A METHODOLOGY FOR MODELING EMERGENCE IN COMPLEX SYSTEMS

By

Nathan Brugnone

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Community Sustainability—Doctor of Philosophy
Computational Mathematics, Science, & Engineering—Dual Major

2023

ABSTRACT

Understanding how social and ecological systems interact and work together as complex adaptive systems is essential for understanding the emergence of environmental and social systems states. However, complex systems domain general methods are not readily accessible to researchers and policymakers limiting our ability to both understand and intervene to address contemporary social and environmental problems. At the same time, rapid new developments in machine learning (ML) have allowed a sea change in extracting information from disparate datasets to make useful for human decision-making. This dissertation develops and studies novel ML methods to understand the emergent properties of complex systems. By proposing new approaches, we show how these emerging technologies can be applied to both (1) theoretical and (2) real-world problems. The first chapter introduces a domain-general unsupervised machine learning method for identifying clusters in data with hierarchical structure like that found in complex systems. Chapter 2 utilizes complex systems theory and theoretical machine learning to drive development of a theoretical framework and method for understanding graphical models of complex systems. The final chapter demonstrates how these methods can be applied through a case study of maternal and child healthcare (MCH) in Gombe, Nigeria.

Complex systems exhibit hierarchical structure that can be studied at various scales. Individuals are members of families who are in turn members of communities which are members of local government areas and so on. Chapter 1 proposes a clustering method that coarsens fine structures to reveal nested hierarchies in complex data. The method is presented in a theoretical framework based upon graph signal processing. The performance of this method is assessed on canonical ground-truth datasets. It is then shown to identify novel structure in real-world complex system data.

Complex systems can also be characterized by the patterns of emergent phenomena that they generate. These patterns often exhibit self-similarity which can be modeled with stochastic processes called textures. Chapter 2 introduces a theoretical framework that gen-

eralizes statistically self-similar random fields to graphs, again via the graph signal processing paradigm. Perhaps surprisingly, this generalization is shown to facilitate the classification of cognitive maps based upon structure. We model cognitive maps with samples from random graph families. We find that the statistical model produces sufficiently rich features to enable the accurate classification of these random graph families, thereby paving the way to application in unsupervised, real-world contexts.

Chapter 3 builds upon Chapters 1 and 2 by introducing a novel clustering method and comparing its performance with two others on real-world complex systems data. The data consists of value-laden statements made by expectant mothers and fathers in Gombe, Nigeria about utilization of the local maternal and child health (MCH) system. We take a methodological approach to identifying groups of individuals expressing similar values, which we seek to improve wisdom-of-crowds estimates of healthcare utilization. Similarities and differences among the identified values-based subpopulations provide insight into challenges and potentials for application of machine learning in similar development contexts.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the support of a great number of generous people. I am indebted to Steven Gray for academically adopting me and cooking an awesome crab dinner around which this document coalesced. Steven introduced me to James Gentile and the amazing complex & social systems research group at Two Six Technologies where some of the research in this dissertation has been conducted. I owe a great debt of gratitude to Matt Hirn and Robby Richardson whose patient guidance, kindness, and curiosity enabled me to pursue ideas at the intersection of complex systems modeling and machine learning. Matt's Herculean teaching and research efforts set a high bar that continues to inspire his students. Shout out to Dirk Colbry who continually welcomed me to teach beside him and proved that mathematics and data science are human-centered disciplines that can improve our well-being. Special thanks to Michael Murillo and his agent-based modeling group for many interesting discussions and encouragement. Thanks also to the MSU Modeling Ecological and Social Systems faculty Laura Schmitt Olabisi and Arika Ligmann-Zielinska whose collaborative approaches continue to inform my perspective. Thanks to Heather Williamson, Gail Vander Stoep, and the staff in CSUS and CMSE for turning untimely paperwork into a fully functioning degree program. Thanks to the members of my research groups: Timmy, Payam, Carissa, Mahdi, Anna, Mike et al. To everyone involved with the Sustainable Michigan Endowed Project–Pat and Paul, Jessica and Laura (CC 'holla), Kyle, Zach, Bethany et al.–thank you for the formative experiences. Thanks also to my mom, dad, and brothers, Bobby and Thomas, not only for the formative experiences, but also for the interest and support–thank you. Thanks to my wife, Danielle, for believing in and pushing me to stick with it through difficult times. To my youngest son, Miles, thank you for the cuddles. And very special thanks to my oldest son, Asa, whose excitement and kindness have been absolutely essential to seeing this dissertation through.

TABLE OF CONTENTS

CHAPTER 1

COARSE GRAINING OF DATA VIA INHOMOGENEOUS DIFFUSION CONDENSATION

1.1 Abstract

Big data often has emergent structure that exists at multiple levels of abstraction, which are useful for characterizing complex interactions and dynamics of the observations. Here, we consider multiple levels of abstraction via a multiresolution geometry of data points at different granularities. To construct this geometry we define a time-inhomogeneous diffusion process that effectively condenses data points together to uncover nested groupings at larger and larger granularities. This inhomogeneous process creates a deep cascade of intrinsic low pass filters on the data affinity graph that are applied in sequence to gradually eliminate local variability while adjusting the learned data geometry to increasingly coarser resolutions. We provide visualizations to exhibit our method as a "continuously-hierarchical" clustering with directions of eliminated variation highlighted at each step. The utility of our algorithm is demonstrated via neuronal data condensation, where the constructed multiresolution data geometry uncovers the organization, grouping, and connectivity between neurons.

1.2 Introduction

A fundamental task in data analysis is to characterize variability that separates informative data relations from disruptive ones, e.g., due to noise or collection artifacts. In predictive tasks such as classification, for example, one might seek to extract and preserve information that enhances class separation, while eliminating intra-class variance. However, in descriptive tasks and data exploration, such knowledge does not *a priori* exist, and instead data processing methods must detect emergent patterns that encode meaningful abstractions of the data. Furthermore, it is often the case that data abstraction cannot be conducted at a single scale, and instead one must consider multiresolution data representations that generate several scales of abstraction – each emphasizing different properties in the data.

The need for multiresolution data representations is of particular importance in biomedical data exploration, where recent technological advances introduce vast amounts of unlabeled data to be explored by limited numbers of domain experts. For example, in single-cell transcriptomics, high-throughput genomic and epigenetic assays have led to an explosion in high-dimensional biological data measured from various systems including imaging (Giesen et al., 2014; Angelo et al., 2014), mass cytometry (Bendall et al., 2011), and scRNA-seq (Shapiro et al., 2013; Kolodziejczyk et al., 2015). To fully utilize this transformative big data availability, computational methods are needed that leverage the intrinsic data geometry (e.g., using manifold learning techniques (Moon et al., 2018)) to enable exploratory analysis upon it).

A common approach towards data abstraction is to use clustering algorithms that provide coarse-grained representations of the data by grouping data points into salient clusters (Levine et al., 2015; Galluccio et al., 2013; Von Luxburg, 2007), either at a single scale or hierarchically (see Section 1.3). However, standard clustering algorithms such as k-means (Lloyd, 1982; Kanungo et al., 2002) or expectation maximization (EM) (Moon, 1996) have many limitations. For example, they fail to perform well on high-dimensional data, or they require a number of assumptions about the underlying structure of the data (Ng et al., 2002). In particular, a primary challenge in clustering is determining the optimal number of clusters or groups. Many algorithms require the user to explicitly choose the number of clusters (as in k-means) or tune a parameter that directly relates to the number of detected clusters (e.g., as in Phenograph (Levine et al., 2015)). In exploratory settings, this makes it particularly challenging to detect small, unique, or otherwise rare data type clusters, and extract new knowledge from them.

Here, we present a new approach to address the challenge of multiscale data coarse graining by using a data-driven time-inhomogeneous diffusion process, which we call diffusion condensation. Our proposed diffusion condensation process learns a "continuous hierarchy" of coarse-grained representations by iteratively contracting the data-points towards a time-

varying data manifold that represents increasingly coarser resolutions. At each iteration, the data points move to the center of gravity of their local neighbors as defined by this data-driven diffusion process (Coifman and Lafon, 2006; Nadler et al., 2005). This in turn alters the next steps of the diffusion process to reflect the new data positions. Across iterations, this construction creates a time-inhomogeneous Markov process on the data, which represents the changing affinities between data points, along with changing granularities. The process eventually collapses the entire data set to a single point. However, intermediate steps in this process produce coarse-grained data representations at particular granularities or abstraction levels. Importantly, our results show that distinct clusters emerge at different scales and each data point (e.g., each cell in transcriptomic data) is represented by a time series of feature vectors that capture multiresolution information in the data. Therefore, the data embedding provided by the constructed diffusion condensation process can be thought of as a dynamic video, as opposed to static snapshots provided by traditional manifold learning, such as diffusion maps (Coifman and Lafon, 2006) and other dimensionality reduction methods (Van Der Maaten et al., 2009).

1.3 Related Work

Typical attempts at providing multiscale data abstraction or summarization rely on hierarchical clustering, which is a family of methods that attempts to derive a tree of clusters based on either recursive agglomeration of datapoints or recursive splitting. Agglommerative methods include the popular linkage clustering, or community detection methods including the Louvain Method (Blondel et al., 2008). Splitting based approaches include recursive bisection (Dasgupta et al., 2006) and divisive analysis clustering (Kaufman and Rousseeuw, 2009). At each iteration, these methods explicitly attempt to discover the best split or merge at each iteration, thereby forcing points together or apart as the case may be. Diffusion condensation by contrast does not force any splits or mergers at any iteration and simply allows datapoints to come together naturally via repeated condensation steps. Thus, there may be many iterations in which a cluster of datapoints remains distinct from other clusters.

This time length under which the cluster persists can itself be a metric of the distinctness of a cluster, and the agglomeration of all such cluster persistence times creates a diagram similar to those created in persistent homology (Wasserman, 2018; Kwitt et al., 2015). Thus the hierarchical tree created by diffusion condensation (displayed as a Sankey diagram in Figure 1.4) has branches whose lengths are meaningful in terms of cluster separation.

1.4 Preliminaries

1.4.1 Manifold learning

High dimensional data can often be modeled as originating from a sampling $Z = \{z_i\}_{i=1}^N \subset \mathcal{M}^d$ of a d dimensional manifold $\mathcal{M}^d$ that is mapped to $n \gg d$ dimensional observations $X = \{x_1, \ldots, x_N\} \subset \mathbb{R}^n$ via a nonlinear function $x_i = f(z_i)$. Intuitively, the reason for this phenomenon is that data collection measurements (modeled here via f) typically result in high dimensional observations, even when the intrinsic dimensionality, or degrees of freedom, in the data is relatively low. This manifold assumption is at the core of the vast field of manifold learning (e.g., (Moon et al., 2018; Coifman and Lafon, 2006; Van Der Maaten et al., 2009; Izenman, 2012), and references therein), which leverages the intrinsic data geometry, modeled as a manifold, for exploring and understanding patterns, trends, and structure in data.

1.4.2 Diffusion geometry

In Coifman and Lafon (2006), diffusion maps were proposed as a robust way to capture intrinsic manifold geometry in data using random walks that aggregate local affinity to reveal nonlinear relations in data and allow their embedding in low dimensional coordinates. These local affinities are commonly constructed using a Gaussian kernel

$$\mathbf{K}(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{\varepsilon}\right), \quad i, j = 1, ..., N \tag{1.1}$$

where $\mathbf{K}$ is an $N \times N$ Gram matrix whose (i, j) entry is denoted by $\mathbf{K}(x_i, x_j)$ to emphasize the dependency on the data X. The bandwidth parameter ε controls neighborhood sizes. A diffusion operator is defined as the row-stochastic matrix $\mathbf{P} = \mathbf{D}^{-1}\mathbf{K}$ where $\mathbf{D}$ is a diagonal

matrix with $\mathbf{D}(x_i, x_i) = \sum_j \mathbf{K}(x_i, x_j)$, which is referred to as the degree of x_i. The matrix $\mathbf{P}$ defines single-step transition probabilities for a time-homogeneous diffusion process (which is a Markovian random walk) over the data, and is thus referred to as the diffusion operator. Furthermore, as shown in Coifman and Lafon (2006), powers of this matrix $\mathbf{P}^t$, for $t > 0$, can be used for multiscale organization of X, which can be interpreted geometrically when the manifold assumption is satisfied.

1.4.3 Diffusion filters

While originally conceived for dimensionality reduction via the eigendecomposition of the diffusion operator, recent works (e.g., Van Dijk et al. (2018); Lindenbaum et al. (2018); Gama et al. (2019); Gao et al. (2019)) have extended the diffusion framework of Coifman and Lafon (2006) to allow processing of data features by directly using the operator $\mathbf{P}$. In this case, $\mathbf{P}$ serves as a smoothing operator, and may be regarded as a generalization of a low-pass filter for either unstructured or graph-structured data. Indeed, consider a vector $\mathbf{v} \in \mathbb{R}^N$ that we think of as a signal $\mathbf{v}(x_i)$ over X. Then $\mathbf{P}\mathbf{v}(x_i)$ replaces the value $\mathbf{v}(x_i)$ with a weighted average of the values $\mathbf{v}(x_j)$ for those points x_j such that $\|x_i - x_j\| = O(\sqrt{\varepsilon})$. Applications of this approach include data denoising and imputation (Van Dijk et al., 2018), data generation (Lindenbaum et al., 2018), and graph embedding with geometric scattering (Gama et al., 2019; Gao et al., 2019).

1.5 Diffusion Condensation

1.5.1 Time inhomogeneous heat diffusion

The matrix $\mathbf{P}$ defines the transition probabilities of a random walk over the data set X. Computing powers of $\mathbf{P}$ runs the walk forward, so that $\mathbf{P}^t$ gives the transition probabilities of the t-step random walk. Since the same transition probabilities are used for every step of the walk, the resulting diffusion process is time homogeneous.

A time inhomogeneous diffusion process arises from an inhomogeneous random walk in which the transition probabilities change with every step. Its t-step transition probabilities

are given by

$$\mathbf{P}^{(t)} = \mathbf{P}_t \mathbf{P}_{t-1} \cdots \mathbf{P}_1$$

where $\mathbf{P}_k$ is the Markov matrix that encodes the transition probabilities at step k.

Suppose the data set X has an additional parameter $t = 0, 1, 2, \ldots$ that results from measurements $X(t) = \{x_1(t), \ldots, x_N(t)\}$ of a time-varying manifold $\mathcal{M}^d(\tau)$ at discretely sampled times $\tau_t = \varepsilon t$. Let $\mathbf{P}_t$ be the resulting Markov matrix derived from $X(t)$, constructed according to the anisotropic diffusion process of (Coifman and Lafon, 2006, Section 3) (which is similar to the construction described in Section 1.4). One can show (Marshall and Hirn, 2018) the resulting inhomogeneous diffusion process $\mathbf{P}^{(t)}$ approximates heat diffusion over the time varying manifold $\mathcal{M}^d(\tau)$. The singular vectors of this process can be used to construct a so-called time coupled diffusion map, which gives time-space geometric summaries of the data X.

The perspective of Marshall and Hirn (2018) is that the data is intrinsically time varying. However, one can also start with a static data set X and construct a series of deformations of the data. In this paper we take the latter perspective and deform the data set according to an imposed, data driven time inhomogeneous process $\mathbf{P}^{(t)}$ that reduces variability within the data over time. The resulting process is referred to as condensation, and is described in the next section.

1.5.2 The diffusion condensation process

Recall from Section 1.4 that the application of the operator $\mathbf{P}$ to a vector $\mathbf{v}$ averages the values of $\mathbf{v}$ over small neighborhoods in the data. In the case of data $X = \{x_1, \ldots, x_N\} \subset \mathbb{R}^n$ measured from an underlying manifold $\mathcal{M}^d$ with the model $x_i = f(z_i)$ for $z_i \in \mathcal{M}^d$, this averaging operator can be directly applied to the coordinate functions $f = (f_1, \ldots, f_n)$. Let $\mathbf{f}_k \in \mathbb{R}^N$ be the vector corresponding to the coordinate function f_k evaluated on the data samples, i.e., $\mathbf{f}_k(z_i) = f_k(z_i)$. The resulting description of the data is given by $\bar{X} = \{\bar{x}_1, \ldots, \bar{x}_N\}$ where $\bar{x}_i = (\mathbf{P}\mathbf{f}_1(z_i), \ldots, \mathbf{P}\mathbf{f}_n(z_i))$. The coordinates of $\bar{X}$ are smoothed versions of the coordinates of X, which dampens high frequency variations in the coordinate functions

and thus removes small perturbations in the data. This smoothing technique is used in Van Dijk et al. (2018) to impute and denoise data.

Here we consider not only the task of eliminating variability that originates from noise, but also coarse graining the data coordinates to provide multiple resolutions of the captured information in them. Therefore, we aim to gradually eliminate local variability in the data using a time inhomogeneous diffusion process that refines the constructed diffusion geometry to the coarser resolution as time progresses. This condensation process proceeds as follows. Let $X(0) = X$ be the original data set with Markov matrix $\mathbf{P}_0 = \mathbf{P}$ and $X(1) = \bar{X}$ the coordinate-smoothed data described in the previous paragraph. We can iterate this process to further reduce the variability in the data by computing the Markov matrix $\mathbf{P}_1$ using the coordinate representation $X(1)$. A new coordinate representation $X(2)$ is obtained by applying $\mathbf{P}_1$ to the coordinate functions of $X(1)$. In general, one can apply the process for an arbitrary number of steps, which results in the condensation process. Let $X(t)$ be the coordinate representation of the data after $t \geq 0$ steps so that $X(t) = \{x_1(t), \ldots, x_N(t)\}$ with $x_i(t) = (\mathbf{f}_1^{(t)}(z_i), \ldots, \mathbf{f}_n^{(t)}(z_i))$, where $\mathbf{f}_k^{(0)} = \mathbf{f}_k$. We obtain $X(t+1)$ by applying $\mathbf{P}_t$, the Markov matrix computed from $X(t)$, to the coordinate vectors $\mathbf{f}_k^{(t)}$. This process results in:

$$\mathbf{f}_k^{(t+1)} = \mathbf{P}_t \mathbf{f}_k^{(t)} = \mathbf{P}_t \mathbf{P}_{t-1} \cdots \mathbf{P}_1 \mathbf{P}_0 \mathbf{f}_k, \quad t \geq 0 \tag{1.2}$$

From (1.2) we see the coordinate functions of the condensation process at time $t+1$ are derived from the imposed time inhomogeneous diffusion process $\mathbf{P}^{(t)} = \mathbf{P}_t \cdots \mathbf{P}_0$. The low pass operator $\mathbf{P}_t$ applies a localized smoothing operation to the coordinate functions $\mathbf{f}_k^{(t)}$. Over the entire condensation time, however, the original coordinate functions $\mathbf{f}_k$ are smoothed by the cascade of diffusion operators $\mathbf{P}_t \cdots \mathbf{P}_0$. This process adaptively removes the high frequency variations in the original coordinate functions. The effect on the data points X is to draw them towards local barycenters, which are defined by the inhomogeneous diffusion process. Once two or more points collapse into the same barycenter, they are identified as being members of the same cluster. In Section 1.6 we demonstrate condensation's dynamic data deformations to remove variability and collapse points into clusters.

Input : $X \leftarrow$ NxM matrix of N data points, M features
$\epsilon \leftarrow$ initial filter bandwidth
Output : $X_t \leftarrow$ NxM data matrix after t condensations
begin
  $i \leftarrow 0$; $i_{prev} \leftarrow -2$
  $Q' \leftarrow I_N$
  $Q_{diff} \leftarrow \infty$
  labels $\leftarrow$ `Range`$(0, N)$
  while $i - i_{prev} > 1$ **do**
    $i_{prev} \leftarrow i$
    while $Q_{diff} >= \ 1 \times 10^{-4}$ **do**
      $i \leftarrow i + 1$
      $D \leftarrow$ `Distance`(X)
      `Merge`(labels[`Where`$(D < 1 \times 10^{-3})$])
      $A \leftarrow$ `Affinity`(D)
      $Q \leftarrow$ `Diag`(`RowSum`(A))
      $K \leftarrow Q^{-1}AQ^{-1}$
      $P \leftarrow$ `RowNormalize`(K)
      $X \leftarrow P \times X$
      $Q_{diff} \leftarrow ||$`Diag(`Q`) - Diag(`Q`')`$||_{l^\infty}$
      $Q' \leftarrow Q$
    end
    $\epsilon \leftarrow \epsilon \times 2$
    $Q_{diff} \leftarrow \infty$
  end
end

Algorithm 1.1 Condensation

1.5.3 Algorithm

Pseudocode is provided in Algorithm 1.1. Although not strictly necessary, cluster convergence may be accelerated by increasing the bandwidth, ϵ, when the l^∞-norm of the difference between densities of the previous and current iterations, $\mathbf{Diag}(Q')$ and $\mathbf{Diag}(Q)$, falls below a threshold.

The present implementation provides proof-of-concept. We see computational complexity is dominated by matrix multiply and is $\mathcal{O}(n^4)$ when $t \geq n$. Thus, more research is needed to scale the algorithm.

1.6 Properties of Diffusion Condensation

1.6.1 Cluster self-organization

Unlike other state-of-the-art clustering algorithms, such as k-means, diffusion condensation does not require the user to *a priori* choose a potentially arbitrary number of data clusters to find. Rather, condensation grows self-organizing cluster hierarchies that emerge through local interactions among the data manifold's sampling density and curvature variation. To disentangle and illustrate such properties, we provide condensation video stills in Figure 1.1. To begin we highlight the hyperuniformly-sampled (i.e., grid-sampled) circle manifold on the top-left of Figure 1.1, which demonstrates the base case of homogeneous data density and constant curvature. Note the absence of cluster formation. Comparing this to the hyperuniformly-sampled ellipse on the right of Figure 1.1, we observe the formation of nontrivial condensation clusters, particularly in the regions of high curvature.

Similarly, the uniformly-sampled circle manifold of constant curvature in the bottom-left of Figure 1.1 exhibits local cluster formation. Hence, we conjecture that nontrivial data density or curvature variation are sufficient conditions for the formation of diffusion condensates.

