors' contributions

XL contributed conceptualization, methodology, investigation, visualization, and writing. NAR contributed conceptualization, methodology, supervision, and writing. GG contributed conceptualization, methodology, supervision, and writing.

Financial support

This work was supported by National Institutes of Health grant R01 HG005855 and National Science Foundation grant BCS-2116322 awarded to NAR, and a grant from the Hebrew University of Jerusalem Center for Interdisciplinary Data Science Research (CIDR) awarded to GG.

Conflicts of interest

Authors declare that they have no competing interests.

Data availability statement

The data that support the findings of this study are available as supplementary files. Functions for processing and visualizing *NetStruct* output are provided, as are example uses of those functions.

References

Acerbi, A., & Bentley, R. A. (2014). Biases in cultural transmission shape the turnover of popular traits. *Evolution and Human Behavior*, 35(3), 228–236.

Aplin, L. M. (2019). Culture and cultural evolution in birds: A review of the evidence. *Animal Behaviour*, 147, 179–187.

Atkinson, Q. D. (2011). Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science*, 332(6027), 346–349.

Barucca, P., Rocchi, J., Marinari, E., Parisi, G., & Ricci-Tersenghi, F. (2015). Cross-correlations of American baby names. *Proceedings of the National Academy of Sciences*, 112(26), 7943–7947.

Berezkin, Y. (2010). Tricksters trot to America: areal distribution of folklore motifs. *Folklore: Electronic Journal of Folklore*, 46, 125–142.

Berezkin, Yu. E., Borinskaya, S. A., Kuznetsova, A. V., & Sen’ko, O. V. (2009). Study of folklore and mythological traditions using intellectual data mining. *Pattern Recognition and Image Analysis*, 19(4), 630–633.

Berger, J., Bradlow, E. T., Braunstein, A., & Zhang, Y. (2012). From Karen to Katie: using baby names to understand cultural evolution. *Psychological Science*, 23(10), 1067–1073.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), P10008.

Bortolini, E., Pagani, L., Crema, E. R., Sarno, S., Barbieri, C., Boattini, A., Sazzini, M., da Silva, S. G., Martini, G., Metspalu, M., Pettener, D., Luiselli, D., & Tehrani, J. J. (2017). Inferring patterns of folktale diffusion using genomic data. *Proceedings of the National Academy of Sciences*, 114(34), 9140–9145.

Boyd, R., & Richerson, P. J. (1985). *Culture and the Evolutionary Process*. Chicago: University of Chicago Press.

Bromham, L. (2017). Curiously the same: Swapping tools between linguistics and evolutionary biology. *Biology & Philosophy*, 32(6), 855–886.

Cavalli-Sforza, L. L., & Feldman, M. W. (1981). *Cultural Transmission and Evolution*. Princeton, NJ: Princeton University Press.

Cavalli-Sforza, L. L., & Piazza, A. (1975). Analysis of evolution: evolutionary rates, independence and treeness. *Theoretical Population Biology*, 8(2), 127-165.

Creanza, N., Ruhlen, M., Pemberton, T. J., Rosenberg, N. A., Feldman, M. W., & Ramachandran, S. (2015). A comparison of worldwide phonemic and genetic variation in human populations. *Proceedings of the National Academy of Sciences*, 112(5), 1265–1272.

Fort, J., & Pérez-Losada, J. (2016). Can a linguistic serial founder effect originating in Africa explain the worldwide phonemic cline? *Journal of The Royal Society Interface*, 13(117), 20160185.

Fortescue, M. (2005). *Comparative Chukotko-Kamchatkan Dictionary*. Berlin: Mouton de Gruyter.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12), 7821–7826.

Gray, R. D., Bryant, D., & Greenhill, S. J. (2010). On the shape and fabric of human history. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1559), 3923–3933.

Greenbaum, G., Rubin, A., Templeton, A. R., & Rosenberg, N. A. (2019). Network-based hierarchical population structure analysis for large genomic data sets. *Genome Research*, 29(12), 2020–2033.

Greenbaum, G., Templeton, A. R., & Bar-David, S. (2016). Inference and analysis of population structure using genetic data and network theory. *Genetics*, 202(4), 1299–1312.