1.6.2 Cluster characterization via spectral decay

In addition to still frames, it is enlightening to consider cluster formation via its correspondence with the spectral decay. Figure 1.2 demonstrates that data condensation corresponds with sudden, rapid spectral decay. Recall that a nested series of hierarchical data representations may be achieved through diffusion maps by taking successive powers of the diffusion operator, $\mathbf{P}^t$ (*not* $\mathbf{P}^{(t)}$), or, equivalently, powering its eigenvalues, $\lambda_i^t \in [0, 1)$ for $i = 2, 3, \ldots, N$, which function as coordinates of the spectral embedding (e.g., $x_j \mapsto \{\lambda_i^t \psi_i(x_j)\}_{i \geq 2}$, for all $x_j \in X$). Of particular interest is the contrast between smooth decay to 0 of the diffusion maps spectrum as $t \to \infty$ and the rapid, finite-time eigenvalue and singular value decays of $\mathbf{P}_t$ and $\mathbf{P}^{(t)}$, respectively, pictured in Figure 1.3. The latter characterization may be useful in the identification of hierarchical condensation events in high dimensions, for ex-

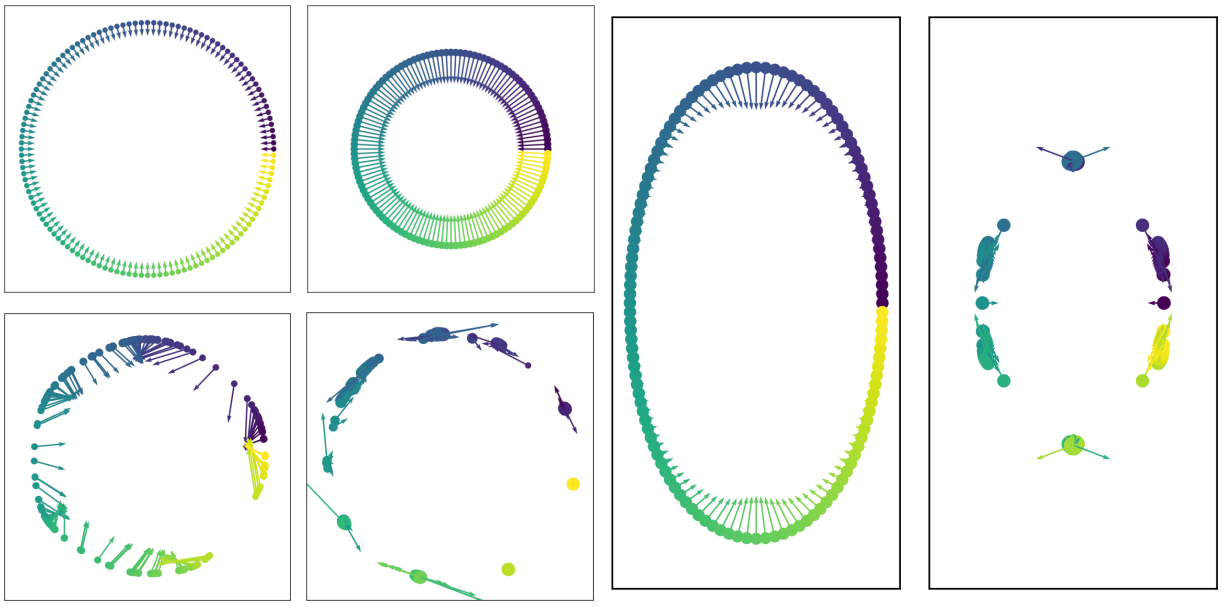

Figure 1.1 Condensation of hyperuniform circle (top-left), uniform circle (bottom-left), and hyperuniform ellipse (right) at early/late iterations (left/right, respectively); point radius corresponds to local density; arrows computed via the infinitesimal generator $\frac{\mathbf{P}_k - I_N}{\epsilon}$ show the gradient field and clearly depict data point acceleration during cluster condensation

ample. We note that while the condensation operator, $P^{(t)}$, is constructed as in Marshall and Hirn (2018), its use in clustering is novel. Spectral characterizations of cluster hierarchy persistence further differentiate the present work. For instance, Figure 1.2 displays many features of interest. Most striking is the correspondence between rapid spectral decay of $\mathbf{P}_t$ and cluster formation, which are depicted just before the moment of condensation. We see three major areas of cluster formation beginning near iteration 15, again near iteration 53, and once again near the last iteration, 100, when the algorithm halts.

1.6.3 Condensation allows multiscale persistence analysis

Since the condensation algorithm naturally allows points to come together via a low pass filter application at each iteration, the time-point in the process at which clusters naturally come together and the length of time for which a cluster persists (without merging) offer notions of cluster metastability. This can be used to derive a partitioning of the dataspace that has mixed levels of granularity. By contrast, most clustering methods are only able

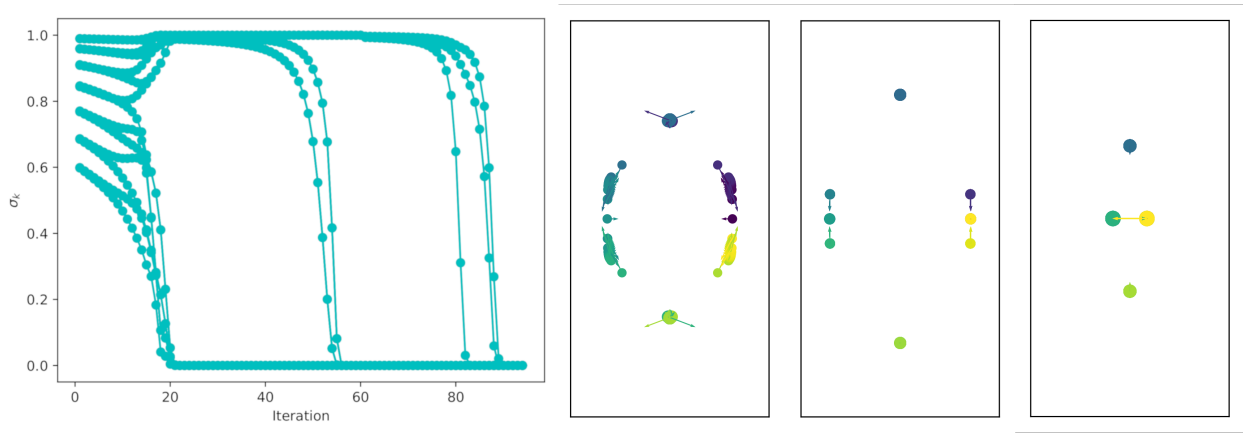

Figure 1.2 Alternative characterization of hyperuniform ellipse cluster formation via top 14 nontrivial singular values of the Markov/diffusion operators, $\{\{\sigma_i(\mathbf{P}_k)\}_{k=0}^{t}\}_{i=2}^{15}$ (far left), and corresponding video stills of hyperuniformly-sampled ellipse condensation at iterations $15, 53$, and 100

to produce results at a particular granularity; for example, k-means tends to favor clusters that roughly divide the data into k partitions of similar sizes. However, different parts of the dataspace may naturally separate at different levels of granularity and this is not visible in other methods. Even hierarchical clustering, due to forced splits and merges, may not reveal the levels of granularity at which data groupings are most distinct. We visualize this persistence information using Sankey diagrams (see Figures 1.4 and 1.5) that show natural groupings of the data. In Section 1.7.1 we use this capability of condensation to suggest a more relevant subtyping of retinal bipolar neurons on the basis of their transcriptomic profile, as compared to previous literature.

1.7 Empirical Results

1.7.1 Single-cell transcriptomics data

A recent study of retinal bipolar neurons using single-cell transcriptomics was performed (Shekhar et al., 2016) to classify cells into coherent subtypes. The study identified 15 cell subtypes by using the method of Blondel et al. (2008), of which 13 were well known and 2 were novel. We use the findings of said study to benchmark the condensation algorithm. From the dataset, we use a randomly selected sample of 20,000 cells with gene expression counts

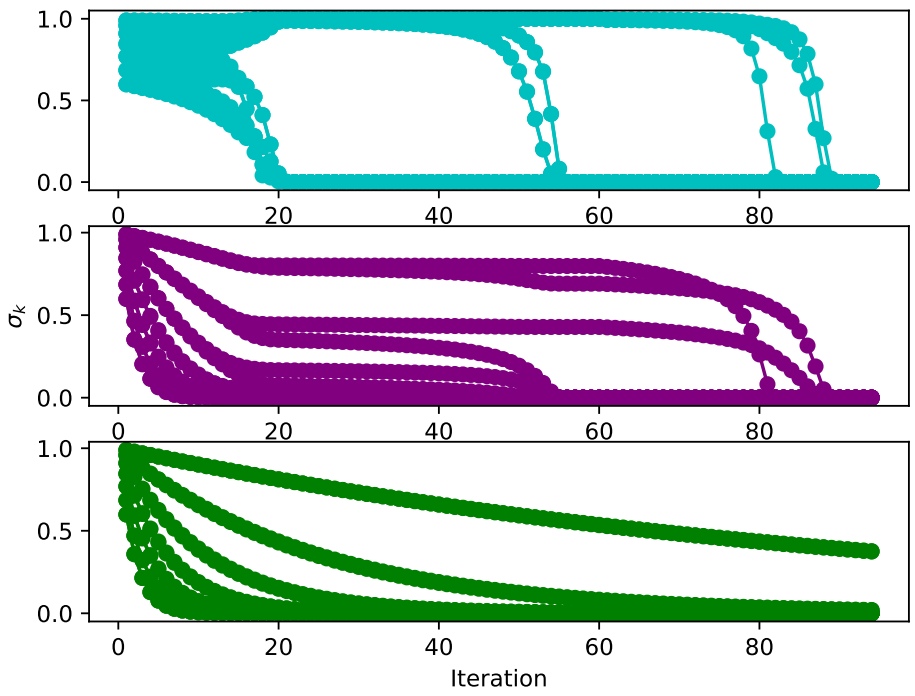

Figure 1.3 Characterization of hyperuniform ellipse cluster formation via top 14 nontrivial singular values of the Markov/diffusion operators, $\{\{\sigma_i(\mathbf{P}_k)\}_{k=0}^{t}\}_{i=2}^{15}$ (top, see also Figure 1.2), $\{\{\sigma_i(\mathbf{P}^{(k)})\}_{k=0}^{t}\}_{i=2}^{15}$ (middle), and $\{\{\sigma_i(\mathbf{P}^{k})\}_{k=0}^{t}\}_{i=2}^{15}$ (bottom, diffusion maps operator)

sequenced to a median depth of 8,200 mapped reads per cell to perform condensation. The condensation ran for 64 iterations until it achieved a metastable state of 12 clusters (close to the 15 reported in Shekhar et al. (2016)). However due to the continuous clustering history offered by condensation, we are able to assess when these metastable clusters first form; see Figure 1.4 for a diagram of iterations 44 to 64. A key advantage of the condensation method is its ability to compute cluster persistence based on the lengths of the clustering tree branches, which we use to reassess the subtyping of retinal bipolar cells performed in Shekhar et al. (2016).

Using community detection methods, Shekhar et al. (2016) found that cluster BC1 (bipolar cone cells, subtype 1) is better described as two clusters, BC1A and BC1B. Shekhar et al. Shekhar et al. (2016) even confirm that morphologically BC1B seems to be a unipolar cluster rather than bipolar. Condensation clustering corroborates this new finding. Indeed, as shown in Figure 1.4, the dark grey BC1 subclusters stay persistently separated until the

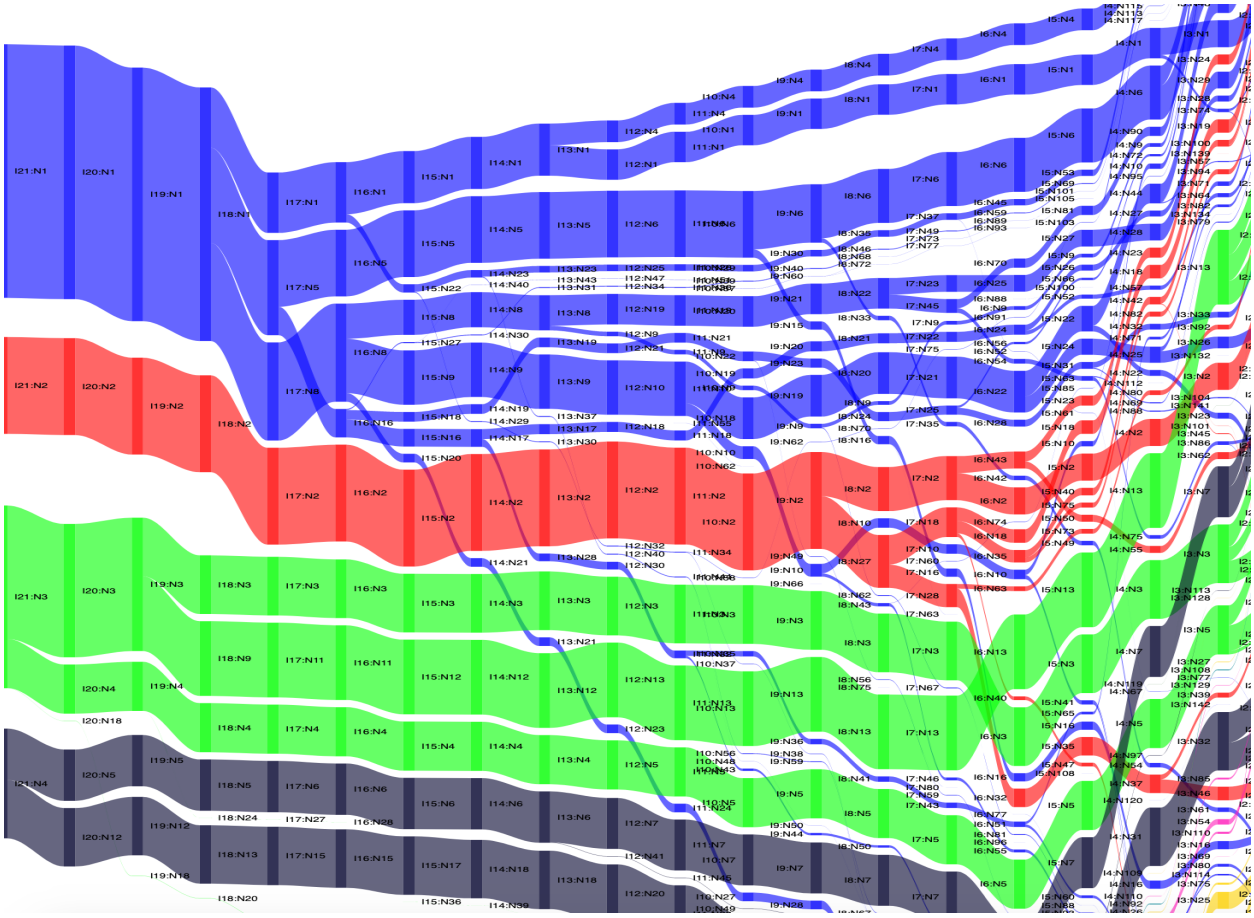

Figure 1.4 Sankey diagram showing results of 20 iterations of diffusion condensation on the scRNA-Seq retinal bipolar dataset; left side representing final clusters and right representing earlier stages of the process; the two dark strands represent BC1B and BC1A sub-populations; red representing BC3A cell type which becomes distinct quite early in the process; light green being BC7, forms from three distinct strands suggesting possible subtle sub-populations

last iteration shown. Therefore, the two subcluster-state is more persistent than the single cluster.

On the other hand, condensation suggests alternative groupings of other clusters not identified by previous papers on retinal bipolar neurons including Shekhar et al. (2016) . Among these, we find that although BC3 has been described in terms of two subcomponents, BC3A and BC3B in biological literature, and in Shekhar et al. (2016), these subclusters merge early (iteration 53) and the transcriptional profiles are not significantly distinct overall, despite certain selective markers such as Erbb4, Nnat being different between the two. Additionally,

we find that our results strongly suggest that BC7 consists of 3 distinct subtypes that persist separately until the last iteration. Previously, the BC7 cell type has been described as a Vstm2b+Casp7+ cone cell that is distinct from other BC types as predicted by Shekhar et al. (2016). Our analysis, however, reveals that there may be multiple sub-populations that are distinct within this cell type designation. While additional experimentation are required to follow up on this finding, condensation provides a way to examine granularities at which data is best organized based on cluster persistence via the whole condensation history.

1.7.2 Neural connectome data

Since the condensation algorithm operates via a series of diffusion operators, which can be regarded as types of adjacency matrices, we sought to understand if the algorithm would apply to coordinate-free spaces. To achieve this we took a datatype that naturally exists as a graph: the neural connectome data of the *Caenorhabditis elegans* brain, a neuropil called the "nerve ring" consisting of 181 neurons. Here an adjacency matrix was created from the contact profiles determined by images along slices of the worm, i.e., neurons that were more frequently in contact with one another were assumed to have a stronger connection and communication with one another. This adjacency matrix was then eigendecomposed to create a coordinate space in order to perform the condensation. The remainder of the algorithm remained as described.

First we sought to test out the robustness of the condensation algorithm by applying it to two complete connectomes of the *Caenorhabditis elegans* brain. Previous comparisons between these connectomes had concluded that they largely share similar structure at the level of cell morphology and synaptic positions (White et al., 1986). We therefore hypothesized that by comparing the output from these similar connectomes we could test the robustness of our algorithm. Specifically, we focused on analyzing the relationship between cell-cell contact profiles for every neuron within the two connectomes (Brittin et al., 2018). Cell-cell contact relationships should define modules with the brain that are bundled together, and we hypothesized that if the algorithm was working as expected, it should extract similar con-

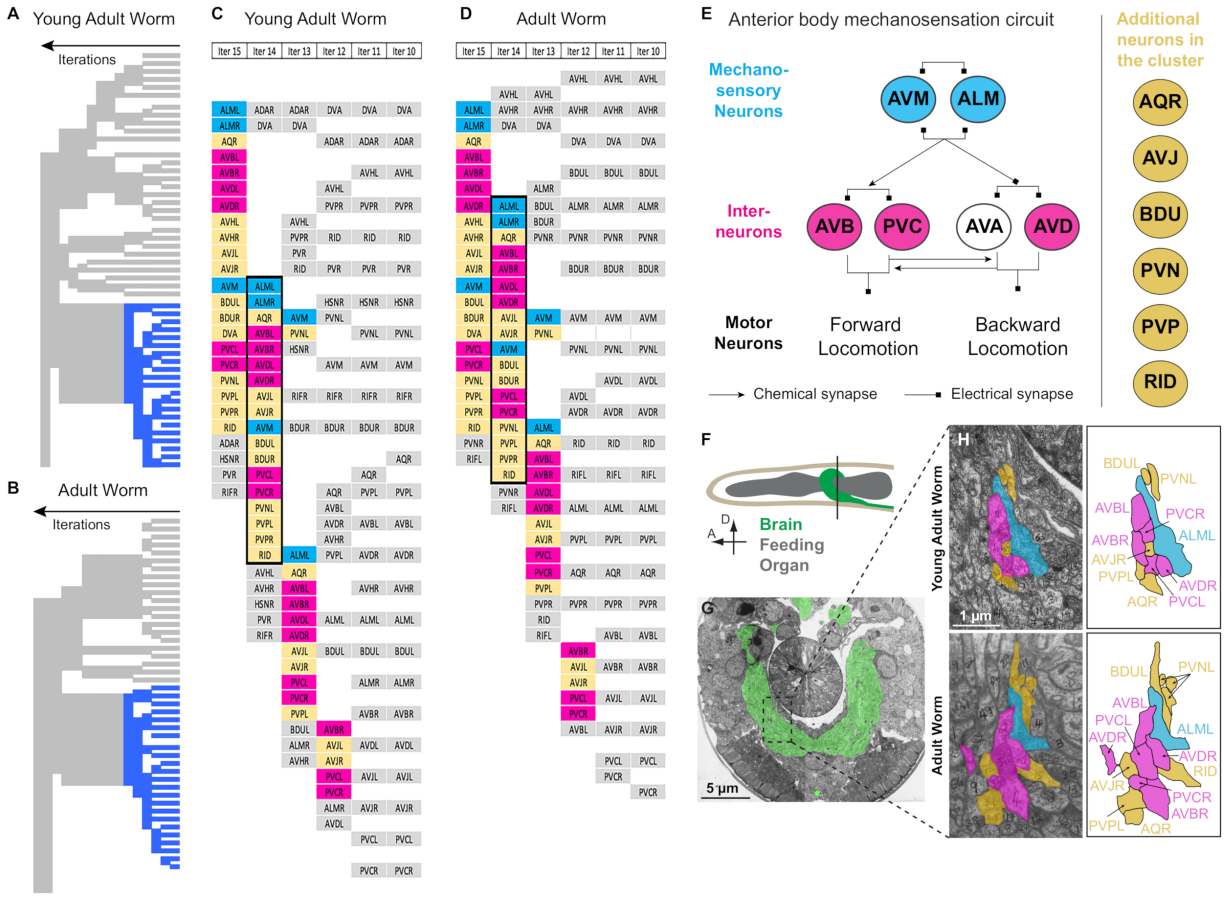

Figure 1.5 A, B) Sankey diagrams of the condensation results for two *C. elegans* connectomes: a young adult (A) and an adult worm (B). C, D) Sankey diagrams of a subset on neurons (blue in A and B) for a young adult (C) and an adult worm (D); letter code corresponds to specific neuron names; cells are pseudocolored based on the function of the specific neuron in the circuit, with blue representing mechanosensory neurons, red representing interneurons, and yellow representing additional neurons unknown to function in this circuit; note similar condense profiles at the level of single cells between both young adult (C) and an adult worm (D) connectomes; cluster outlined in black is the cluster analyzed further in (E); E) Circuit diagram for the anterior body mechanosensation circuit; we color the specific neurons according to function as in (C and D); circuit diagram adapted from Girard et al. (2006); F) Cartoon depicting the head of *C. elegans*; the vertical black line shows where the electron micrograph serial section was collected; G) Serial electron micrograph image; neurons corresponding to the brain of the animal are highlighted green; H) Cropped view of a cross section from the serial electron micrographs corresponding to the anterior body mechanosensation circuit (represented in E); neurons are pseudocolored as in (C-E); note how both the relative positions contact profiles of these neurons are similar between both animals, as predicted by the algorithm

tact profiles among the two connectomes. We observed that our algorithm produced similar condense profiles for the two connectomes (See Figure 1.5), suggesting that our method can be used to robustly analyze connectomics data. We see that the Sankey diagrams preserve much of the structure including an important mechanosensory circuit. To quantify the similarity between condensation clusters generated from the two connectomes, we compute the adjusted Rand index (ARI) at each condensation iteration from 0 to 24 and then take the mean. This yields an ARI $= 0.7$, for $-1 \leq \text{ARI} \leq 1$, where the closer to ARI $= 1$ the better.