Gureckis, T. M., & Goldstone, R. L. (2009). How you named your child: understanding the relationship between individual decision making and collective outcomes. *Topics in Cognitive Science*, 1(4), 651–674.

Haasl, R. J., & Payseur, B. A. (2011). Multi-locus inference of population structure: A comparison between single nucleotide polymorphisms and microsatellites. *Heredity*, 106(1), 158–171.

Hahn, M. W., & Bentley, R. A. (2003). Drift as a mechanism for cultural change: An example from baby names. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(suppl_1), S120–S123.

Kessler, D. A., Maruvka, Y. E., Ouren, J., & Shnerb, N. M. (2012). You name it – how memory and delay govern first name dynamics. *PLoS One*, 7(6), e38790.

Kolodny, O., Feldman, M. W., & Creanza, N. (2018). Integrative studies of cultural evolution: Crossing disciplinary boundaries to produce new insights. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 373(1743), 20170048.

Korotaev, A. V., Berezkin, Y., Kozmin, A., & Arkhipova, A. (2006). Return of the white raven: Postdiluvial reconnaissance motif A2234.1.1 reconsidered. *Journal of American Folklore*, 119(472), 203–235.

Korotayev, A. V., Berezkin, Y. E., Borinskaya, S. A., Davletshin, A. I., & Khaltourina, D. A. (2017). Genes and myths: which genes and myths did the different waves of the peopling of Americas bring to the New World? In L. E. Grinin, A. V. Korotayev, eds., *History & Mathematics: Economy, Demography, Culture, and Cosmic Civilizations* (pp. 9–77). Volgograd: Uchitel Publishing House.

Kretzschmar Jr., W. A., McDavid, V. G., Lerud, T. K., & Johnson, E. (1993). *Handbook of the Linguistic Atlas of the Middle and South Atlantic States*. Chicago: University of Chicago Press.

Lee, J., & Kretzschmar Jr, W. A. (1993). Spatial analysis of linguistic data with GIS functions. *International Journal of Geographical Information Systems*, 7(6), 541–560.

Mesoudi, A. (2016). Cultural evolution: A review of theory, findings and controversies. *Evolutionary Biology*, 43(4), 481–497.

Mesoudi, A., Whiten, A., & Laland, K. N. (2006). Towards a unified science of cultural evolution. *Behavioral and Brain Sciences*, 29(4), 329-347.

Nerbonne, J. (2015). Various variation aggregates in the LAMSAS south. In M. D. Picone & C. E. Davies, eds., *Language Variety in the South III* (pp. 369-382). Tuscaloosa, AL: University of Alabama Press.

Nerbonne, J., & Kleiweg, P. (2003). Lexical distance in LAMSAS. *Computers and the Humanities*, 37(3), 339–357.

Nerbonne, J., & Kretzschmar, W. (2003*).* Introducing computational techniques in dialectometry. *Computers and the Humanities*, 37(3), 245–255.

O'Dwyer, J. P., & Kandler, A. (2017). Inferring processes of cultural transmission: The critical role of rare variants in distinguishing neutrality from novelty biases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1735), 20160426.

Pagel, M. (2009). Human language as a culturally transmitted replicator. *Nature Reviews Genetics*, 10(6), 405–415.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., & Reich, D. (2012). Ancient admixture in human history. *Genetics*, 192(3), 1065-1093.

Pérez-Losada, J., & Fort, J. (2018). A serial founder effect model of phonemic diversity based on phonemic loss in low-density populations. *PLoS One*, 13(6), e0198346.

Pickrell, J., & Pritchard, J. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics* 8: e1002967.

Pomorski, M., Krawczyk, M. J., Kułakowski, K., Kwapień, J., & Ausloos, M. (2016). Inferring cultural regions from correlation networks of given baby names. *Physica A: Statistical Mechanics and Its Applications*, 445, 169–175.

Rendell, L., & Whitehead, H. (2005). Spatial and temporal variation in sperm whale coda vocalizations: Stable usage and local dialects. *Animal Behaviour*, 70(1), 191–198.