A major advantage of the diffusion condensation algorithm 1.1 is that it allows analyses of computational iterations to extract biologically relevant information informing the clustering steps. We hypothesized that these iterative steps could reveal units of circuit architecture underlying the brain. To test this, we examined the clusters for well described circuits, specifically, for the anterior body mechanosensation circuit (Girard et al., 2006). The anterior body mechanosensation circuit contains 2 classes of mechanosensory neurons and 4 classes of command interneurons that contact and connect to each other, and based on their contact profile, should be identified by the algorithm (Chalfie et al., 1985; Wicks and Rankin, 1995). Indeed, iteration 14 (Figure 1.5) identified the circuit in both worms, revealing the predicted relationships between these connecting neurons. Interestingly, iteration 14 also contains neurons of unknown function that, upon closer inspection, are closely associated to the circuit, but have not been implicated in mechanosensory behaviors. Therefore inspection of the condensation algorithm not only extracted the known circuit, but also motivated a new hypothesis regarding the function of unknown neurons associated to the circuit. Together, our analyses demonstrate that the method can be used to compare connectomics data across organisms, to extract biologically relevant units of circuit architecture and even to inform new experiments and discoveries of biological importance. We propose that this method will be broadly useful for systems level analysis of connectomics data.

1.7.3 Algorithm comparison

We compare condensation at two times with Mini Batch K-Means, Agglomerative Clustering with Ward linkage, and Agglomerative Clustering with average linkage. The two condensation times are early and later, where early is half the iterations of later. The computational experiments are conducted on part of the scikit-learn clustering dataset with default, tuned parameter values and datasets. The variance of the center blob in the Gaussian blobs dataset (Figure 1.6, row two) was decreased from 2.5 to 1.5 for separability.

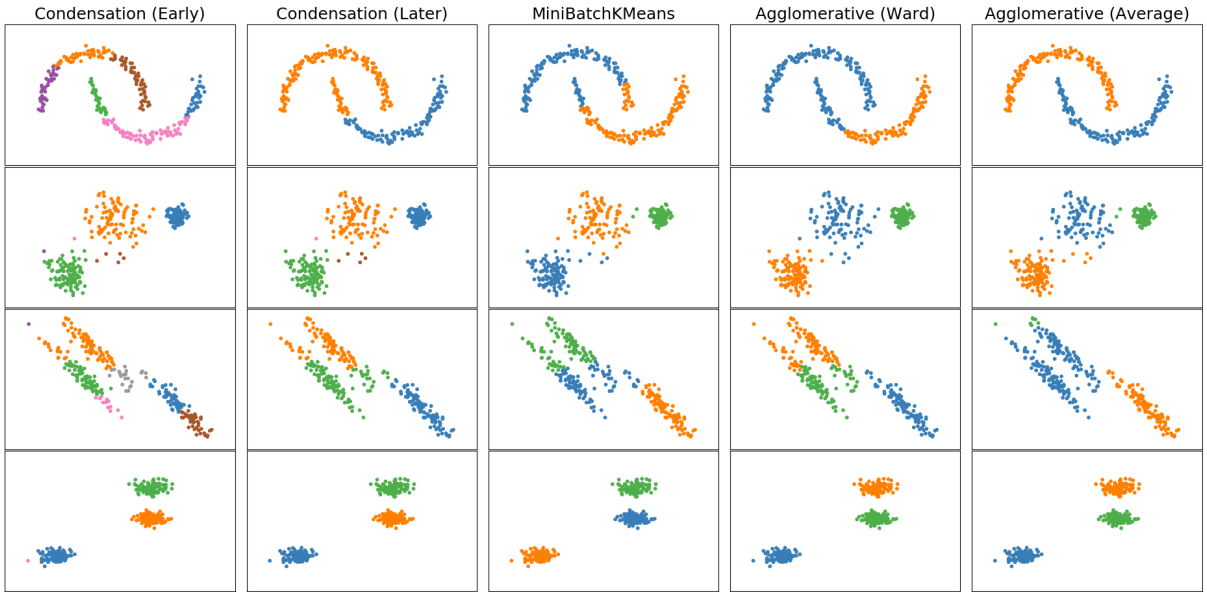

Figure 1.6 Condensation results as compared with Mini Batch K-Means (center), Agglomerative Clustering with Ward linkage (center-right), and Agglomerative Clustering with average linkage (right) on the scikit-learn clustering dataset ($N = 300$); the early condensation snapshot (left) is taken at half the iterations of the later (center-left)

In Figure 1.6 we see the earlier iteration of condensation exhibits finer clustering by curvature than the later. Similarly, row three of Figure 1.6 exhibits coarser clustering in the later condensation labeling. These examples demonstrate the multiscale nature of clusters assigned via condensation. We note that while we employ only the Euclidean metric in these examples, preliminary tests using other metrics yield promising results.

1.8 Conclusion

We presented a multiresolution data abstraction approach based on a time-inhomogeneous diffusion condensation process that gradually coarse grains data features along the intrinsic data geometry. We demonstrated the application of this method to biomedical data analysis, in particular in single cell transcriptomics. Furthermore, the presented diffusion condensation can be seen as a cascade of data-driven lowpass filters that gradually eliminates variations in the data to extract increasingly abstract features. Indeed, under this interpretation, the abstraction provided by the condensation process can be related to common intuitions of features extracted by hidden layers of deep convolutional networks, e.g., in image processing. Such features are commonly considered as increasing in abstraction capabilities together with the depth of the network. However, we note that while convolutional networks typically employ relatively-simple pointwise nonlinearities, here the nonlinearity we employ is the reconstruction of the diffusion geometry based on the coarse grained features along the cascade. Therefore, our cascade both learns a multiresolution data geometry and extracts multiresolution characterizations of groupings based on invariant features at each iteration. Finally, we note the increasing interest in geometric deep learning, which aims to tie together filter training in deep networks with non-Euclidean geometric structures that often exist intrinsically in modern data. While our approach here relies only on lowpass filters, it opens interesting directions in employing trained filters (or even designed diffusion wavelets, as done in diffusion and geometric scattering (Gama et al., 2019; Gao et al., 2019) together with geometric reconstruction for multiscale feature extraction from data.

1.9 Acknowledgements

In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of Michigan State University's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/

publications/rights/rights_link.html to learn how to obtain a License from RightsLink. If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

BIBLIOGRAPHY

Angelo, M., Bendall, S. C., Finck, R., Hale, M. B., Hitzman, C., Borowsky, A. D., Levenson, R. M., Lowe, J. B., Liu, S. D., Zhao, S., et al. (2014). Multiplexed ion beam imaging of human breast tumors. *Nature medicine*, 20(4):436.

Bendall, S. C., Simonds, E. F., Qiu, P., Amir, E.-a. D., Krutzik, P. O., Finck, R., Bruggner, R. V., Melamed, R., Trejo, A., Ornatsky, O. I., Balderas, R. S., Plevritis, S. K., Sachs, K., Pe, D., Tanner, S. D., and Nolan, G. P. (2011). Single-Cell Mass Cytometry of Differential. *Science*, 332(May):687–695.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Brittin, C. A., Cook, S. J., Hall, D. H., Emmons, S. W., and Cohen, N. (2018). Volumetric reconstruction of main caenorhabditis elegans neuropil at two different time points. *bioRxiv*, page 485771.

Brugnone, N., Gonopolskiy, A., Moyle, M. W., Kuchroo, M., van Dijk, D., Moon, K. R., Colon-Ramos, D., Wolf, G., Hirn, M. J., and Krishnaswamy, S. (2019). Coarse graining of data via inhomogeneous diffusion condensation. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2624–2633. IEEE.

Chalfie, M., Sulston, J. E., White, J. G., Southgate, E., Thomson, J. N., and Brenner, S. (1985). The neural circuit for touch sensitivity in caenorhabditis elegans. *Journal of Neuroscience*, 5(4):956–964.

Coifman, R. R. and Lafon, S. (2006). Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30.

Dasgupta, A., Hopcroft, J., Kannan, R., and Mitra, P. (2006). Spectral clustering by recursive partitioning. In *European Symposium on Algorithms*, pages 256–267. Springer.

Galluccio, L., Michel, O., Comon, P., Kliger, M., and Hero, A. O. (2013). Clustering with a new distance measure based on a dual-rooted tree. *Information Sciences*, 251:96–113.

Gama, F., Ribeiro, A., and Bruna, J. (2019). Diffusion scattering transforms on graphs. In *International Conference on Learning Representations*. arXiv:1806.08829.

Gao, F., Wolf, G., and Hirn, M. (2019). Geometric scattering for graph data analysis. To appear in the *Proceedings of the 36th International Conference on Machine Learning*, arXiv:1810.03068.

Giesen, C., Wang, H. A., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., Schüffler,

P. J., Grolimund, D., Buhmann, J. M., Brandt, S., et al. (2014). Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature methods*, 11(4):417.

Girard, L. R., Fiedler, T. J., Harris, T. W., Carvalho, F., Antoshechkin, I., Han, M., Sternberg, P. W., Stein, L. D., and Chalfie, M. (2006). Wormbook: the online review of caenorhabditis elegans biology. *Nucleic acids research*, 35(suppl_1):D472–D475.

Izenman, A. J. (2012). Introduction to manifold learning. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(5):439–446.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., and Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 24(7):881–892.

Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., and Teichmann, S. A. (2015). The technology and biology of single-cell rna sequencing. *Molecular cell*, 58(4):610–620.

Kwitt, R., Huber, S., Niethammer, M., Lin, W., and Bauer, U. (2015). Statistical topological data analysis-a kernel perspective. In *Advances in neural information processing systems*, pages 3070–3078.

Levine, J. H., Simonds, E. F., Bendall, S. C., Davis, K. L., El-ad, D. A., Tadmor, M. D., Litvin, O., Fienberg, H. G., Jager, A., Zunder, E. R., et al. (2015). Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197.

Lindenbaum, O., Stanley, J., Wolf, G., and Krishnaswamy, S. (2018). Geometry based data generation. In *Advances in Neural Information Processing Systems*, pages 1400–1411.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Marshall, N. and Hirn, M. J. (2018). Time-coupled diffusion maps. *Applied and Computational Harmonic Analysis*, 45(3):709–728.

Moon, K. R., Stanley, J., Burkhardt, D., van Dijk, D., Wolf, G., and Krishnaswamy, S. (2018). Manifold learning-based methods for analyzing single-cell rna-sequencing data. *Current Opinion in Systems Biology*, 7:36–46.

Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60.

Nadler, B., Lafon, S., Kevrekidis, I., and Coifman, R. R. (2005). Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. In Weiss, Y., Schölkopf, P. B., and Platt, J. C., editors, *Advances in Neural Information Processing Systems 18*, pages 955–962. MIT Press.

Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In Dietterich, T. G., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, pages 849–856. MIT Press.

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14(9):618.

Shekhar, K., Lapan, S. W., Whitney, I. E., Tran, N. M., Macosko, E. Z., Kowalczyk, M., Adiconis, X., Levin, J. Z., Nemesh, J., Goldman, M., McCarroll, S. A., Cepko, C. L., Regev, A., and Sanes, J. R. (2016). Comprehensive Classification of Retinal Bipolar Neurons by Single-Cell Transcriptomics. *Cell*, 166(5):1308–1323.e30.

Van Der Maaten, L., Postma, E., and Van den Herik, J. (2009). Dimensionality reduction: a comparative. *J Mach Learn Res*, 10(66-71):13.

Van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A., Burdziak, C., Moon, K. R., Chaffer, C. L., Pattabiraman, D., et al. (2018). Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3):716–729.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

Wasserman, L. (2018). Topological data analysis. *Annual Review of Statistics and Its Application*, 5:501–532.

White, J. G., Southgate, E., Thomson, J. N., and Brenner, S. (1986). The structure of the nervous system of the nematode caenorhabditis elegans. *Philos Trans R Soc Lond B Biol Sci*, 314(1165):1–340.

Wicks, S. R. and Rankin, C. H. (1995). Integration of mechanosensory stimuli in caenorhabditis elegans. *Journal of Neuroscience*, 15(3):2434–2444.

CHAPTER 2

SELF-SIMILAR GRAPH SIGNALS & SYSTEM CLASSIFICATION

2.1 Abstract

Complex systems are characterized by emergent patterns. These patterns often exhibit self-similarity that arises from interactions across various scales. Self-similar patterns can be modeled as random fields called *textures*. In this paper we generalize a class of self-similar random fields that are indexed by time to random fields that are indexed by the vertices of a graph. We take the perspective of graph signal processing and draw upon recent developments in the theory of self-similar random fields on manifolds. by invoking the duality between graphs and graph signals, we then empirically demonstrate how the proposed model can support the modeling of complex social-ecological systems phenomena through data-driven classification of system models. We conclude by suggesting multiple use cases.

2.2 Introduction

From cauliflower to clouds to coastlines, self-similar patterns surround us. Mandelbrot (1967) popularized the study of *self-similarity* through enchanting patterns called fractals. Statistical self-similarity refers to properties of distributions the consistently reappear at various scales. In natural systems, mountain elevations, river water levels, and forest tree densities are said to be spatially self-similar. Self-similarity appears in tandem with *long-range dependence*, which refers to (second-order stationary) signals with correlations that persist across great distances (Pipiras and Taqqu, 2017). Self-similarity and long-range dependence are characteristic of complex systems and notably emerge through agent-based simulations. Such patterns are observed, for instance, in the size distribution of forest fires and the spatial distribution of communities in Schelling's segregation model (Batten, 2001). In economic systems, Pareto studied (Amoroso, 1938; Schoenberg and Patel, 2012) self-similar patterns in the distribution of income and wealth.

According to Grimm et al. (2005, p.1):

> ...patterns are defining characteristics of a system and often, therefore, indicators of essential underlying processes and structures. Patterns contain information on the internal organization of a system, but in a coded form.

A major task for systems modelers, then, is to account for these patterns. We explore this pattern-oriented approach to systems from the perspective of random processes.

2.3 Background

2.3.1 Random Processes & Fields

A *random process* can be defined as a collection of real-valued random variables $\{X(t)\}_{t\in\mathcal{I}}$ that are indexed by a set, $\mathcal{I}$, which contains an origin at $t = 0$ as well as a linearly ordered structure like $\mathbb{R}$ and $\mathbb{Z}$. Random processes are often used to model time-evolving phenomena such as prices. A *Euclidean random field* is an extension of random processes to index sets that have not only an origin, but also higher-dimensional vector space structure and larger symmetry groups. For instance, $\mathbb{R}^d$ is a d-dimensional vector space whose symmetries include rotation about the origin. The algebraic structure associated with an index set is often used to endow a random field with certain properties. A canonical example is the Brownian motion on $\mathbb{R}$, which has an ordered structure that is defined with respect to the origin. In general, $\mathcal{I}$ could be any set with or without additional structure, for instance, the vertex set of a graph.

A random field is called *strictly stationary* if $X(s) \overset{d}{=} X(t)$ for every $s, t \in \mathcal{I}$, where $\overset{d}{=}$ denotes equality of all finite dimensional distributions. In applications, stationarity functions as a modeling assumption that can facilitate parameter estimation and theoretical analysis. Strict stationarity, however, can be too restrictive for modeling purposes, which motivates consideration of *second-order stationary* fields for which only the mean and covariance are assumed to be stationary (Pipiras and Taqqu, 2017), i.e., for every $s, t \in \mathcal{I}$ with $\mathcal{I}$ a vector

space and $a \in \mathbb{R}$,

$$\mathbb{E}\left[X\left(s\right)\right] = a,$$

and,

$$\begin{aligned}\mathrm{Cov}\left(X\left(s\right), X\left(t\right)\right) &:= \mathbb{E}\left[\left(X\left(s\right) - \mathbb{E}\left[X\left(s\right)\right]\right)\left(X\left(t\right) - \mathbb{E}\left[X\left(t\right)\right]\right)\right] \\ &= \mathbb{E}\left[\left(X\left(0\right) - \mathbb{E}\left[X\left(0\right)\right]\right)\left(X\left(t-s\right) - \mathbb{E}\left[X\left(t-s\right)\right]\right)\right] \\ &= \mathrm{Cov}\left(X\left(0\right), X\left(t-s\right)\right).\end{aligned}$$

The field is called *centered* if $a = 0$. Vector space structure also enters into random fields via *isotropy* (Ayache, 2018), which says $X\left(\mathbf{Q}t\right) \overset{d}{=} X\left(t\right)$ for each $t \in \mathbb{R}^d$ and every orthogonal operator $\mathbf{Q}$ on $\mathbb{R}^d$. Following Gelbaum (2014) it is then tempting to define a (centered) second-order stationary random field as one for which,

$$\mathrm{Cov}\left(X\left(\iota\left(s\right)\right), X\left(\iota\left(t\right)\right)\right) = \mathrm{Cov}\left(X\left(s\right), X\left(t\right)\right),$$

for every $s, t \in \mathcal{I}$ and every isometry ι on $\mathcal{I}$. One could even neglect the metric structure and define a second-order G-stationary random field as one for which,

$$\mathrm{Cov}\left(X\left(g * s\right), X\left(g * t\right)\right) = \mathrm{Cov}\left(X\left(s\right), X\left(t\right)\right),$$

for every $s, t \in \mathcal{I}$ and $g \in G$, with G a subgroup of $\mathrm{Aut}\left(\mathcal{I}\right)$, the group of automorphisms on $\mathcal{I}$, and $g*$ denoting the group action. We will say a random field $\left\{X\left(t\right)\right\}_{t\in\mathcal{I}}$ has *stationary increments* if,

$$\left\{X\left(g * s\right) - X\left(g * t\right)\right\} \overset{d}{=} \left\{X\left(s\right) - X\left(t\right)\right\},$$

for $s, t \in \mathcal{I}$ and $g \in \mathrm{Aut}\left(\mathcal{I}\right)$. We will similarly say a field is G-stationary if,

$$\left\{X\left(g * t\right)\right\} \overset{d}{=} \left\{X\left(t\right)\right\}.$$

for all $g \in G$ for G as above. If $\mathcal{I} = \mathbb{R}^d$, for example, and G consists of all orthogonal transformations, then the field is isotropic. If $\mathcal{I}$ is the integer lattice modulo periodic translations

and G consists of all translations, then the field is stationary in the traditional sense. When the field is centered and characterized by second order moments, our notion of stationarity corresponds to graph weak stationarity as defined in Marques et al. (2017).

2.3.1.1 Statistical Self-Similarity

We define a *statistically self-similar* Euclidean random field as one for which given a scalar, $c \in \mathbb{R}_{\geq 0}$,

$$X(ct) \stackrel{d}{=} c^H X(t), \tag{2.1}$$

where H is the self-similarity, or Hurst, parameter (Ayache, 2018; Pipiras and Taqqu, 2017; Gelbaum, 2014). It is notable that this definition implies that there are no stationary statistically self-similar Euclidean random fields since $c^H X(t) \to \infty$ as $c \to \infty$. Furthermore, $X(0) = 0$.

2.3.2 Extension to graphs

We would like to use equation 2.1 to develop models for classes of complex social-ecological systems phenomena. The immediate challenge encountered when attempting to generalize this definition to graphs and manifolds, however, is how to interpret the expression $X(ct)$. Let V denote the vertex set of an undirected graph, $\Gamma = (V, E)$, with E the edge set. Changing notation emphasizes the issue, as $X(cv)$ for $v \in V$ is nonsensical. This familiar challenge arises when attempting to extend wavelets to graphs using the spatial/temporal notions of dilation and translation (Hammond et al., 2011). In the context of manifolds, Gelbaum (2014) circumvents this issue by utilizing the heat kernel to extend Euclidean fractional Brownian fields to complete Riemannian manifolds and showing that distributional scaling can be understood with respect to the underlying metric. We will draw on this approach and introduce a weaker form of self-similarity based upon wavelet theory.

2.3.2.1 Wavelets

Wavelets are spatially localized waveforms with concentrated frequency support. Their localized nature enables measurement and representation of fine details in images and audio

(Mallat, 1999) as well as signal data supported on graphs and manifolds (Hammond et al., 2011; Gao et al., 2019; Perlmutter et al., 2020). The study of wavelets has yielded insights into the empirical successes of deep convolutional neural networks (both Euclidean and graph) via the scattering paradigm (Bruna and Mallat, 2013; Mallat, 2016; Perlmutter et al., 2019), and parallel work has discovered interpretable models of non-Gaussian random fields (Morel et al., 2022; He and Hirn, 2021; Bruna and Mallat, 2019; Zhang and Mallat, 2021).

Formally, we can define a wavelet as a zero-mean function,

$$\begin{aligned} \psi : \mathcal{I} &\to \mathbb{C} \\ t &\mapsto \psi(t), \end{aligned}$$

with $\int |\psi(t)|\, dt = 1$ for some nice measure dt. For our purposes, it will be convenient to consider a wavelet as a convolution operator on vectors, f, and write this as an inner product,

$$\begin{aligned} \langle f, \psi_{v,j} \rangle_{L^2(\mathcal{I})} &= \int f(t)\, \psi^*_{v,j}(t)\, dt \\ &= \mathbf{\Psi}_j f(v), \quad v \in \mathcal{I}, \end{aligned}$$

where $\psi_{v,j}$ denotes $\mathbf{\Psi}^*_j \delta_v$, and can be interpreted as a wavelet with spatial support concentrated in a neighborhood of $v \in \mathcal{I}$ at a scale 2^j for $j \leq J$ the maximal scale. For all wavelets we will consider we have,

$$\psi_{v,j}(t) = \psi_{g*v,j}(g * t), \quad g \in G, \tag{2.2}$$

for G a subgroup of the group of automorphisms on $\mathcal{I}$. This corresponds to translation and/or rotational invariance about the origin in the Euclidean case.