Rogers, D. S., & Ehrlich, P. R. (2008). Natural selection and cultural rates of change. *Proceedings of the National Academy of Sciences*, 105(9), 3416–3420.

Ross, R. M., Greenhill, S. J., & Atkinson, Q. D. (2013). Population structure and cultural geography of a folktale in Europe. *Proceedings of the Royal Society B: Biological Sciences*, 280(1756), 20123065.

Takezaki, N., & Nei, M. (1996). Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics*, 144(1), 389–399.

Tehrani, J., & Collard, M. (2002). Investigating cultural evolution through biological phylogenetic analyses of Turkmen textiles. *Journal of Anthropological Archaeology*, 21(4), 443–463.

Tehrani, J. J. (2013). The phylogeny of Little Red Riding Hood. *PLoS One*, 8(11), e78871.

Thuillard, M., Quellec, J. L., D'huy, J., & Berezkin, Y. (2018). A large-scale study of world myths. *Trames: A Journal of the Humanities and Social Sciences*, 22(4), 407–424.

Westphal, E. O. J. (1971). The click languages of Southern and Eastern Africa. In T. A. Sebeok, ed., *Linguistics in Sub-Saharan Africa* (pp. 367–420). The Hague: De Gruyter Mouton.

Figures

Fig. 1: Hierarchical features of variation in English pronunciation in the middle and south Atlantic region of the United States. (A) Hierarchical tree of the pronunciation similarity network. Informants are marked on the map by the color of the finest-scale cluster to which they belong. (B, C) Two major clusters detected at the first level of the hierarchy in (A), each re-colored with the full color interval. (D, E) Two finer-scale clusters of the hierarchy in (C). In these panels, colors are assigned based on placement in the area of the hierarchy circled in (C), with all descendants of a child in the circled area assigned the same color. The colors in (D) correspond to 1/4 and 3/4 on the unit interval, and the colors in (E) correspond to 1/14, 3/14, 5/14, 7/14, 9/14, 11/14, and 13/14. For convenience, the child clusters associated with specific internal nodes in the tree diagrams are numbered. Birth-year distributions of informants in these child clusters appear at right.

Fig. 2: Hierarchical features of variation in folklore motifs across cultures. (A) Hierarchical tree of the motif similarity network. Regions are marked on the map by the color of the finest-scale cluster to which they belong. (B) Distributions of normalized mutual information (NMI) between hierarchies extracted from sampled subsets of motifs and from all motifs, with 100 subsets of 20, 50, 100, and 500 motifs each. (C) Geographic distributions of five motifs that occur most frequently in the 200 of 5,000 subsets of 20 motifs that produce hierarchies with highest NMI to the hierarchy produced by all motifs. These motifs drive the hierarchy at higher levels, separating regions into major clusters. (D) Distributions of NMI between the leaves of hierarchies extracted from subsets of motifs (those from B) and the leaves extracted from all motifs. (E) Geographic distributions of three motifs that occur most frequently in the 200 subsets that produce hierarchies whose leaf clusters produce highest NMI to those produced by all motifs. These motifs are more specific to the hierarchy in lower levels and potentially capture fine-scale regional differences.

Fig. 3: Hierarchical features of phonemic variation. (A) Hierarchical tree of the phoneme similarity network. Major branches that contain most of the languages are assigned distinct colors, and other branches are colored gray. (B) Language map. Languages are marked by the color of the finest-scale cluster to which they belong. Three regions are magnified. (C) Northeast Asia. (D) Northeastern Siberia. (E) East Africa.