2.3.2.2 Graph Wavelets

Given the adjacency matrix, $\mathbf{A}$, of a graph, we define the averaging operator,

$$\mathbf{P} = \frac{1}{2}\left(\mathbf{I} + \mathbf{D}^{-1}\mathbf{A}\right), \tag{2.3}$$

where $\mathbf{D}$ is the diagonal matrix of degrees and $\mathbf{I}$ is the identity. Given a signal x on a graph, $\mathbf{P}x$ averages x in 1-hop neighborhoods of each vertex. We next define the wavelet operator

at a scale 2^j as,

$$\mathbf{\Psi}_j = \mathbf{P}^{2^{j-1}} - \mathbf{P}^{2^j}, \quad j \leq J. \tag{2.4}$$

This difference between averaging operators applied to a signal yields the details lost to averaging between scales 2^{j-1} and 2^j. Our wavelets are related to those utilized in graph scattering (Gama et al., 2019; Gao et al., 2019) and can be interpreted within the framework of Perlmutter et al. (2019). We note that our wavelets as well as those considered in the above works satisfy equation 2.2. This is easily seen since adjacency is preserved under automorphisms.

Wavelets are powerful tools for processing random fields as we show in the following result.

Theorem 1. *Let X be a random field on $\mathcal{I}$ with G-stationary increments, and let dt be the density of a measure on $\mathcal{I}$ that is invariant to the action of G. Then $\mathbf{\Psi}X$ is G-stationary.*

Proof. Let $g \in G$ be arbitrary. Since a wavelet has zero mean,

$$\begin{aligned}
\mathbf{\Psi}_j X(v) &= \int X(t)\, \psi^*_{v,j}(t)\, dt \\
&= \int X(g*t)\, \psi^*_{v,j}(g*t)\, dt \\
&= \int X(g*t)\, \psi^*_{g^{-1}*v,j}(t)\, dt, \quad \text{(by 2.2)} \\
&= \int X(g*t)\, \psi^*_{g^{-1}*v,j}(t)\, dt - X(g*v) \int \psi^*_{g^{-1}*v,j}(t)\, dt \\
&= \int [X(g*t) - X(g*v)]\, \psi^*_{g^{-1}*v,j}(t)\, dt \\
&\overset{d}{=} \int [X(t) - X(v)]\, \psi^*_{g^{-1}*v,j}(t)\, dt \\
&= \int X(t)\, \psi^*_{g^{-1}*v,j}(t)\, dt \\
&= \mathbf{\Psi}_j X\left(g^{-1}*v\right).
\end{aligned}$$

□

This is meaningful in the context of manifolds, for instance, when dt corresponds to the Riemannian volume element, G corresponds to a group of isometries, and the wavelets are

constructed as in Perlmutter et al. (2020). This perspective may, in turn, be interpreted to subsume the Euclidean cases where, for instance, wavelet filtering of the Brownian motion on $\mathbb{R}$ produces stationary random processes due to the stationarity of increments of Brownian motion.

We note that result 1 is most interesting when the automorphism group on $\mathcal{I}$ is nontrivial. However, 1 is also interesting when considered across *all graphs*, as it indicates that there are asymptotically few nonstationary processes with stationary increments since the proportion of graphs on n vertices with nontrivial automorphism groups tends to zero as n tends to infinity (Godsil and Royle, 2001).

2.3.2.3 Graph wide sense self-similarity

As noted by Morel et al. (2022) in the Euclidean setting, the scaling definition 2.1 for self-similarity can be too restrictive for modeling and is impossible to numerically verify. This motivates a weaker definition based upon wavelets.

Definition 2.3.1 (Wide-Sense Self-Similarity)**.** *A graph random field will be called wide sense self-similar up to a maximum scale 2^J if there exist coefficients c_1, c_2, ζ_1, and ζ_2 such that for all $j \leq J$,*

$$\mathbb{E}\left[|\mathbf{\Psi}_j X(v)|\right] = c_1 2^{j\zeta_1}, \tag{2.5}$$

and,

$$\mathbb{E}\left[|\mathbf{\Psi}_j X(v)|^2\right] = c_2 2^{j\zeta_2}. \tag{2.6}$$

2.3.3 Multiscale statistics

Complex systems (e.g., economic, climate, and social systems) exhibit cross-scale dependencies that generate emergent phenomena. Wavelets isolate fluctuations at different scales. To study interactions between scales it is tempting to consider Gram statistics between these scales, i.e. $\langle \mathbf{\Psi}_j X, \mathbf{\Psi}_k X \rangle$ for $j, k \leq J$. However, interpreting our wavelets as polynomials on the spectrum of the normalized graph Laplacian (e.g., Perlmutter et al. (2019)), we see that

$\mathbf{\Psi}_j$ and $\mathbf{\Psi}_k$ share very little spectral overlap, and hence $\langle \mathbf{\Psi}_j X, \mathbf{\Psi}_k X \rangle \approx 0$ for $j \neq k$. We can, however, apply the ReLU operator to wavelet filtered signals to push higher frequency components into a lower range so as to capture cross-scale statistics. Extending the Euclidean model of He and Hirn (2021), we introduce our main statistical model for self-similar graph signals.

Definition 2.3.2 (Nonlinear Graph Wavelet Model)**.** *Given a statistically self-similar random field, X, we represent it as,*

$$X \mapsto \left\{ \left\langle \sigma\left(\gamma \mathbf{\Psi}_j X\right), \sigma\left(\gamma \mathbf{\Psi}_k X\right) \right\rangle_{L^2(\mathcal{I})} : 1 \leq j \leq k \leq J,\ \gamma \in \{-1, 1\} \right\}. \tag{2.7}$$

Note that the ReLU (rectified linear unit) operator, defined as the pointwise nonlinearity $\sigma(f) = \max\{f, 0\}$, fulfills,

$$|f| = \sigma(f) + \sigma(-f), \tag{2.8}$$

Hence, the ReLU graph wavelet statistics subsume the modulus wavelet statistics.

2.4 Methods & Materials

We employ the statistical model 2.7 to study classification of systems based upon structural similarity. System structure indicates function (Meadows, 2008). This structure can be encoded in graphical models that are constructed via interview, collaborative modeling processes, and even via natural language processing of scholarly texts. As we discuss in chapter 3, there are certain estimation tasks that can be improved by aggregating knowledge over expert subpopulations. In such applications, subpopulation identification amounts to an unsupervised process. In this section, we study self-similar graph signal statistics in a supervised setting to understand potentials for generalization to unsupervised tasks.

2.4.1 From Modeling to Classifying Systems

Just as the function of a system is encoded in its structure, patterns of cognition about system behavior can be represented in models called fuzzy cognitive maps (FCMs) (Gray et al., 2013) and causal loop diagrams (CLDs) (Voinov et al., 2018). An example FCM is depicted in Figure 2.1.

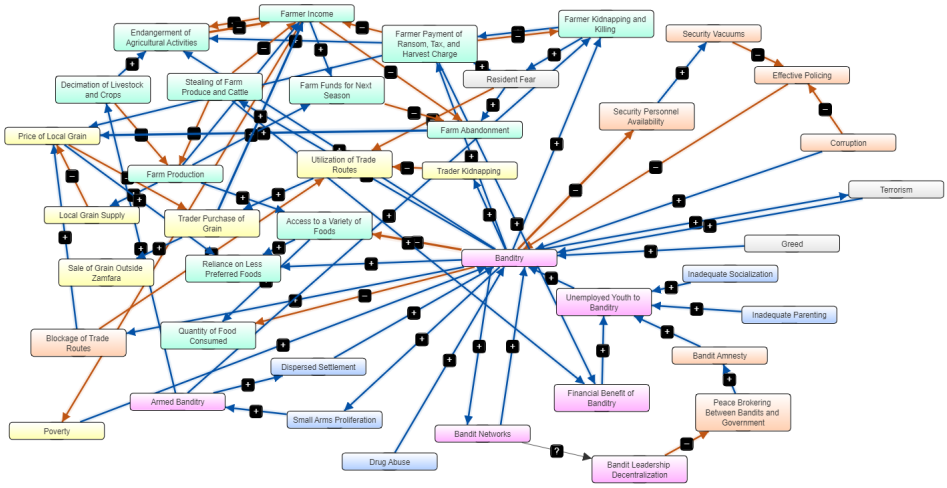

Figure 2.1 Manually drawn FCM describing food commodity prices and conflict dynamics in Zamfara, Nigeria; based upon Mmahi and James (2023)

These models represent causal knowledge about systems as graphs. Using the statistical model in 2.7, we can place these models of potentially different sizes and providence into a common framework for comparison.

2.4.1.1 Random Graphs

We generate graphical surrogates of FCMs and CLDs using random graph models. This technique is used, for instance, in Aminpour et al. (2020). In our case, surrogates allow for assessment with respect to ground truths for which we may vary parameters and study performance in a controlled manner. We use random graphs drawn from the Erdös-Renyi, the Barabasi-Albert, and the Watts-Strogatz model families (pictured in Figure 2.2). Each of these families reproduces aspects of the FCM/CLD construction process. For instance, the Barabasi-Albert process mimics the construction of FCMs by individual experts and small groups of like-minded stakeholders, where new nodes are sequentially connected to more central concepts with higher probability. By contrast, the Erdös-Renyi family follows a construction process that mimics construction by individuals possessing more general knowledge as well as larger, more diverse audiences engaged in collaborative modeling where there is less emphasis placed upon a small handful of central concepts and more on the connectivity of

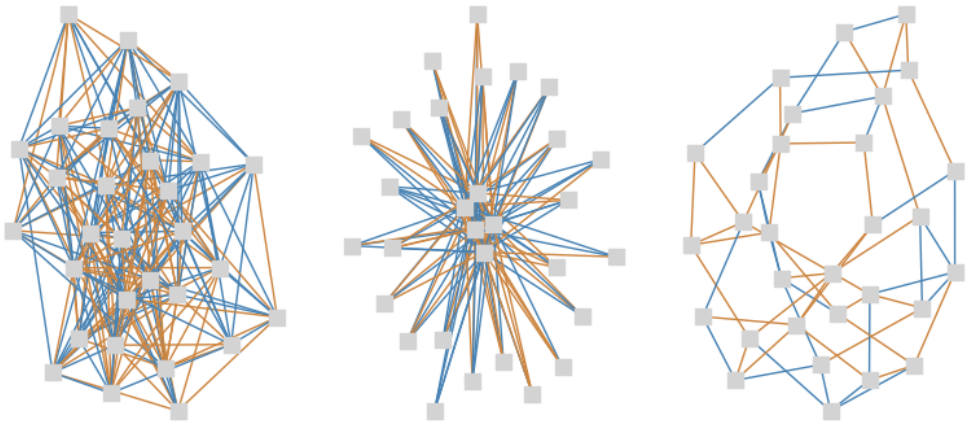

Figure 2.2 Random graph models from the Erdös-Renyi (left), Barabasi-Albert (center), and Watts-Strogatz (right) families

the entire system. Finally, the Watts-Strogatz small-world model mimics a reflective process wherein a base model is presented and participants suggest corrections to the proposal.

2.4.1.2 Experimental Setup

Graphs are sampled from random graph families. We vary the number of nodes, sample sizes, and parameters of the random graph models to understand performance of the method. Except where noted in the small-sample experiments, we draw 40 samples from each graph family. Where relevant and possible, we numerically tune the densities of these graph families so they are as close as possible on average; this prevents this simple summary statistic from being used as a proxy. Note the density of a graph is defined as the number of edges divided by the total number of possible edges. Graph families on the same number of nodes with approximately equal mean densities therefore have the same number of edges on average as well. So, our problem really becomes one of studying topology, i.e., system structure.

Each graph is represented via the covariance structure of a set of qualitatively self-similar signals defined on the nodes of the graph. As in Gao et al. (2019), for each graph we compute the node-wise eccentricity and clustering coefficient signals. We then form the graph wavelets described in equation 2.4 and compute the ReLU graph wavelet Gram statistics for

both signals as in equation 2.7. We also sum these statistics (interpreted as a 1-1 convolution) and compute the ReLU graph wavelet Gram statistics. This is essentially a 2-layer graph convolutional neural network in which nothing is learned. These representations are concatenated to form coordinates to which support-vector machines (SVM) is applied. We use nested 5-fold cross-validation to tune hyperparamters and assess prediction uncertainty. Hyperparameters are selected via grid search.

2.5 Results

2.5.1 Single-Class, Varied Parameters

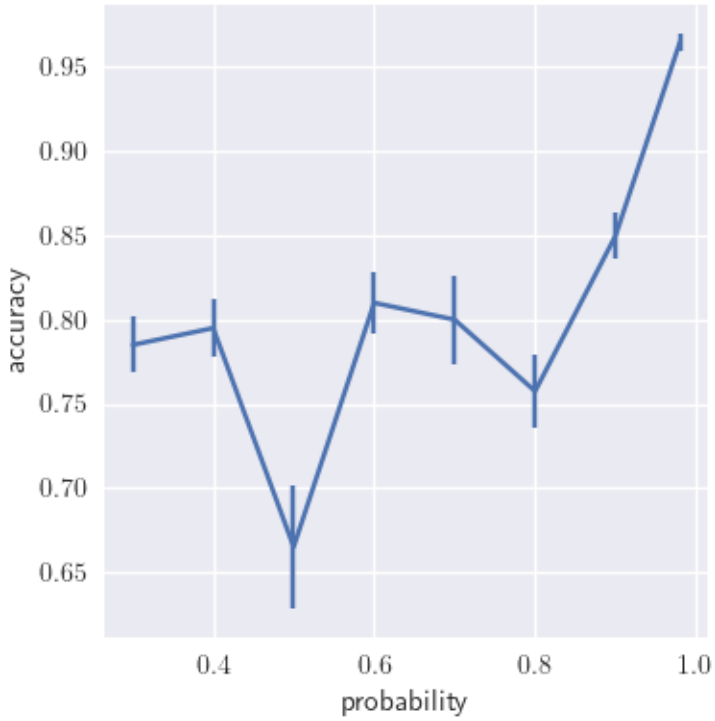

Figure 2.3 Classification accuracy for Erdös-Renyi graphs with edge probabilities p and $p + 0.01$. The lower of the probabilities is reported as the x coordinate.

Figure 2.3 shows the results of classification of Erdös-Renyi graphs for pairs of graph samples ($n = 40$ each) where the edge probability parameter of the first sample is p, and that of the second sample is $p + 0.01$. p is varied over the range $\{.3, .4, .5, .6, .7, .8, .9, .98\}$. The method is able to distinguish most successfully outside of a neighborhood of $p = 0.5$

and with highest accuracy for the largest values of p.

Experiment	Accuracy
Low-density	0.82 ± 0.020
Medium-density	0.85 ± 0.018
High-density	0.99 ± 0.004

Table 2.1 Barabasi-Albert results

To assess multiple parameters from the same class, we apply the method to multiple samples of Barabasi-Albert graphs with similar densities in a sequence of three experiments enumerated by mean density. Parameters for these three experiments appear in Appendix 2.7. Table 2.5.1 displays these results. Accuracy is observed to increase with density.

2.5.2 Single-Class, Varied Size

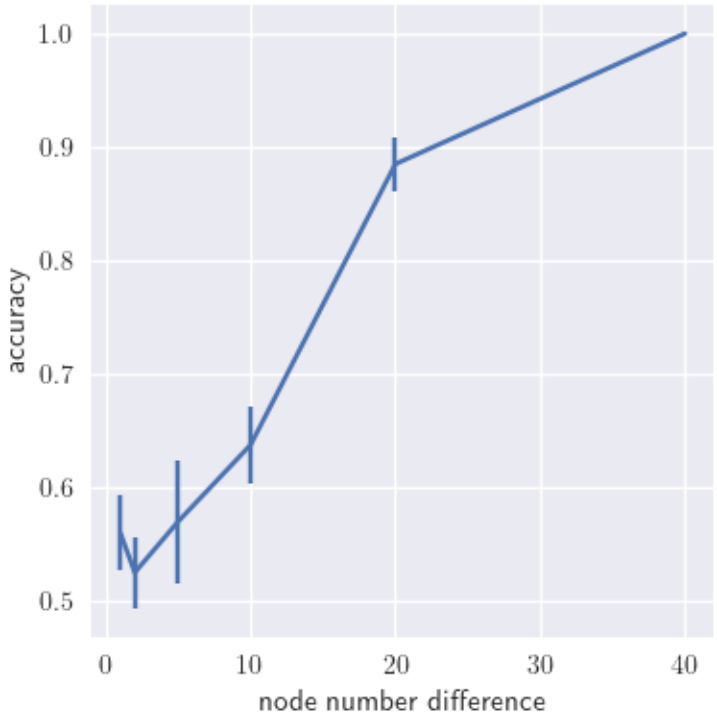

Figure 2.4 Classification accuracy on pairs of samples from Erdös-Renyi graphs whose number of nodes differs by the value along the x-axis (i.e., zero corresponds to identical families)

To understand method performance at distinguishing graph size, we draw samples from

two classes of Erdös-Renyi graphs with the same edge probability parameter, p, and vary the number of nodes. For each experiment, samples are drawn from one family with node count fixed at 100, while another sample is drawn from a family with a number of nodes equal to 60, 80, 90, 95, 98, and 99. This corresponds to six pairs of samples in total. Figure 2.4 shows that classification accuracy decreases as the graph families become more similar. There appears to be a qualitative increase in the variance of the predictive accuracy as well.

2.5.3 Multiclass Classification

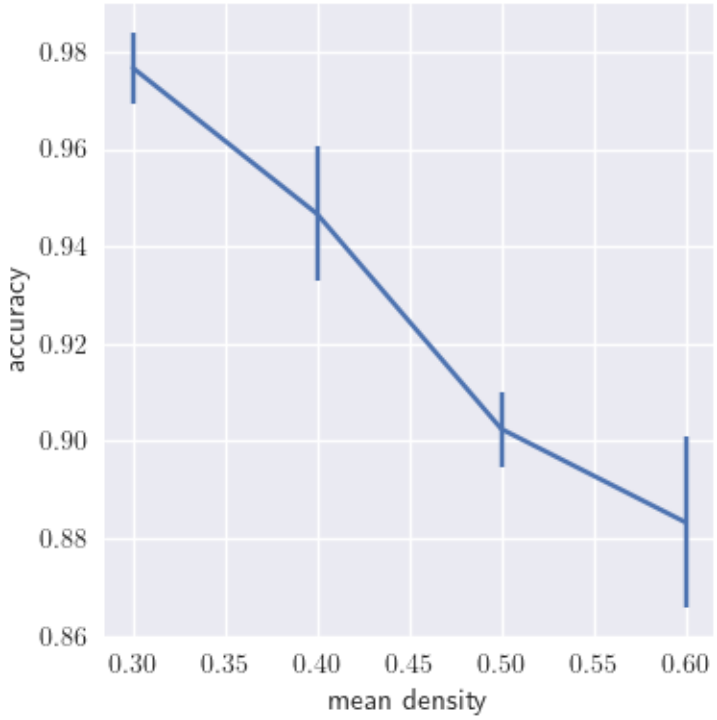

Figure 2.5 Multiclass classification accuracy as a function of mean density of the graph families

To assess performance on a presumably more complex task, we introduce a multiclass classification problem. For each of four experiments we draw six sets of samples from:

1. Erdös-Renyi with 100 nodes
2. Erdös-Renyi with 50 nodes
3. Barabasi-Albert with 100 nodes

4. Barabasi-Albert with 50 nodes
5. Watts-Strogatz with 100 nodes
6. Watts-Strogatz with 50 nodes.

These families are each tuned to have approximately the same density on average with the exception of the highest density experiment for which the Barabasi-Albert model achieved only a maximum mean density of 0.5. The other two graph family densities coincided. Figure 2.5 indicates a drop in performance as density increases. However, the worst case mean performance is conservatively lower-bounded by 86%.

2.5.4 Small Sample Sizes

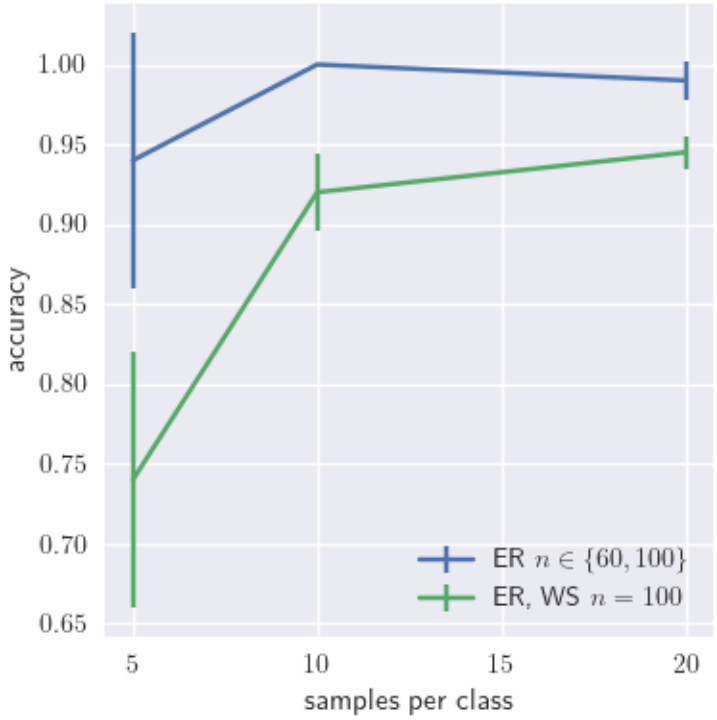

Figure 2.6 Small-sample classification accuracy of Erdös-Renyi and Watts-Strogatz

Finally, we study the effect of small sample sizes through a sequence of pairs of experiments. For sample sizes 20, 10, and 5, we compare classification of Erdös-Renyi graphs with different numbers of nodes (60 and 100) and the same edge probability, $p = 0.5$. We also compare classification of Erdös-Renyi and Watts-Strogatz families on 100 nodes at these

sample sizes. As depicted in Figure 2.6, smaller sample sizes result in a significant drop in accuracy only at the most extreme end.

2.6 Discussion

The ReLU graph wavelet statistics are able to differentiate families of graphs remarkably well across a variety of parameter regimes, in multiclass tasks, and crucially for development applications, in the low-data regime. The statistical model exhibits notably fine resolution when distinguishing Erdos-Renyi graphs where edge probability differs by only 0.01. We see the same capabilities in the classification of Barabasi-Albert models with immediately adjacent parameters.