Fig. 4: Hierarchical features of time series for frequencies of female names (A-D) and male names (E-H). (A, E) Hierarchical tree of similarity in time series for name frequencies. Major branches are assigned distinct colors. Time series of annual national frequencies appear below the trees, with two names selected from each major branch highlighted. Node area is proportional to the number of names in a cluster, except that clusters containing greater than 25 names are set to a fixed size and are colored half-transparently. (B, F) Recoding of the hierarchies in (A, E) by states of highest frequency. Each cluster shows a pie chart tabulating the states in which names in the cluster have the highest frequency. Time series of name frequencies appear below the hierarchies. (C, G) Recoding of the hierarchies in (A, E) by regions of highest normalized frequency. The states are grouped into four regions: West (AK, AZ, CA, CO, HI, ID, MT, NM, NV, OR, UT, WA, WY), Midwest (IA, IL, IN, KS, MI, MN, MO, ND, NE, OH, SD, WI), South (AL, AR, DC, DE, FL, GA, KY, LA, MD, MS, NC, OK, SC, TN, TX, VA, WV), and Northeast (CT, MA, ME, NH, NJ, NY, PA, RI, VT). The normalized frequency of a name in a region is the count of the name in the region normalized by the total number of individuals in the region. The steps to obtain the states of highest frequency and regions of highest normalized frequency are described in *Methods.* (D, H) Recoding of the hierarchies in (A, E) by mean number of syllables of names in clusters.

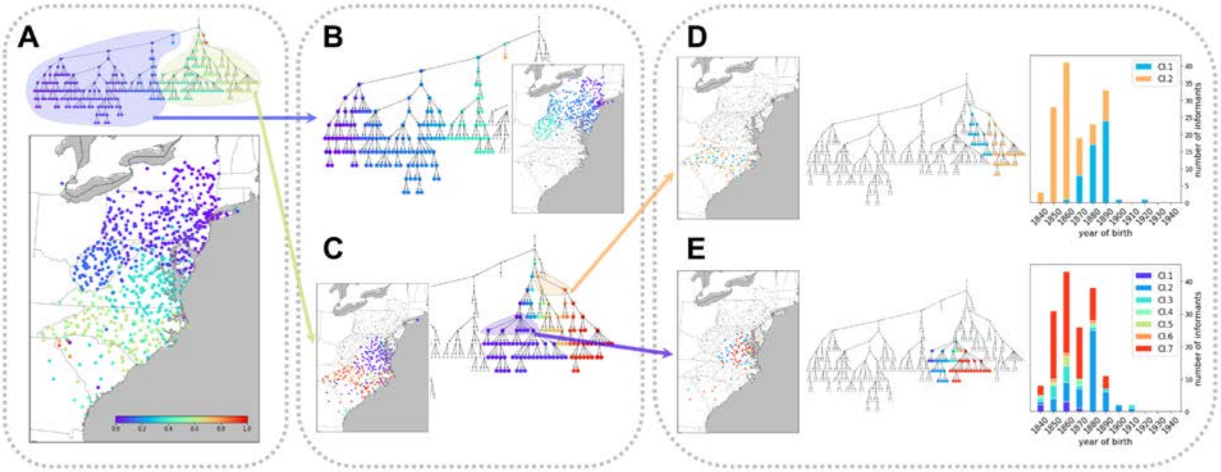
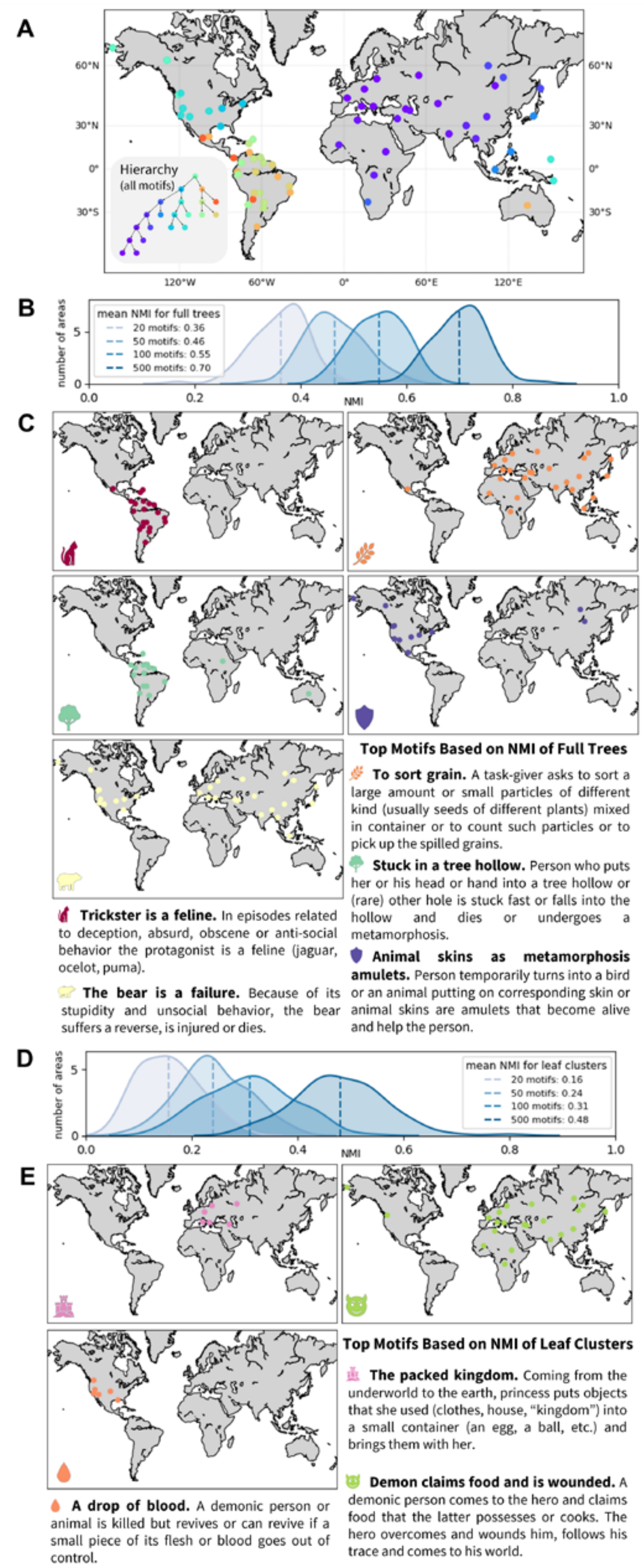

A
B
C
D
E
0.0
0.2
0.4
0.6
0.8
1.0
CL.1
CL.2
number of informants
year of birth
1840
1850
1860
1870
1880
1890
1900
1910
1920
1930
1940
0
5
10
15
20
25
30
35
40
CL.1
CL.2
CL.3
CL.4
CL.5
CL.6
CL.7
number of informants
year of birth

Figure 1.



Top Motifs Based on NMI of Full Trees

To sort grain. A task-giver asks to sort a large amount or small particles of different kind (usually seeds of different plants) mixed in container or to count such particles or to pick up the spilled grains.

Stuck in a tree hollow. Person who puts her or his head or hand into a tree hollow or (rare) other hole is stuck fast or falls into the hollow and dies or undergoes a metamorphosis.

Animal skins as metamorphosis amulets. Person temporarily turns into a bird or an animal putting on corresponding skin or animal skins are amulets that become alive and help the person.

Trickster is a feline. In episodes related to deception, absurd, obscene or anti-social behavior the protagonist is a feline (jaguar, ocelot, puma).

The bear is a failure. Because of its stupidity and unsocial behavior, the bear suffers a reverse, is injured or dies.

Top Motifs Based on NMI of Leaf Clusters

The packed kingdom. Coming from the underworld to the earth, princess puts objects that she used (clothes, house, "kingdom") into a small container (an egg, a ball, etc.) and brings them with her.

Demon claims food and is wounded. A demonic person comes to the hero and claims food that the latter possesses or cooks. The hero overcomes and wounds him, follows his trace and comes to his world.

A drop of blood. A demonic person or animal is killed but revives or can revive if a small piece of its flesh or blood goes out of control.

Figure 2.