The drop in performance seen while reducing the difference in number of nodes between classes while holding edge probability fixed (and equal) suggests that these statistics differentiate systems based upon structure as encoded in the graph topology. This is emphasized by (1) the fine detail captured when node numbers and families are equal and edge formation probability is similar, and (2) the low performance in distinguishing 99 from 100 nodes. This is desirable as we want to learn emergent structure, and topologically we expect that there is not much difference between two Erdos-Renyi models with the same edge formation probability and very similar node counts. The analogous result holds for the other models considered, and it is reasonable to expect this generally when there is some logic, formal process, or consistent heuristics underlying construction. Given that we want to treat systems as dynamically similar even if they differ by a small number of nodes, this can be seen as a strength of the method and framework.

A potential drawback of the method is that it treats directed, signed, weighted graphs as undirected, unsigned, weighted graphs. Extension to directed graphs is straightforward and there are a number of avenues of exploration that we leave to future work.

2.7 Conclusion

In this work we used complex systems theory and theoretical machine learning to guide development of a statistical model of self-similar emergent phenomena on graphs and more

general spaces. We then empirically demonstrated *in silico* that the theory leads to a meaningful framework via a new method for classifying systems based upon structure. We connected this with perspectives from collaborative systems modeling and suggested how it could be implemented to support estimation tasks as described in chapter 3. This framework opens a number of exciting possibilities. Preliminary work incorporating the texture synthesis paradigm (He and Hirn, 2021) suggests that our statistical model might be successful at supporting wisdom-of-crowds type forecasts of complex dynamical phenomena such as the interactions among food prices and social-ecological processes in low-data environments

BIBLIOGRAPHY

Aminpour, P., Gray, S. A., Jetter, A. J., Introne, J. E., Singer, A., and Arlinghaus, R. (2020). Wisdom of stakeholder crowds in complex social–ecological systems. *Nature Sustainability*, 3(3):191–199.

Amoroso, L. (1938). Vilfredo pareto. *Econometrica: Journal of the Econometric Society*, pages 1–21.

Ayache, A. (2018). *Multifractional stochastic fields: wavelet strategies in multifractional frameworks*. World Scientific.

Batten, D. F. (2001). Complex landscapes of spatial interaction. *The Annals of Regional Science*, 35:81–111.

Bruna, J. and Mallat, S. (2013). Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886.

Bruna, J. and Mallat, S. (2019). Multiscale sparse microcanonical models. *Mathematical Statistics and Learning*, 1(3):257–315.

Gama, F., Ribeiro, A., and Bruna, J. (2019). Diffusion scattering transforms on graphs. In *International Conference on Learning Representations*.

Gao, F., Wolf, G., and Hirn, M. (2019). Geometric scattering for graph data analysis. In *International Conference on Machine Learning*, pages 2122–2131. PMLR.

Gelbaum, Z. (2014). Fractional brownian fields over manifolds. *Transactions of the American Mathematical Society*, 366(9):4781–4814.

Godsil, C. and Royle, G. F. (2001). *Algebraic graph theory*, volume 207. Springer Science & Business Media.

Gray, S. A., Zanre, E., and Gray, S. R. (2013). Fuzzy cognitive maps as representations of mental models and group beliefs. In *Fuzzy cognitive maps for applied sciences and engineering: From fundamentals to extensions and learning algorithms*, pages 29–48. Springer.

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W. M., Railsback, S. F., Thulke, H.-H., Weiner, J., Wiegand, T., and DeAngelis, D. L. (2005). Pattern-oriented modeling of agent-based complex systems: lessons from ecology. *science*, 310(5750):987–991.

Hammond, D. K., Vandergheynst, P., and Gribonval, R. (2011). Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150.

He, J. and Hirn, M. (2021). Texture synthesis via projection onto multiscale, multilayer

statistics. *arXiv preprint arXiv:2105.10825*.

Mallat, S. (1999). *A wavelet tour of signal processing*. Elsevier.

Mallat, S. (2016). Understanding deep convolutional networks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150203.

Mandelbrot, B. (1967). How long is the coast of britain? statistical self-similarity and fractional dimension. *science*, 156(3775):636–638.

Marques, A. G., Segarra, S., Leus, G., and Ribeiro, A. (2017). Stationary graph processes and spectral estimation. *IEEE Transactions on Signal Processing*, 65(22):5911–5926.

Meadows, D. H. (2008). *Thinking in systems: A primer*. chelsea green publishing.

Mmahi, O. P. and James, F. T. (2023). Brigandage and criminal victimization in nahuche community, zamfara state: impact on food security. *Environment, Development and Sustainability*, pages 1–18.

Morel, R., Rochette, G., Leonarduzzi, R., Bouchaud, J.-P., and Mallat, S. (2022). Scale dependencies and self-similarity through wavelet scattering covariance. *arXiv preprint arXiv:2204.10177*.

Perlmutter, M., Gao, F., Wolf, G., and Hirn, M. (2019). Understanding graph neural networks with asymmetric geometric scattering transforms. *arXiv preprint arXiv:1911.06253*.

Perlmutter, M., Gao, F., Wolf, G., and Hirn, M. (2020). Geometric wavelet scattering networks on compact riemannian manifolds. In *Mathematical and Scientific Machine Learning*, pages 570–604. PMLR.

Pipiras, V. and Taqqu, M. S. (2017). *Long-range dependence and self-similarity*, volume 45. Cambridge university press.

Schoenberg, F. P. and Patel, R. D. (2012). Comparison of pareto and tapered pareto distributions for environmental phenomena. *The European Physical Journal Special Topics*, 205(1):159–166.

Voinov, A., Jenni, K., Gray, S., Kolagani, N., Glynn, P. D., Bommel, P., Prell, C., Zellner, M., Paolisso, M., Jordan, R., et al. (2018). Tools and methods in participatory modeling: Selecting the right tool for the job. *Environmental Modelling & Software*, 109:232–255.

Zhang, S. and Mallat, S. (2021). Maximum entropy models from phase harmonic covariances. *Applied and Computational Harmonic Analysis*, 53:199–230.

APPENDIX

EXPERIMENTAL PARAMETERS

Parameters for these three experiments are (where n denotes the number of nodes and m the Barabasi-Albert parameter),

1. Low-Density Experiment
    a) $n = 100$, $m = 22$
    b) $n = 100$, $m = 23$
    c) $n = 100$, $m = 21$
    d) $n = 50$, $m = 9$
    e) $n = 50$, $m = 10$
    f) $n = 50$, $m = 11$
2. Medium-Density Experiment
    a) $n = 100$, $m = 44$
    b) $n = 100$, $m = 45$
    c) $n = 100$, $m = 43$
    d) $n = 50$, $m = 22$
    e) $n = 50$, $m = 23$
    f) $n = 50$, $m = 21$
3. High-Density Experiment
    a) $n = 100$, $m = 65$
    b) $n = 100$, $m = 66$
    c) $n = 100$, $m = 67$
    d) $n = 50$, $m = 35$
    e) $n = 50$, $m = 34$
    f) $n = 50$, $m = 33$

CHAPTER 3

FROM 'OUGHT' TO 'IS': A COMPARISON OF UNSUPERVISED METHODS FOR VALUES-INFORMED WISDOM OF CROWDS

3.1 Abstract

Many social and ecological problems require us to consider objectively verifiable phenomena as well as subjective states of knowledge and associated value systems. When approximating the facts of reality, the wisdom of crowds phenomenon demonstrates that many pooled estimates can be more accurate than individual or expert estimates. For complex and social systems, wisdom of crowd approaches are improved by aggregating knowledge over subpopulations. In this paper we introduce three unsupervised methods for identifying subpopulations based upon value-laden statements in narrative data from hyperlocal maternal and child health (MCH) contexts in Gombe State, Nigeria. We employ data science techniques and compare methods to assess the stability of inferences. We find the hypothesized groups to be method dependent and discuss implications for wisdom-of-crowd estimates in sustainable development contexts.

3.2 Introduction

Our most pressing social and ecological problems require consideration of not only objectively verifiable phenomena, but also more subjective states of knowledge and value systems. When approximating the facts of reality, the wisdom of crowds phenomenon empirically demonstrates that the pooled estimates of a group of individuals can often provide a more accurate model than most individual estimates (Yi et al., 2012; Galton, 1907). The key to leveraging the wisdom of crowds rests in pooling estimates. The seminal research by Galton (1907) on the wisdom of crowds presents evidence that the median of a group of scalar estimates of the weight of an ox provides a more sensible model than the mean due to its greater stability to outliers. Yi et al. (2012) subsequently showed that individual estimates of solutions to higher dimensional problems—specifically the Euclidean minimum spanning

tree and traveling salesman problems—can be combined by reference to the majority to better approximate the optimal. Each of these tasks are designed around a ground truth with respect to which estimates can be compared. This ground truth constitutes a uniquely specified problem ontology, the *is* of the problem.

Such an unambiguous specification is generally not possible, however, in complex social and ecological problems. As observed by Levin et al. (2021), many sustainability problem contexts exhibit what Breiman (2001) calls the *Rashomon effect* wherein multiple plausible models account for the objectively verifiable aspects and yet remain mutually inconsistent. This is further complicated by the *wickedness* of such problems, which signifies disagreement among stakeholders regarding the facts, the nature of the problem, and/or the manner in which it ought to be addressed (Rittel and Webber, 1973). Hence, a "ground truth" may be ill-defined. Nevertheless, human actors must engage with these problems, which compels wisdom-of-crowds researchers to develop frames of reference with respect to which progress may be assessed. Aminpour et al. (2020) do so by comparing models aggregated from fisheries stakeholders to that from a group of traditional scientific "experts." Similar to Galton (1907) and Yi et al. (2012), Aminpour et al. (2020, 2021) show that better models are attained by first averaging over stakeholder subpopulations and then aggregating based upon the median of these subgroup models, allowing noise to be reduced and the expertise to be represented. In hyperlocal contexts relevant to the present study, however, formal expert models are lacking. Instead, local stakeholders possess the expert knowledge. Researcher outsiders must, in turn, develop methodologies that operationalize this knowledge while accounting for heterogeneity among participant value systems and cognitions.

In this paper we investigate a process for hypothesizing salient subgroups within human populations based upon values that individuals express in narrative. We call these subgroups *cognitographic clusters*. Inspired by replicability issues in general science as well as the rapid expansion of machine learning into social science, we study three unsupervised methods: the traditional method of K-means, an emerging method utilizing hierarchical

density-based spatial clustering of applications with noise (HDBSCAN) and uniform manifold approximation and projection (UMAP), as well as a new method that draws upon optimal transport (OT) theory and spectral clustering. We compare method performance on data from the maternal and child health (MCH) service utilization contexts of three local government areas (LGAs) in Gombe State, Nigeria. At each step, we reiterate upon the traditional-and-emerging-method theme by employing a traditional, statistical method of analysis and another emerging method from the field of *explainable artificial intelligence* (AI). To understand the extent to which inferences are method-agnostic or -dependent we compare across methods.

3.3 Background

In hyperlocal sustainable development contexts, there is misalignment among formal expert predictions and local behavior. This stems not only from constrained disciplinary perspectives underlying expert models, but also uncertainty about, and heterogeneity among, local behavior. Failure to account for this mismatch produces undesirable outcomes that perpetuate old problems and generate new. For instance, failure to consider local burial practices in the early stages of the 2014 Ebola outbreak in West Africa led to insufficient projections of contagion, which generated the largest such epidemic in history (Maxmen, 2015). Similarly, reduction of the maternal mortality ratio has stagnated in the last decade (WHO, 2023), particularly in northern Nigeria (GSMoH, 2010) where demand for maternal and child health (MCH) services is influenced by sociocultural factors (Sinai et al., 2017). In these contexts, behavior tied to heterogeneous local value systems produces outcomes at odds with development goals.

For outsider researchers tasked with decision-making support for sustainable development missions, there is relatively little information available about cultural and system-level factors that generate changes in behavior. There is, furthermore, uncertainty about how local subpopulations understand and explain changes as a function of interactions among these factors. This Rashomon effect (Breiman, 2001) has been observed in complex social system

contexts (Levin et al., 2021) where it intertwines with the wicked problems phenomenon (Rittel and Webber, 1973) to present fundamental modeling challenges. Estimating a systems model requires accounting for this epistemic heterogeneity. What if modelers were able to leverage this diversity of perspective to improve sustainable development operations?

Transdisciplinary research methodologies like participatory modeling (PM) (Gray et al., 2015; Voinov et al., 2018) are often used to uncover connections among diverse values and knowledge, and to identify factors that explain how these manifest in behavior. Recent research in collective intelligence and PM suggests that local stakeholder knowledge may be profitably combined across subpopulations to account for scientifically validated features of complex social-ecological systems (Aminpour et al., 2020, 2021). We draw upon this wisdom of stakeholder crowds to better understand hyperlocal drivers of change in MCH utilization.

3.3.1 Wisdom of Crowds

The *wisdom of crowds* refers to the ability of groups to outperform individuals in decision-making and estimation tasks. Wisdom of crowd inferences are typically generated by pooling yes-no or scalar valued estimates made by individuals. The INFER project (INtegrated Forecasting and Estimates of Risk) and Cosmic Bazaar, for example, use crowd-sourced event probability estimates to produce early warning signals for policymakers (Team, 2023). Investigations by Yi et al. (2012) show that success in the scalar regime can generalize to certain higher dimensional phenomena, and in particular to the properties of graphs embedded in Euclidean space. Aminpour et al. (2020) further suggests that the wisdom of stakeholder crowds is able to approximate scientifically validated aspects of complex social and ecological phenomena. In the context of MCH utilization, individual estimates of the system are causal explanations. They relate system components to causal relationships among them. Fuzzy cognitive maps (FCMs) can be used to tap into this locally specific causal knowledge (Aminpour et al., 2020) of the MCH system.

3.3.2 Fuzzy Cognitive Mapping

FCMs are signed, digraph causal models that provide representations of complex dynamical systems. Factors are represented as nodes whose values qualitatively increase and decrease, and causal relationships are encoded as edges decorated with signs and weights corresponding to causal direction and magnitude, respectively. The FCM framework was introduced by Kosko (1986) to extend Axelrod's cognitive maps in a manner that accounts for the fuzziness of natural language. Fuzziness arises when attempting to precisely disambiguate, for instance, "a lot of rain" from "a little bit of rain." FCMs have been employed to understand human decision making in a wide range of complex contexts including war games, organizational management, engineering (Papageorgiou and Salmeron, 2012), and social and ecological systems (Özesmi and Özesmi, 2004; Gray et al., 2015). The graphical structure yields dynamical models that may be simulated to inform users of counter-intuitive dynamics, to identify potential leverage points and intervention strategies, and to explore the consequences of such.

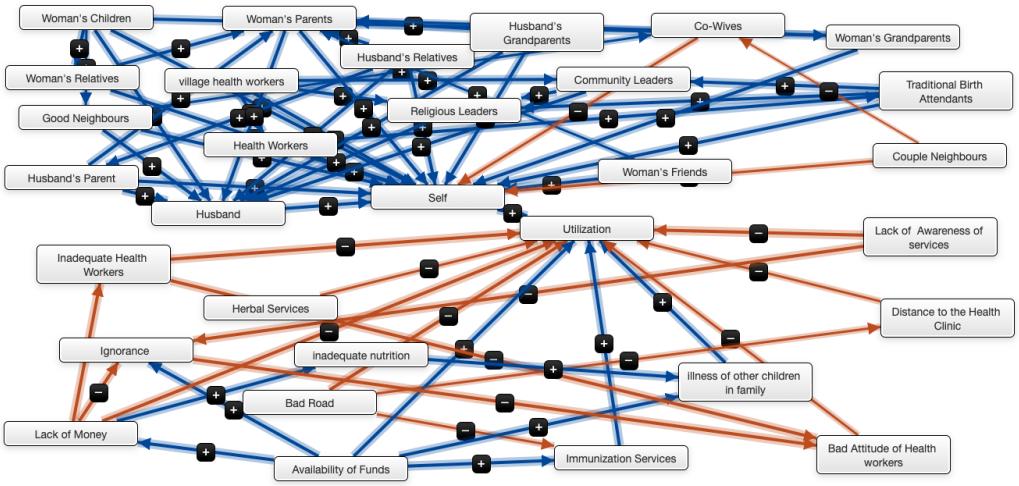

Figure 3.1 FCM from a focus group discussion in Bojude, Gombe, Nigeria

Aminpour et al. (2020, 2021) investigate the usage of FCM as wisdom of crowd models of complex social and ecological systems. Similar to previous work, they find that a well-chosen estimate pooling method is key to effectively harnessing the collective intelligence

of stakeholders. By first averaging over subpopulation models and then aggregating by their median, stakeholder models can accurately reproduce scientifically validated system models (Aminpour et al., 2020). In highly localized MCH contexts, however, locally adapted, scientifically validated models are lacking, and, hence, a formally proposed ground truth is absent. Instead, implicit representations of the target system–specifically, stakeholder explanations for system behavior–can make for useful formal models when made explicit. Supposing one has access to these mental models, how does one identify subgroups exhibiting consensus that can be meaningfully differentiated from that of other subpopulations for incorporation into wisdom of crowd models?

One approach is to begin by considering features that are universal all human systems. Since FCMs capture heterogeneous cognitions about, and explanations for, system behavior (i.e., what constitutes the system), we can augment with beliefs about how the system ought to behave (e.g., preferences, desires). By explicitly incorporating stakeholder senses of "is" and "ought," we may have some hope of successfully addressing challenges presented by the wicked Rashomon effect and thereby improve MCH outcomes along sociocultural lines. This motivates our consideration of human values.

3.3.3 Values

Anthropologist David Graeber (Graeber, 2001, p.47) defines *values* as:

> ...the way people represent the importance of their own actions to themselves: normally, as reflected in one or another socially recognized form.

Values correspond to *oughts* expressed by individuals and groups, as in, "the federal government ought not restrict access to reproductive health services." Through beliefs, norms, attitudes, and behavioral intentions, values indicate–and are indicated in–behavioral tendencies and shared ideals of human groups (Schwartz, 2006). The behavioral implications of values suggests their consideration may be useful for modeling interventions in uncertain sustainable development contexts like localized MCH utilization. We can draw on domain

general quantitative approaches like the World Value Survey (WVS) to support these efforts.

3.3.3.1 World Values Survey

Inglehart's model of cultural values relies on data collected through the World Value Survey for the last 40 years in 60 countries around the world (Inglehart and Welzel, 2010). The model covers individual and country-level dimensions and theoretically rests upon the assumption that economic development is linked to distinct value orientations. The two most important cross-national dimensions that emerged through the analysis of a broad range of political, social and religious norms and beliefs are a polarization between traditional/sacred versus secular/rational orientations towards authority and survival versus self-expression values, providing a comprehensive measurement of all major areas of human concern, from religion to politics to economic and social life. Each society can be located on a global map of cross-cultural variations based on these two dimensions in Figure 3.2 (Inglehart, 2020).

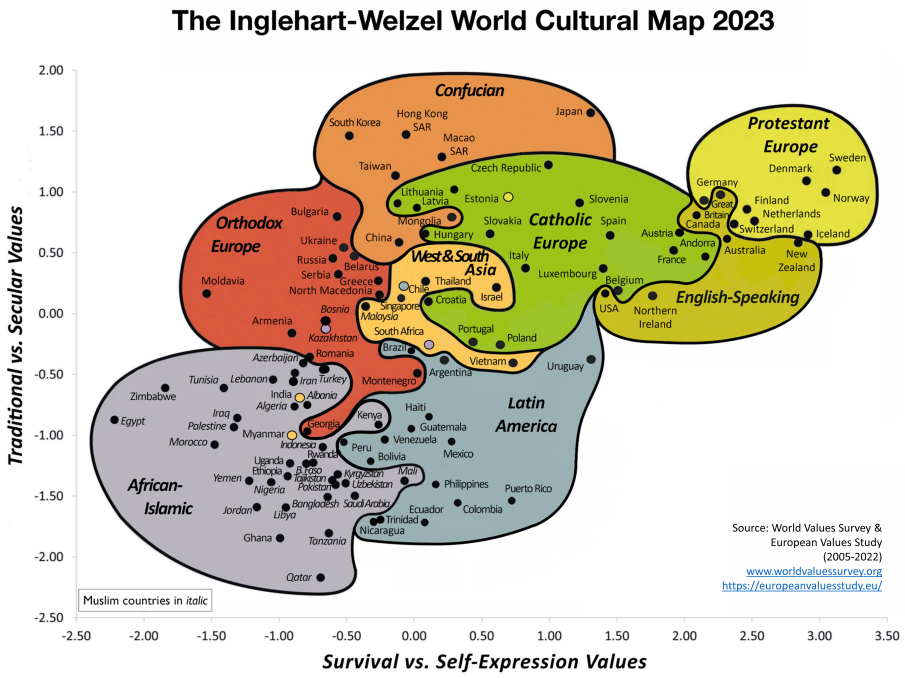

Figure 3.2 Inglehart-Welzel World Cultural Map 2023, reproduced from Inglehart (2020)

A particular point on the cultural map reflects the relative position on a number of topics that often form the basis of influence campaigns because they are tied to core values. Tradi-

tional values emphasize the importance of religion, parent-child ties, deference to authority and traditional family values. People who embrace these values also reject divorce, abortion, euthanasia and suicide. Secular-rational values have the opposite preferences to the traditional values. These societies place less emphasis on religion, traditional family values and authority. Divorce, abortion, euthanasia and suicide are seen as relatively acceptable. Survival values place emphasis on economic and physical security. They are linked with a relatively ethnocentric outlook and low levels of trust and tolerance. Self-expression values, by contrast, give high priority to environmental protection, growing tolerance of foreigners, gays and lesbians and gender equality, and rising demands for participation in decision-making in economic and political life.

Using factor analysis these dimensions can be reduced to 10 items, which cover approximately 70% of all cross-national variation of values (Inglehart et al., 2000). However, societies vary widely. For example, in Pakistan or Nigeria, 90% of the population say that God is extremely important in their lives, while in Japan only 6% take this position. Similarly, countries where survival values are high also exhibit low levels of subjective well-being, poor health, low interpersonal trust, intolerance of outgroups, and low support for gender equality. These countries tend to perceive cultural diversity as threatening, emphasize materialist values, absolute rules, and familiar norms, and favor authoritarian forms of government. Foreigners tend to be seen as dangerous outsiders who reduce already scarce resources. The opposite is seen in countries with low survival measures. Because these dimensions are highly correlated, if people in a given population place a strong emphasis on religion, the relative position of that population on many other variables can be predicted.

Recently, researchers in the field of natural language processing (NLP) trained a model to help identify WVS values that resonate within narrative or ethnographic text (Benkler et al., 2022). In this paper, we explore the possibility of generating cultural clusters directly from text utilizing this model alongside other automated techniques. These are used to support modeling of MCH utilization.

3.3.4 Maternal and Child Healthcare (MCH) Utilization in Nigeria

Improving outcomes and addressing inequalities in maternal and child health (MCH) has been a central component of global health initiatives in the last two decades, figuring prominently in the Millennium and later Sustainable Development Goals and international humanitarian and development efforts. However, while the global maternal mortality ratio (MMR) declined by 33% between 2000 and 2015, it has remained stubbornly stagnant since 2016, despite myriad initiatives around the world to improve global maternal outcomes (WHO, 2023).

Significant disparities persist in maternal outcomes across the world: In 2020, the lifetime risk of maternal death in low-income countries was 1 in 49, compared to 1 in 5,300 in high-income countries (WHO, 2023). Nigeria's national MMR is one of the highest in the world at 917 per 100,000 live births (WHO, 2023), and in Gombe State the MMR is even higher at 1,002 per 100,000 live births (GSMoH, 2010). Quality maternal health services can prevent nearly all maternal death, but availability of, demand for, and utilization of these services is low in many of the countries that experience the highest rates of maternal mortality (WHO, 2023). Nigeria and Gombe State is no exception here either: the WHO standard for healthcare worker density is 4.45 healthcare workers per 1000 population (WHO, 2023), but the density in Nigeria is just 2.52 per 1,000 (Uzochukwu, 2017), and the density in Gombe State is even lower at about 1 per 1,000 (GSMoH, 2010).

Demand for maternal health services in Northern Nigeria may be explained by socio-cultural factors including religion, tradition, urbanization, education, family structure, and marital practices (Sinai et al., 2017) as well as communal and cultural norms (WHO, 2023). Understanding interactions among social determinants can support development of interventions that reduce social and structural barriers to utilization, which disproportionately impact socially marginalized women and girls (WHO, 2023). However, highly granular social determinant data is lacking and resource intensive to collect. Furthermore, locally specific cultural factors may not be accounted for within existing scholarly literature, which often

span national or multinational regions. To better understand drivers of change in MCH utilization as a function of cultural practice in hyperlocal spaces, such as at the local government area (LGA) level, stakeholder knowledge may be combined across subpopulations.

3.3.5 Summary

In this study, we wish to simultaneously account for the cognitions and value systems of local stakeholders to inform wisdom of crowd models of maternal and child healthcare utilization within three local government areas (LGAs) in Nigeria. Tapping into the wisdom of the crowds in such complex social system contexts requires a method of pooling estimates that leverages both consistencies in understanding and diversities of perspective (Aminpour et al., 2021) and produces low variance estimates. Aminpour et al. (2020, 2021) find that this diversity can be accounted for by first averaging over subpopulation FCMs and then aggregating by the median of the averaged maps. We propose to perform what we term a *cognitographic analysis*, that is, to identify subpopulations who express similar value systems. To this end, we introduce three unsupervised methods of cognitographic clustering and compare using a variety of measures.

3.4 Methods & Materials

3.4.1 Overview

We introduce a narrative dataset in which expectant mothers and fathers self-signify espoused values and cognitions. We then describe an automated procedure called *recognizing value resonance* (RVR) which uses *natural language processing* (NLP) to efficiently identify values expressed within these narratives. Each document is associated with a feature vector of densities over value hypotheses. To understand mesoscale similarities and differences among these features, we employ three clustering methods that attend to different aspects of the data.

Our methodology is motivated by replicability issues in general science. As such we adopt a data science perspective, defined as *the science of learning from data* (Donoho,

2017). We are specifically motivated by a result of Bernau et al. (2014) as discussed in Donoho (2017). Bernau et al. (2014) presents a meta-analysis of multiple medical studies that evaluates a family of models across a family of data sets and assesses their performance. Interestingly, the most generally successful model is found to have been developed on the most troublesome data set. Donoho (2017) further identifies a cross-workflow analysis in medicine, Madigan et al. (2014), which demonstrates the instability of inferences to various workflows applied to the same data set. This instability produces not only differing inferences but also contradicting conclusions. Similarly, our approach studies and compares three clustering methods on a single data set. Analyses are performed with traditional and emerging methods.

3.4.2 Recognizing Value Resonance in Narrative

3.4.2.1 Narrative Completion Data

The text data which we analyze for value resonance in this work is a collection of story completions generated by local participants surveyed in Gombe state, Nigeria as a response to narrative prompts provided by survey workers in the field. Narrative prompts for story completion is a method aimed at collecting rich linguistic data containing information about respondents' implicit knowledge and worldviews in a specific domain, in contrast to responses obtained by posing explicit questions devoid of context and imagery. The narrative prompts used in data collection were specifically designed to elicit locals' reasoning in the context of maternal and child health care (MCH) as practiced in their geographical locale. Each participant was provided (in sequence) with three narrative prompts randomly selected from a pool of 45 prompts created to evoke a plausible scenario in the local context. Of these, 22 prompts were targeted toward male respondents, and 23 were targeted toward female respondents. The respondents were then asked to complete the story in a few sentences, which were audio-recorded by the enumerators and later translated from Hausa into English by a professional translator.

The narrative prompts were written in accordance with the precede-proceed model of public health behavior (Green, 1974; Green and Kreuter, 1991), which offers a typology of

three top-level categories of factors influencing health care related beliefs and behaviors on the individual and group levels. The three main sets of factors are:

1. Predisposing (e.g. education, rootedness in one's own culture, prior experience)
2. Reinforcing (e.g. influence of family members, friends, community leaders, the media, and cultural authorities)
3. Enabling (e.g. Physical and financial security, access to health providers, quality of available care)

Respondent sex	Narrative prompt	Story completion
Female	Sarah is not progressing well during labor, and the staff in the local clinic call for help at a bigger hospital. They tell her husband, who...	... decides to take her to the bigger hospital as advised by the hospital staff, there emergency care is given to her by the doctor, and she safely delivers a baby boy.
Male	Asah runs a clinic in his village. He notices that recent mothers nearby are not bringing their children in for postnatal visits. He asks around to find out why that is. He learns...	...that in the clinic the women do not receive appropriate medical care, others also complained that their husbands don't give them permission to go because even if they go they are not cared for.

Table 3.1 Examples of narrative prompts and story completions from male and female perspectives

Table 3.1 shows two examples of a narrative prompt[1] intended for a respondent and a story completion from one of the respondents. Factors present in both the prompt and the response are highlighted in colors representing their categorization according to the precede-proceed model (red for predisposing, green for reinforcing, and blue for enabling).

While the goal of the story completion approach is to instantiate otherwise implicit knowledge and beliefs devoid of bias introduced by explicit lines of questioning, the scenarios

[1]Note: predisposing in red, reinforcing in green, and enabling in blue

present in the prompts reflect factors a priori believed to bear significant weight on locals' reasoning in the MCH context. The set of factors embedded in these hypothetical situations included gender imbalance in household decision-making power in the region, prevailing low levels of financial security, limited physical access to health care, and the common use of traditional health practice in contrast to institutional care, among others.

To collect the story completion data, the survey team randomly selected three wards[2] in Gombe state. The target population within each ward was defined using a stratified cluster sample. Survey enumerators then utilized a systematic walk to sample participants from the target community. Apart from recording the story completions, enumerators collected a rich set of demographic information for each respondent. Answers to 37 different demographic questions were solicited, including individual biological, social and cultural characteristics, family information, and living conditions. Figure 3.3 shows the frequency of demographic categories for a subset of these questions.

All experimental protocols were approved by Health Media Lab (HML), a U.S.-based IRB registered with the US Department of Health and Human Services DHHS OHRP, and by a Health Research Ethics Committee (HREC) managed by the Gombe State Ministry of Health (SMOH) in Gombe, Nigeria. The team carried out the research in precise accordance with the proposed approach, using the consent forms, instruments, items, and materials reviewed and approved by HML and the Gombe SMOH HREC, and in accordance with the legal and research ethical guidelines and requirements as stated by HML and the Gombe State Ministry of Health. The approach, materials, consent forms, instruments, items, and activities were also reviewed and overseen by the Army Human Research Protection Office (HRPO) in accordance with Department of Defense Human Subjects Research (HSR) requirements. All participants were given study information including privacy protection and anticipated risk information, as well as contact information for the study PI, the Gombe SMOH HREC, and the Health Media Lab, and provided informed consent before participating in the study.

[2]Lowest level administrative unit in the state, usually centered around an *anguwa* ("town" or "village")

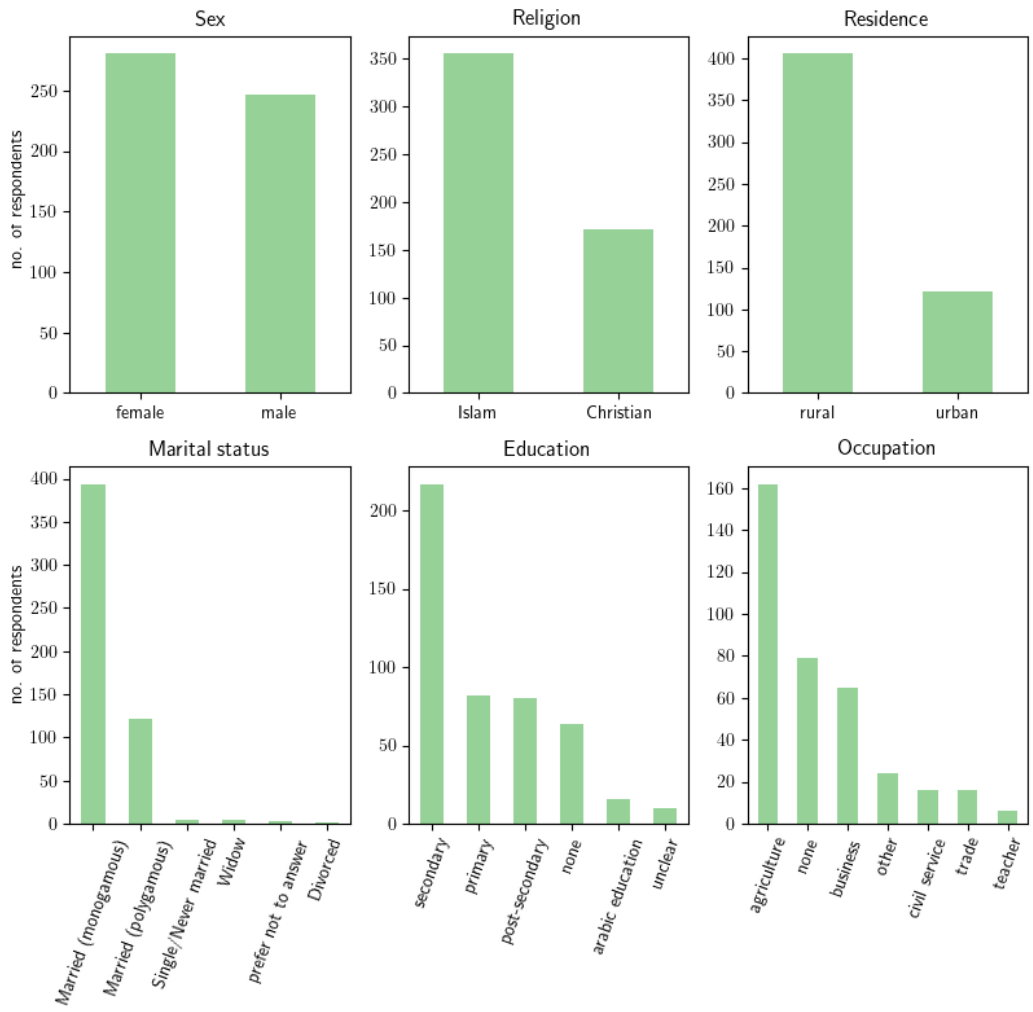

Figure 3.3 Sample demographic frequencies

3.4.2.2 Recognizing Value Resonance Model

Recognizing Value Resonance (RVR) is a natural language processing task concerned with detecting implicit endorsement of, rejection of, or neutrality towards a certain belief given a span of text (Benkler et al., 2022). Given a belief hypothesis H and textual premise P, H 'resonates' with P if P communicates the speaker's belief in H, 'conflicts' with P if P communicates the speaker's opposition to H, and is 'neutral' if neither is true. A 2022 study published a hand-annotated dataset, the World Values Corpus (WVC), and a fine-tuned model, Resonance-Tuned RoBERTa designed to model the task of RVR (Benkler et al., 2022). The value hypotheses in the WVC were derived from the World Values Survey

questions. We utilize a technique described in the study's analysis of folktales to extract the document level value-density vectors behind our cluster analyses presented in this paper.

We extract document-level value density vectors as follows. First, each document is parsed out sentence by sentence. Next, each sentence is paired with each of the 384 value hypotheses from the WVC. Each ⟨premise sentence, hypothesis⟩ pair is then scored for RVR using Resonance-tuned RoBERTa. Finally, for each value hypothesis we calculate the proportion of sentences in which it was scored as "resonates", and the proportion of sentences in which it was scored "conflicts." This essentially returns two vectors of length 384 where each vector details the document-level density of a single RVR score for every value hypothesis in the WVC $(H_1, \ldots, H_{384})$. We then prepend the "resonates" vector to the "conflicts" vector to create a vector of length 768 delineating the sentence level density distribution of all the WVC values within each document.

We focus on value hypotheses that associate with religion, social and gender roles, institutions, as well as self-expression, safety, and tradition (see Appendix 3.8 for full list). In an attempt to reduce the dimensionality of the space, we calculate the story level density distribution of all the WVC values and select values that resonate or conflict, at minimum, in a fraction q of the story completions as dimensions along which the k-means clustering is to be performed. For the experiments which we detail in this work we used $q = 0.05$.

3.4.3 Clustering

3.4.3.1 Overview

Clustering is an unsupervised machine learning approach to pattern recognition that assigns a finite set of labels to data points relative to some measure of (dis)similarity. Formally, given a dataset, $\mathcal{X}$, and data points, $x_i, x_j \in \mathcal{X}$, clusters are generated based upon,

$$d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$$
$$(x_i, x_j) \mapsto d(x_i, x_j),$$

a mapping symmetric in its arguments (Hastie et al., 2009). In Euclidean space, for instance, we have $d(x_i, x_j) = ||x_i - x_j||_2$. Clusters are then identified with respect to this local measure based upon statistical, geometrical, and/or other globally defined decision heuristics. The following introduces three unsupervised clustering methods: The traditional method of k-means, an emerging method utilizing HDBSCAN and UMAP, and a method based upon optimal transport and spectral clustering.

3.4.3.2 k-means

The k-means clustering method is one of the most widely used classical tools in the machine learning problem space. The aim of k-means is to partition a set of n observations into k clusters such that each observation belongs to the cluster whose centroid (the mean of its members) is closest to it in terms of the L^2-norm. This problem is NP-hard, however, efficient heuristics eliciting locally optimal solutions exist. The number of clusters k is a free parameter whose "optimal" or "appropriate" value is context-dependent and often ill-defined. The method assumes real-valued data on an interval scale; thus, when working with categorical or ordinal data, mappings to interval scales are necessary.

We use the k-means method here due to its widespread familiarity and relative accessibility both in terms of applying the basic algorithm as well as interpreting the resulting partition. Focusing on interpretability in particular, one may simply think of the clusters as compact "clouds" of points in Euclidean (albeit often highly dimensional) space that highlight the existing separation or gaps in the population along one or more dimensions.

In the context of partitioning our target population (recent mothers and fathers in Gombe state, Nigeria) into clusters based on the value sets that their members (implicitly) espouse, we first determine the resonance of or conflict with each value from the WVC in all of the collected narrative prompt story completions[3] using the RVR model. For each WVC entry (counting resonance and conflict with each value as two separate values), we introduce a new data dimension, with numeric values of 1 (where WVC values resonate/conflict in a given

[3]After removing items with obvious data entry errors and null or otherwise invalid story completions.

text) and 0 (where WVS values do not resonate in the given manner in a response).

We then perform k-means clustering along the selected value resonance dimensions with $k = 1, 2, \ldots, 20$[4] and compute the sum of squared errors (SSE) between points and their centroids for each k-partition. We then select a partition that lies on the "elbow" of the partitions arranged in the two-dimensional (k, SSE) space, that is, the first partition into k' clusters, such that the improvement in SSE between the k' +1 and k' partitions is below some threshold value[5].

3.4.3.3 HDBSCAN & UMAP

Uniform manifold approximation and projection (UMAP) and hierarchical density-based spatial clustering of applications with noise (HDBSCAN) are two popular machine learning algorithms used for dimensionality reduction and clustering tasks, respectively. UMAP is a nonlinear dimensionality reduction algorithm that constructs a high-dimensional graph representation of the data and then optimizes a low-dimensional projection of this graph, preserving both global and local structure of the data (McInnes et al., 2018). HDBSCAN is a clustering algorithm that automatically determines the number of clusters in a dataset, unlike traditional clustering algorithms that require the user to specify the number of clusters in advance. HDBSCAN identifies high density regions of data and groups together points in these regions, making it highly useful for handling datasets with varying densities and noise (McInnes et al., 2017). Both UMAP and HDBSCAN have gained popularity in recent years due to their performance on large datasets (Allaoui et al., 2020; Blanco-Portals et al., 2022; Pealat et al., 2021; Asyaky and Mandala, 2021).

These algorithms are commonly used together in machine learning workflows for several reasons. First, by reducing the number of dimensions, the data becomes more manageable for clustering algorithms like HDBSCAN, which can be computationally expensive on high-dimensional data. Second, UMAP can improve the separation of clusters by preserving the

[4]Higher values were not tested under the assumption that such high numbers of clusters would be of little value to analysts pursuing further insight into cognito-cultural determinants of MCH utilization (and most other domains for that matter),

[5]This choice is, once again, subjective and to some extent arbitrary.

local structure of the data. This can make it easier for HDBSCAN to identify clusters and reduce the likelihood of false positives or false negatives. Lastly, UMAP can help to reduce noise in the data by separating out irrelevant features. This can help to improve the accuracy of the clustering results by reducing the impact of noisy data points. In short, by combining these two techniques, it is possible to gain a more comprehensive understanding of the structure of high-dimensional datasets and identify meaningful patterns and relationships within the data.

The use of UMAP and HDBSCAN together is promising for clustering analysis of our RVR output vectors for three reasons. Firstly, HDBSCAN is robust with respect to noise, reducing the potential for noisy input documents to bias the output clusters. Secondly, since HDBSCAN does not require pre-specification of the number of output clusters, it may be less influenced by researcher bias. Thirdly, UMAP enables high-dimensional data to be represented in fewer low dimensions, which supports the density-based clustering approach of HDBSCAN. To grasp the latter benefit, observe that our RVR output vectors include both 'resonant' and 'conflict' label densities for all hypotheses, and clustering the raw vectors may be problematic due to multicollinearity. UMAP can reduce our RVR vectors to two-dimensional coordinate pairs, with the x-axis representing the document position in the UMAP 'resonant' space, and the y-axis representing the position in the 'conflict' space. HDBSCAN can, then, cluster RVR vectors based on their location in a two-dimensional 'resonant' vs. 'conflict' space.

Below we describe our process using UMAP and HDBSCAN to cluster the RVR output vectors. Let RVR denote the set of all RVR output vectors for all the narrative documents we process. For HDBSCAN, an individual RVR output vector for document d is conceptualized as such $\text{rvr}_d = [r_1, ..., r_{45}, c_1, ..., c_{45}]$, where r_h is the proportion of sentences in document d in which hypothesis h is labeled 'resonant,' and c_h is the proportion of sentences in which hypothesis h is labeled 'conflicts.' We begin this clustering approach by dividing the space into two subspaces, RVR_r, the 'resonant' set of all RVR output vectors, and RVR_c, the

'conflicting' set of all RVR output vectors. We then use UMAP to approximate each of these spaces independently and return a one-dimensional representation for each vector's position in its corresponding reduced space. We then combine these representations to form the set of coordinate pairs $\text{RVR}_{\text{reduced}}$ representing all the narrative documents' positions in the UMAP reduced 'resonant' vs 'conflicting' space. We then utilize HDBSCAN with the Minkowski distance metric to cluster this set of coordinate pairs.

3.4.3.4 Optimal Transport & Diffusion Maps

To amplify similarities and differences based upon the RVR data structure, we introduce a novel method based upon optimal transport (OT) theory. For this method, individual value resonance densities are conceptualized as normalized histograms encoding the frequency of 'resonance,' 'neutrality,' and 'conflict' with a particular value hypothesis by all sentences within a set of narrative documents produced by an individual participant.

OT was introduced to model the efficient allocation of resources (Peyré et al., 2019). Given two probability distributions on a metric space, it provides the most efficient plan for transporting (or transforming) one into the other subject to some costs of doing so. With this transport plan comes a distance quantifying the total cost, the *optimal transport distance*. Intuitively, given a mound of soil and a hole of equal volume, this distance gives the lowest cost for filling the hole with the soil. Larger distances correspond to higher costs. Formally, given two value resonance densities μ and ν, a symmetric matrix M encoding transport costs, and denoting the set of all couplings with marginals μ and ν as $\Pi(\mu, \nu)$, the OT distance solves,

$$W(\mu, \nu) = \min_{\gamma \in \Pi(\mu,\nu)} \langle \gamma, M \rangle_F,$$

where $\langle \cdot, \cdot \rangle_F$ denotes the Frobenius inner product. In the present context, it ought to cost more to change a narrative 'resonance' into a 'conflict' (and vice versa) than to a neutrality. That is, expressing, "I believe in God," is further from, "I do not believe in God," than it is silence on the matter. This is assured by defining a cost matrix, M, for which there is an

order of magnitude difference between the cost of transporting from 'resonate' to 'conflict' (and vice versa) and the cost of transporting between 'resonate' or 'conflict' and ''neutrality', e.g.,

$$M = \begin{bmatrix} 0 & 10 & 100 \\ 10 & 0 & 10 \\ 100 & 10 & 0 \end{bmatrix},$$

where ordered rows/columns correspond to 'resonance,' 'neutrality,' and 'conflict.' To amplify areas of greatest participant-expressed diversity within data, each pairwise value hypothesis distance is weighted according to the sample population-level entropy defined below. Then, given these locally defined OT distances between value resonance densities, the data is macroscopically organized for clustering via *spectral embedding*.