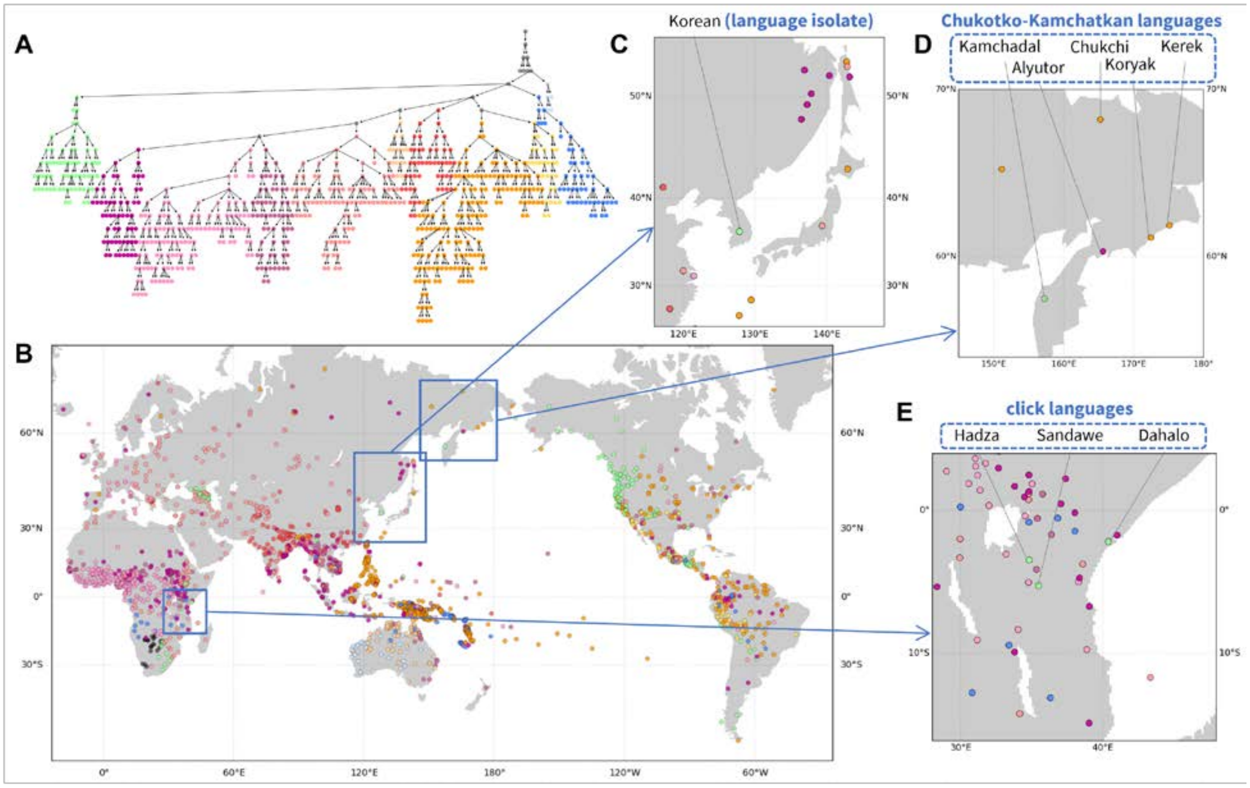

A
B
C
Korean (language isolate)
D
Chukotko-Kamchatkan languages
Kamchadal
Alyutor
Chukchi
Koryak
Kerek
E
click languages
Hadza
Sandawe
Dahalo

Figure 3.

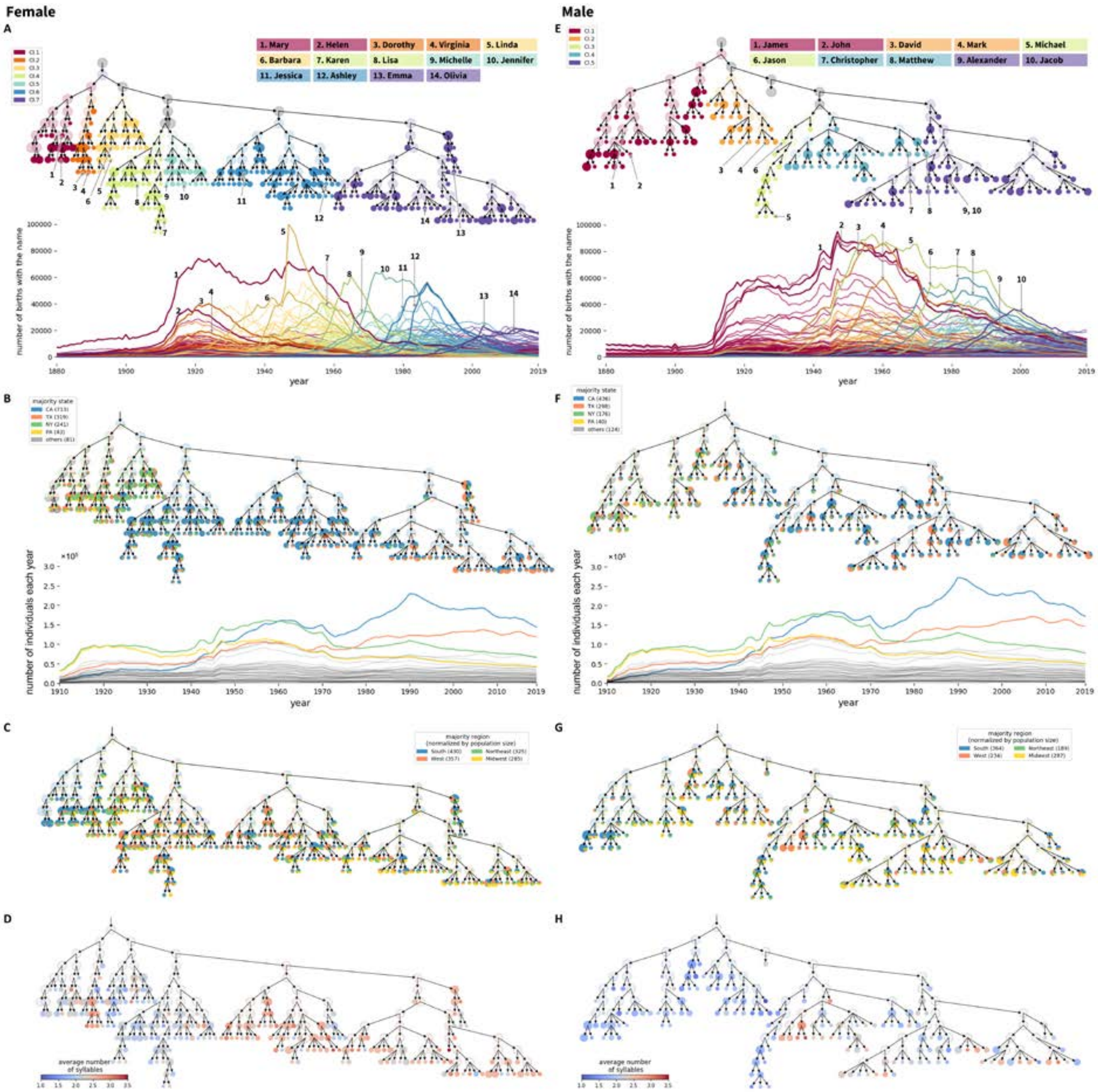

Female
A
Cl.1
Cl.2
Cl.3
Cl.4
Cl.5
Cl.6
Cl.7
1. Mary
2. Helen
3. Dorothy
4. Virginia
5. Linda
6. Barbara
7. Karen
8. Lisa
9. Michelle
10. Jennifer
11. Jessica
12. Ashley
13. Emma
14. Olivia
number of births with the name
100000
80000
60000
40000
20000
0
1880
1900
1920
1940
1960
1980
2019
year
B
majority state
CA (713)
TX (319)
NY (241)
PA (42)
others (81)
number of individuals each year
×10⁵
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2019
year
C
majority region
(normalized by population size)
South (430)
West (357)
Northeast (325)
Midwest (285)
D
average number of syllables
1.0
1.5
2.0
2.5
3.0
3.5
Male
E
Cl.1
Cl.2
Cl.3
Cl.4
Cl.5
1. James
2. John
3. David
4. Mark
5. Michael
6. Jason
7. Christopher
8. Matthew
9. Alexander
10. Jacob
9, 10
F
majority state
CA (436)
TX (298)
NY (176)
PA (40)
others (124)
G
majority region
(normalized by population size)
South (364)
West (234)
Northeast (189)
Midwest (287)
H
average number of syllables

Figure 4.