Spectral clustering refers to a family of methods that cluster data by applying k-means not to the data coordinates themselves, but to the coordinates of an embedding of the data into an Euclidean space where distances correspond to more general metrics on the data (Hastie et al., 2009). The embedding functions map similar data points nearer to each other and dissimilar points further apart via the coordinates of eigenfunctions of Laplacian type operators on inner product spaces of functions defined on the data points.

Initial exploratory analyses indicated the global geometry of the data to be influenced by regions of high sampling density, which motivated a density normalized approach inspired by diffusion maps (Coifman and Lafon, 2006). The diffusion maps framework accounts for sampling by normalizing via kernel density estimation. This allows for the partial disentanglement of statistics and geometry and thereby facilitates separation based upon differences among features and coalescence upon similarity. A diffusion map treats each data point not as a point, but as a one-parameter family of probability densities that can be thought of as 'fuzzy' combinations of data points and their neighbors. This yields a nested collection of distances that correspond to the overlap (or lack thereof) of posterior densities of t-step random walks centered at each data point on the value resonance manifold.

To implement diffusion maps, we convert the OT distances–measures that increase with greater *dissimilarity*–to affinities, which increase with greater *similarity*. We employ the standard kernel machine method of exponentiating the negative of a scaled squared metric, sometimes called a radial basis function when applied to Hilbertian metric spaces (Smola and Schölkopf, 1998). This affinity takes the form,

$$K(\mu, \nu) = e^{-\frac{d(\mu,\nu)^2}{\epsilon}},$$

where $d(\cdot, \cdot)$ is the entropy-weighted linear combination of OT distances between value resonance densities, μ and ν, corresponding to two participants. That is,

$$d(\mu, \nu) = \sum_{h \in \mathcal{H}} \alpha_h d_h(\mu, \nu),$$

where $\mathcal{H}$ is the set of value hypotheses, $d_h(\cdot, \cdot)$ is the OT distance between the h^{th} value hypothesis density of value resonance densities μ and ν, and α_h is the normalized entropy associated to the h^{th} value hypothesis. Normalization ensures $\sum_{h \in \mathcal{H}} \alpha_h = 1$, which is standard when defining data-adapted dissimilarity measures (Hastie et al., 2009). We observe that taking $\alpha_h = \frac{1}{|\mathcal{H}|}$ for all h yields an aggregate distance, $d(\cdot, \cdot)$, that is essentially the OT distance between densities on the product space of the values hypotheses with infinite cost of transporting between different hypotheses. ϵ is a small positive hyperparameter, called the bandwidth, that must be tuned. We do so by the multiscale eigenvalue approach to spectral clustering introduced in Little et al. (2020). This simultaneously tunes ϵ and $\hat{K}$, the estimated number of clusters, by considering eigenvalues, λ_k, of the operator introduced below as a function of ϵ (note: De Plaen et al. (2020) proves there is a regime of ϵ for which this kernel is positive semi-definite).

Following the diffusion maps approach, we form the diffusion operator, $\mathbf{P} = \mathbf{D}^{-1}\mathbf{K}$, where $\mathbf{D}$ is the diagonal matrix of row sums of $\mathbf{K}$. The i^{th} row of $\mathbf{P}$ is a probability density centered at data point $x_i \in \mathcal{X}$. Powers of $\mathbf{P}$ are taken to denoise the data wherein it acts as a low pass filter. This denoising approach has been similarly applied to the study of noisy, high-dimensional biological data (Moon et al., 2019). This simultaneously spreads each

probability density over neighboring data points and attenuates high frequency fluctuations, allowing for dominant meso- and macroscopic patterns to emerge. The eigenfunctions of P are orthogonal on a weighted inner product space of functions on the data and provide the embedding coordinates for clustering via k-means. The hypothesized number of clusters, $\hat{K}$, is chosen by observing the largest difference between adjacent eigenvalues, known as the spectral gap statistic, for which all other eigenvalues are small (note: this formalism corresponds to observing the bottom eigenvalues of $\mathbf{I}-\mathbf{P}$, which are $1-\lambda_k$ for all $\lambda_k \in \Lambda(\mathbf{P})$, the spectrum of $\mathbf{P}$). This is known to be optimal under certain assumptions when a ground truth exists (Little et al., 2020), and it is a sensible geometrically informed heuristic in general.

3.4.4 Cluster Comparative Analysis

3.4.4.1 Within-Method

Clustering methods are assessed with respect to 1) the generated clusters, 2) the between-cluster statistics of demographic and value resonance features, and 3) SHAP values, a method intended to increase model interpretability.

To understand similarities and differences between clusters, we consider between-cluster statistics of demographic features and value hypothesis resonances. These assess the degree to which clustered subpopulations may be differentiated according to objectively verifiable information. Demographic metadata associated with each participant is held out for clustering. Dummy variables are generated for respondent categorical demographics, and scalar demographic features are centered and standardized to unit variance. Cluster means are then compared, and demographic variables are identified for which this contrast is significant for at least one pair. The tested demographics include (a full list is provided in the Appendix 3.8):

1. Age of respondent
2. Number of children in household
3. Monthly income

4. Religion (Christian, Islam)
5. Marital status (married monogamous, married polygamous)
6. Urban or rural household

Between-cluster differences in value hypothesis resonance and conflict are also tested.

To gain qualitative insight into the clusters and methods, a surrogate model is introduced. Surrogate models are simplified models built to approximate the behavior of complex, black-box algorithms and provide insight into their predictions. They are generally used in explainable AI to help humans understand and interpret the predictions of the complex models (Danilevsky et al., 2020). We employ a surrogate modeling system to understand, explain, and interpret clustering results.

The first step in our surrogate modeling system is training a random forest classifier to predict the cluster label of each document using their corresponding value density vector as the random forest's input. A random forest (RF) classifier is an ensemble learning method that creates a large number of decision trees, each of which is trained on a subset of the training data and a random subset of the features. This randomness helps reduce overfitting and increases the generalization performance of the model. An RF prediction is made by aggregating the predictions of all individual decision trees. The main advantages of the RF algorithm are its relative ease of use, minimal hyperparameter tuning required, robustness to outliers and missing data, and high explainability[6].

After constructing our random forest surrogate model we apply a model explainability technique called SHAP (SHapley Additive exPlanations). SHAP is a technique for explaining the predictions of machine learning models (Lundberg and Lee, 2017). SHAP values are based on the concept of Shapley values from cooperative game theory, which assign a contribution to each player in a coalition game based on their marginal contribution to the coalition's success. SHAP values decompose model predictions into linear contributions from each feature. The SHAP value of a feature is the average change, over all possible feature

[6]It is worth qualifying that RF classifiers are complex for humans to interpret given the forest complexity, but highly explainable with mechanized processing.

combinations, in a prediction as the feature value is varied.

The SHAP values of our random forest surrogate model communicate the contribution of each feature (value hypothesis) to classifying a document as belonging to each cluster according to the random forest approximation of our clustering methods. SHAP value outputs indicate:

1. The most important features in document classification, ranked both globally (across all clusters) and locally (within-clusters).
2. The local direction and strength of feature impact. A positive SHAP value indicates a positive contribution to classification in the relevant cluster; a negative SHAP value indicates a feature negatively contributes.
3. The magnitude of the SHAP value, which indicates the strength of the feature's contribution to a document's classification.

3.4.4.2 Between-Method

We introduce the Jaccard score to quantify similarities between clustering methods as a function of consistencies between clusters as sets. Given two clusterings of the same data set, A and B, the Jaccard score between two clusters $a \in A$ and $b \in B$ is defined as,

$$J(a, b) = \frac{|a \cap b|}{|a \cup b|}.$$

The Jaccard similarity score maps cluster pairs to the interval $[0, 1]$, with 0 assigned to clusters that share no members and 1 to identical clusters.

To evaluate the similarity of different clustering methods based on predicted relative feature importance in cluster assignment, we utilize pairwise Rank Biased Overlap (RBO) scores (Webber et al., 2010). We first create ordered lists of features for each clustering algorithm using the average impact of each feature on model output magnitude (mean(|SHAP|)). We then evaluate the pairwise similarity between each pair of lists using a RBO score. The RBO score is a robust measure of similarity between two rankings that considers rank positions and depth of overlap between two lists. To account for top weighted-ness, we specify that

the top 10 most important features should be weighted as 85% of the final similarity score. RBO scores range from 0 to 1, with higher values indicating greater similarity between the two rankings.

3.5 Results

3.5.1 Exploratory Analysis

Densities of mean value hypothesis coefficient magnitudes are depicted in Figure 3.4. Vertical bars are the sample means for each value hypothesis. We see the majority of narratives are identified by the RVR model as neutral to most value hypotheses. This suggests that expressions of resonance and/or conflict with value hypotheses are the most informative, i.e., data regions where participant expressed diversity is greatest.

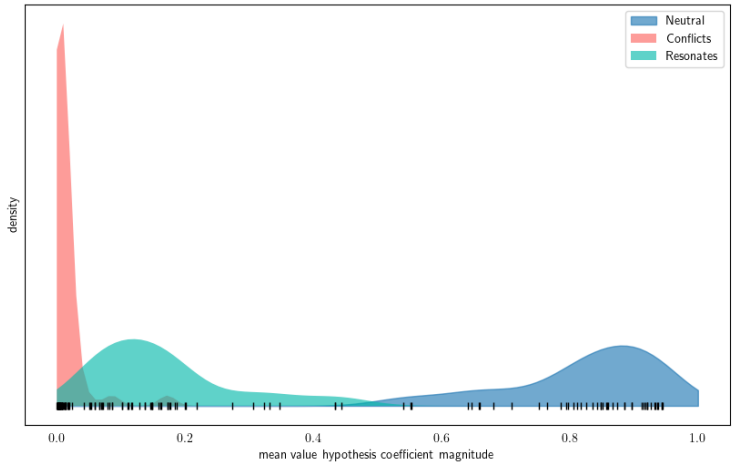

Figure 3.4 Sample level value hypothesis densities, averaged over the sample for each value hypothesis

In all of the following, clusters are sequentially numbered by size with 0 corresponding to the largest cluster.

3.5.2 k-means

3.5.2.1 Overview

Employing the elbow-plot heuristic to choose a reasonable number of clusters, we observe that the sum of squared errors (SSE) drops off less rapidly when moving from 4 to greater numbers of clusters. Hence, the number of clusters is estimated as $\hat{K} = 4$.

3.5.2.2 Statistical Differences Between Clusters

Demographics Demographic characteristics for which there is at least one statistically significant difference between clusters ($p < 0.1$) appear in Figure 3.5. These include respondent sex, highest level of educational attainment (post-secondary), occupation (business), age of head of household, and number of deceased children.

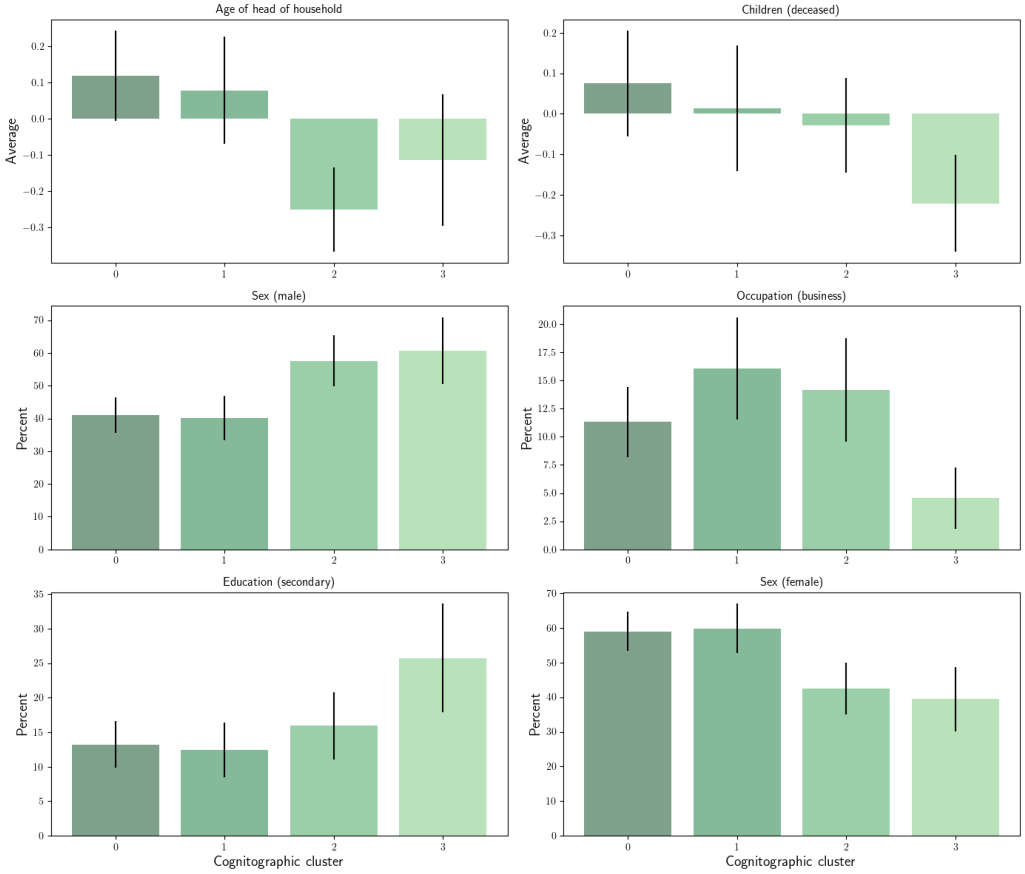

Figure 3.5 Significant between-cluster demographics for k-means

Clusters 0 and 1 exhibit a larger proportion of female respondents than clusters 2 and 3.

Cluster 2 respondents report younger heads of household than larger clusters. As compared with larger clusters, cluster 3 respondents report more post-secondary education and fewer deceased children, with a smaller proportion identifying as businesspersons.

Value Hypotheses Displayed in Figure 3.6 are cluster mean value resonances and conflicts for which one or more between clusters difference is statistically significant ($p < 0.1$). Clusters are ordered largest to smallest, left to right.

Figure 3.6 Significant value hypotheses for k-means

K-means identifies a largest, neutral cluster for which the magnitude of most coefficients is small. This contrasts with cluster 3, the smallest, which exhibits largest overall coefficient magnitudes, particularly for value hypotheses related to social roles and tradition. The greater proportion of male respondents in cluster 2 coincides with larger value hypothesis coefficients favoring men. The larger magnitudes could indicate a tendency towards expression of values.

3.5.2.3 Cluster Explainability

Figure 3.7 shows mean SHAP value magnitudes aggregated across clusters and ordered from largest. All value hypotheses are notably resonances, denoted with the suffix ".RES." These are largely related to gender roles, social actors, and healthcare services.

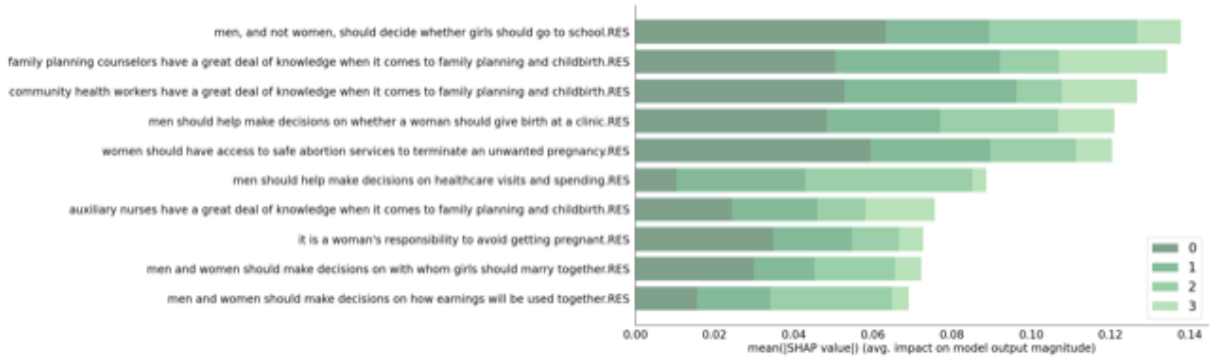

Figure 3.7 Ordered SHAP values for k-means

As with aggregate SHAP values, individual clusters present largest SHAP values for resonances (.RES). Furthermore, value hypotheses with strongest impact relate to gender and healthcare roles. Smaller clusters exhibit larger resonance magnitudes (bright red in the SHAP plots) associated with cluster membership, qualitatively reiterating the statistical findings Figure 3.5. SHAP values associated with cluster 3 notably emphasize the knowledge of social actors, whereas cluster 2 emphasizes gender roles. Clusters 0 and 1 exhibit mixtures of these.

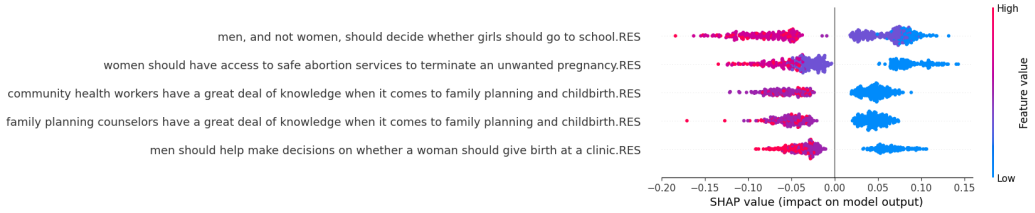

Figure 3.8 Top 5 SHAP values for k-means cluster 0

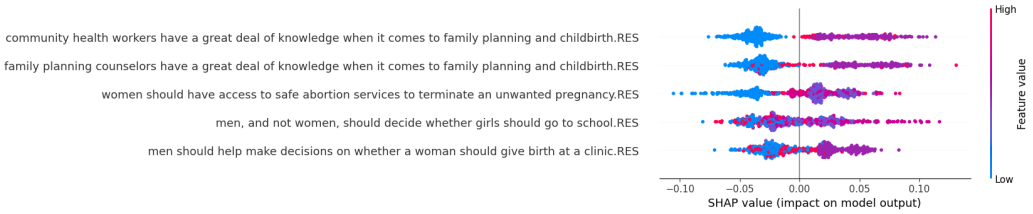

Figure 3.9 Top 5 SHAP values for k-means cluster 1

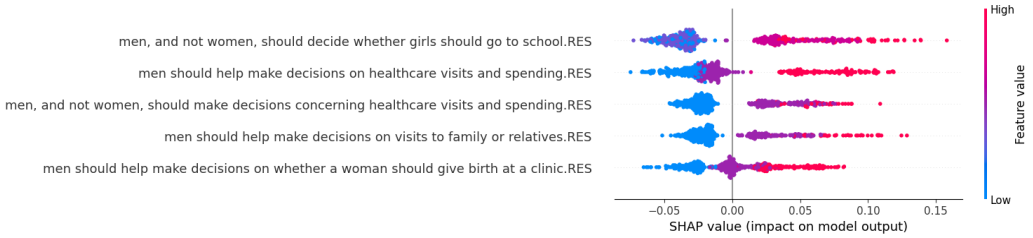

Figure 3.10 Top 5 SHAP values for k-means cluster 2

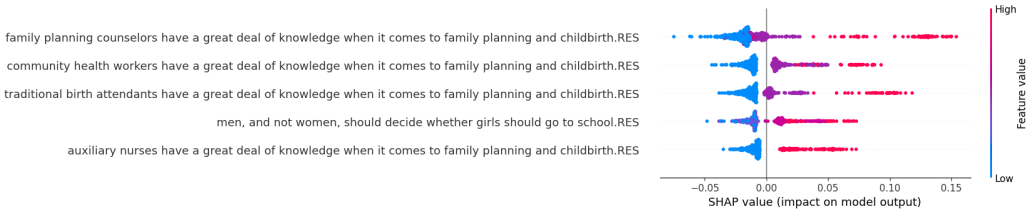

Figure 3.11 Top 5 SHAP values for k-means cluster 3

3.5.3 HDBSCAN & UMAP

3.5.3.1 Statistical Differences Between Clusters

Demographics Demographic characteristics for which at least one between-cluster difference is statistically significant ($p < 0.1$) appear in Figure 3.12. HDBSCAN identifies a signal related to respondent occupation and number of children under 18 years of age or number of children in school. Cluster 4 is associated with a larger proportion reporting agricultural occupations, whereas cluster 3 has a larger representation of teachers. Cluster 2 exhibits a larger proportion of respondents reporting trade as occupation and primary education as the highest level of attainment. There is also some indication of cluster 2 respondents parenting younger children.

Value Hypotheses Per-cluster mean value resonances and conflicts showing at least one

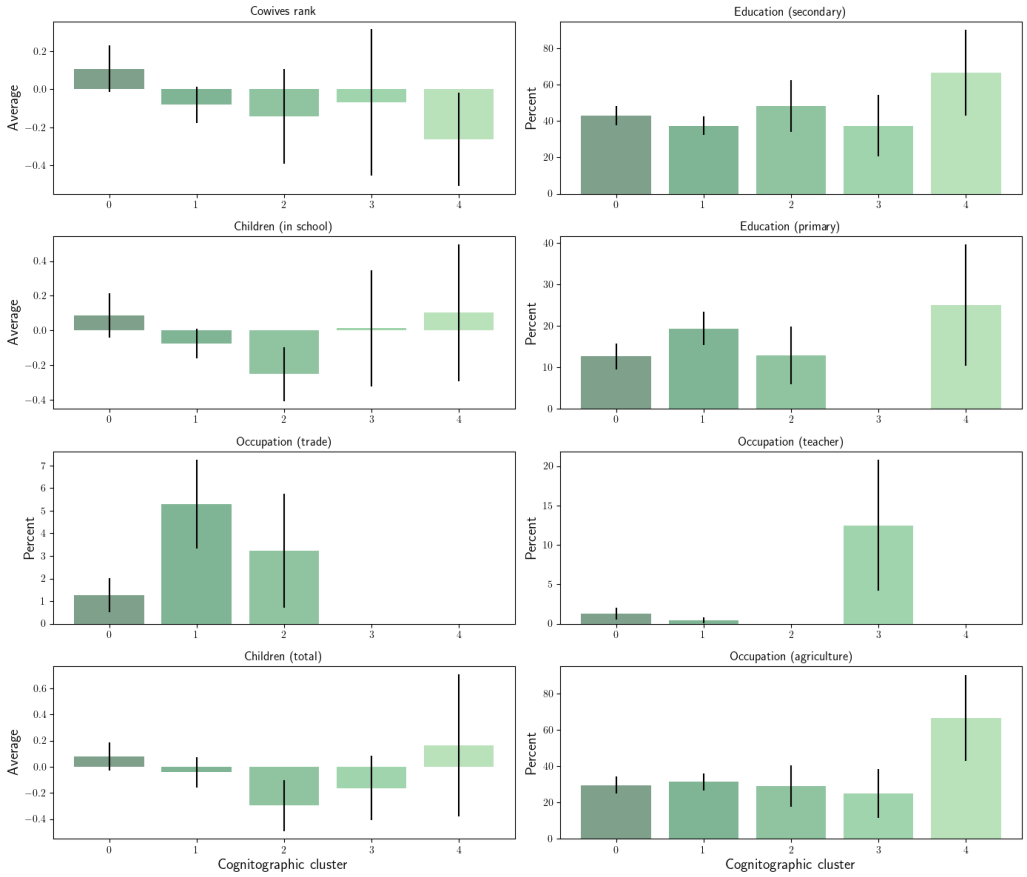

Figure 3.12 Significant between-cluster demographics for HDBSCAN

statistically significant difference between clusters ($p < 0.1$) appear in Figure 3.13. Clusters are ordered largest to smallest, left to right. Cluster 1 exhibits a stronger on average conflict with "the government should prohibit people from immigrating." Clusters 0 and 4 resonate with "it is not important to know about science in one's daily life." Cluster 3 indicates a tendency towards neutrality with the notable exception of conflict with "it is not important to know about science in one's daily life." Cluster 4 exhibits the highest degree of resonance-conflict heterogeneity.

3.5.3.2 Cluster Explainability

Figure 3.14 indicates cluster membership is broadly accounted for by conflict with the value hypotheses "the government should prohibit people from immigrating" and "it is not important to know about science in one's daily life." These accumulated coefficient mag-

Figure 3.13 Significant value hypotheses for HDBSCAN

nitudes eclipse all others. Among top represented value hypotheses, both resonance and conflict appear.

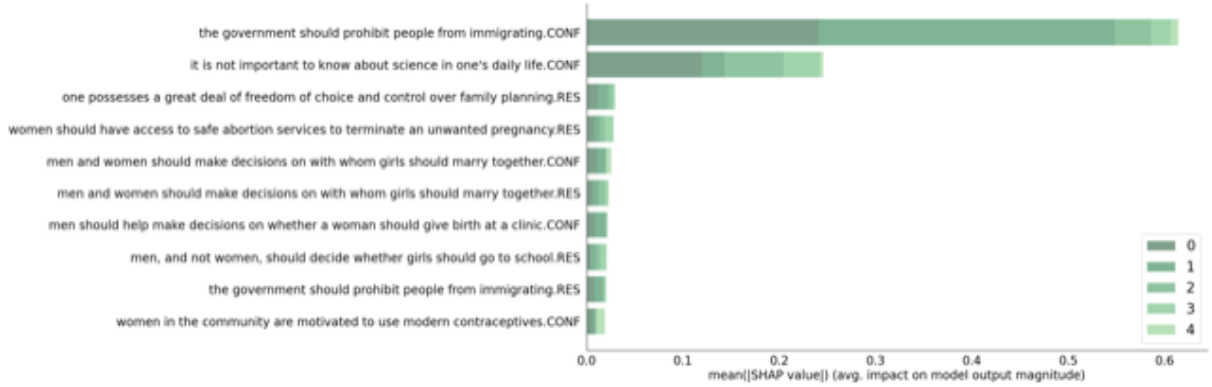

Figure 3.14 Ordered SHAP values for HDBSCAN

Clusters 0 and 1 differ by the SHAP value magnitude of conflict with "the government should prohibit people from immigrating," with cluster 1 showing larger values. This between-cluster pattern is repeated for conflict with "it is not important to know about sci-

ence in one's daily life" and resonance with "one possesses great deal of freedom of choice and control over family planning," and furthermore with each indicated value hypothesis. There is notable across-cluster consistency in the top value hypotheses identified by SHAP analysis.

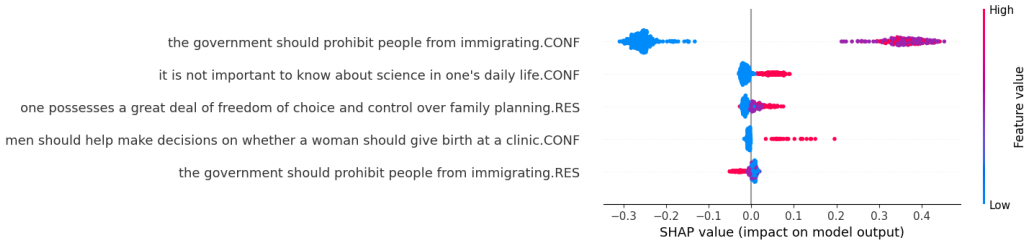

Figure 3.15 Top 5 SHAP values for HDBSCAN cluster 0

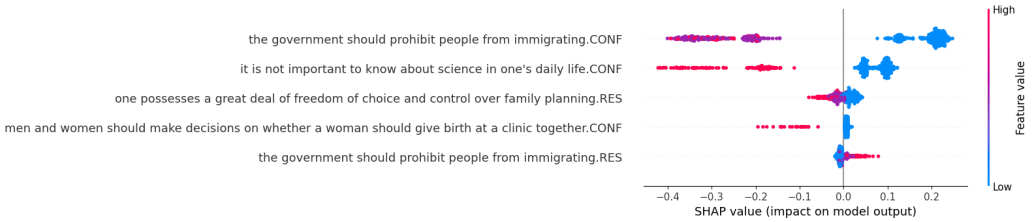

Figure 3.16 Top 5 SHAP values for HDBSCAN cluster 1

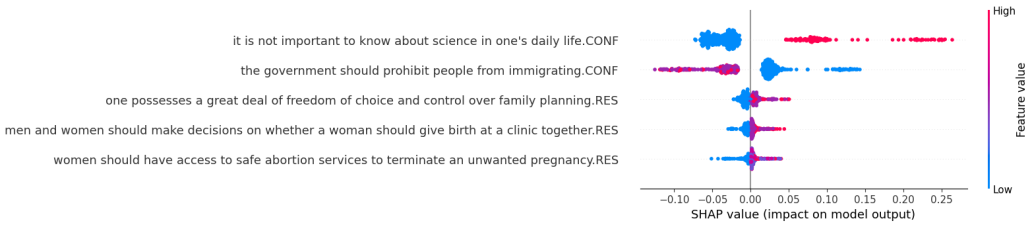

Figure 3.17 Top 5 SHAP values for HDBSCAN cluster 2

SHAP values for institution-, knowledge-, and gender-related value hypotheses are larger overall. Clusters 0 and 1 present complementary SHAP values, where larger magnitudes are associated with membership in 1. Clusters 2 and 3 show large SHAP values for greater conflict with "it is not important to know about science in one's daily life," but complement each other on gender role hypotheses. Cluster 4 SHAP values indicate the importance

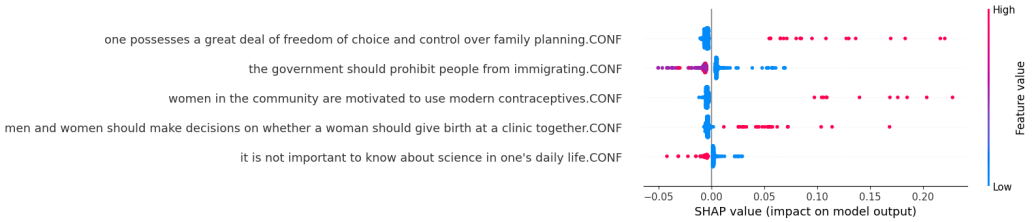

Figure 3.18 Top 5 SHAP values for HDBSCAN cluster 3

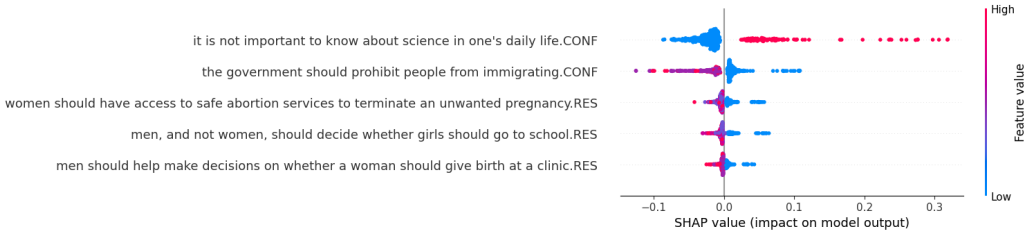

Figure 3.19 Top 5 SHAP values for HDBSCAN cluster 4

of value hypothesis conflicts to cluster membership. This reiterates statistical results of demographics.

3.5.4 Optimal Transport & Diffusion Maps

3.5.4.1 Overview

To identify value hypotheses for which diversity of value resonance is greatest, we combine all participant value resonances across each value hypothesis, compute the entropy of each, and represent in Figure 3.20 (note: density plots utilize Gaussian kernels and are intended as qualitative visual representations rather than quantitative density estimates). Higher resonance entropy at the upper tail of this distribution corresponds to greater diversity and suggests more informative value hypotheses. To inform selection of the bandwidth

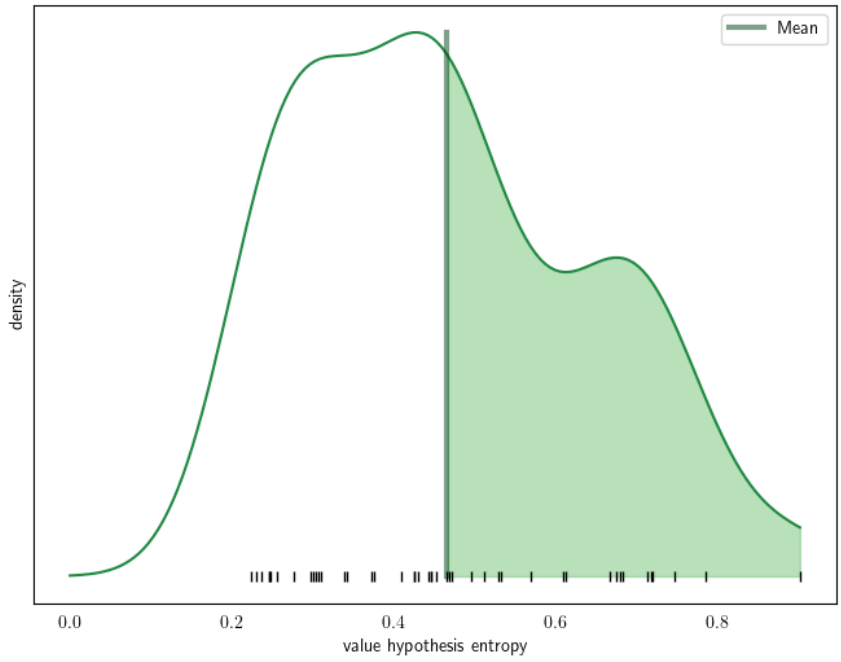

Figure 3.20 Value hypothesis entropy density

parameter, ϵ, the distribution of weighted OT distances between each pair of resonance densities is considered. ϵ should be large enough so that the underlying data graph remains connected but not so large that important information is lost. Hence we explore a range that lies within the highest density region of OT distances. Then the bottom 10 eigenvalues of $\mathbf{I} - \mathbf{P}$ and their differences are plotted as a function of ϵ in Figures 3.21 and 3.22.

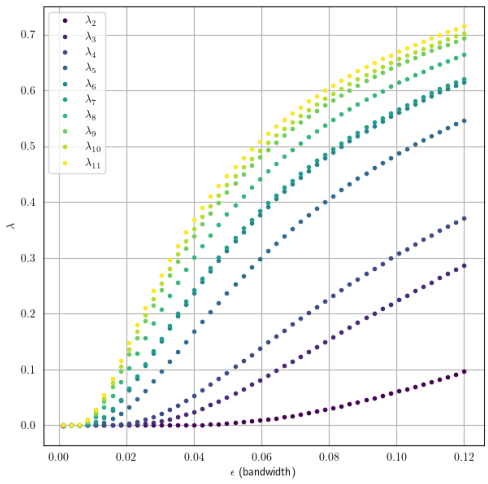

Figure 3.21 Multiscale eigenvalues; darker colors correspond to lower frequency eigenvectors

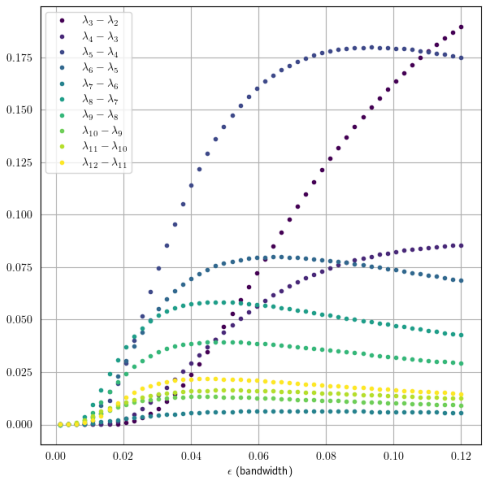

Figure 3.22 Multiscale eigenvalue differences

The largest gaps between adjacent eigenvalues for which all smaller eigenvalues are close to zero occur between 4 and 5 as well as 7 and 8. These yield cluster number estimates of $\hat{K} = 4$ and $\hat{K} = 7$; we find that these are robust to a large range of low-pass filtering as parameterized by t, the number of times $\mathbf{P}$ is powered to smooth. A value of $t = 256$ is chosen as it falls on the lower end of this stability range, thus preserving more detail. $\epsilon_4 = 0.09$ and $\epsilon_7 = 0.045$ are then identified at the corresponding peaks of the eigenvalue differences.

3.5.4.2 Statistical Differences Between Clusters

Demographics Figure 3.23 displays demographic characteristics for which at least one between-cluster difference is statistically significant for the 4-cluster outcome ($p < 0.1$). Clusters 0 and 1 can be differentiated on the basis of sex. Cluster 0 has an approximate 60% female representation and may be differentiated from clusters 1 and 2 by a greater prevalence of members reporting no occupation (20% versus 9% and 10%, respectively).

Significant demographic features for the 7-cluster outcome appear in Figure 3.24. We see more culturally indicated features emerging with this greater granularity. Arabic education

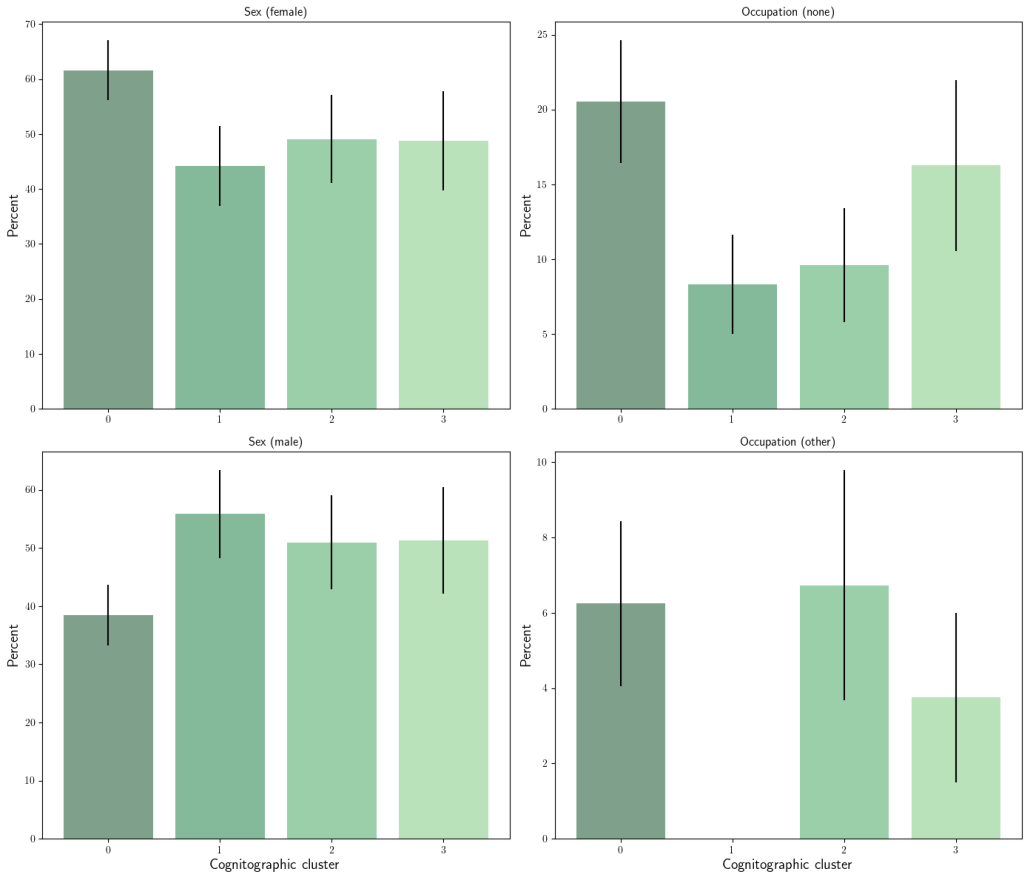

Figure 3.23 Significant between-cluster demographics for diffusion maps 4

accounts for a relatively large proportion of cluster 1 and is notable given the relative size of the cluster and relatively small preponderance of Arabic education in the data. This cluster may be further differentiated from cluster 5 on the basis of number of children either living or deceased, and from cluster 3 by a larger proportion reporting Islam for religion. Cluster 3 demographics suggest a more rural population with a larger proportion identifying as Christian and reporting more secondary education as compared with larger clusters. There is also a small trend towards younger heads of household and fewer young children when moving from cluster 1 to smaller clusters. Cluster 5 participants report slightly lower income than 4.

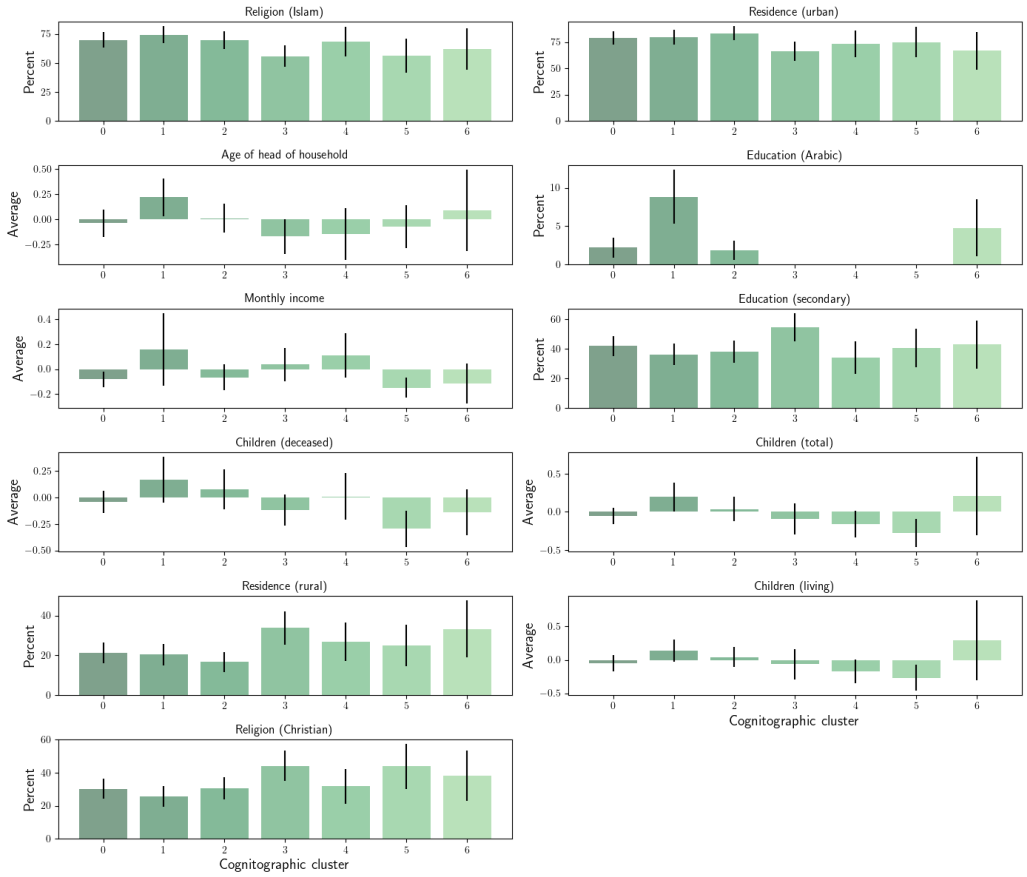

Figure 3.24 Significant between-cluster demographics for diffusion maps 7

Value Hypotheses Displayed in Figure 3.25 and Figure 3.26 are per-cluster mean value resonances and conflicts for which at least one between-cluster difference is statistically significant for the 4 and 7 hypothesized clusters ($p < 0.1$), respectively.

Cluster 1 exhibits larger resonance magnitudes with gender-related value hypotheses. This correlates with the larger male representation identified through demographics. Clusters 0, 1, and 2 exhibit similar patterns of value resonance, with cluster 0 exhibiting the smallest coefficients. Clusters 1 and 3 show the largest conflict with "the government should prohibit people from immigrating." Cluster 3 has a more uniform distribution of coefficient magnitudes with the exception of a large conflict with "it is not important to know about science in one's daily life."

For the 7-cluster outcome, cluster 5 displays a visually striking resonance with "it is

Figure 3.25 Significant value hypotheses for diffusion maps 4

not important to know about science in one's daily life." This is in contrast to cluster 3, which shows a relatively strong conflict here. This cluster also exhibits a larger proportion of Christians. Cluster 5 furthermore exhibits a moderate conflict with "the government should prohibit people from immigrating." Cluster 3 presents no conflict with "locally, one can generally feel secure," which is in contrast to the majority of clusters. Cluster 6 uniquely and moderately conflicts with "the health services, clinic, or nearest hospital can be relied upon to deliver" and slightly with "the midwives at the clinic have a great deal of knowledge when it comes to family planning and childbirth." Cluster 4 uniquely shows no resonance with "men should help make decisions on major household purchases," although those resonance magnitudes are generally small across clusters. Cluster 1 exhibits overall smaller magnitude coefficients. It also resonates least with many gender-related questions. Cluster 6 presents the most resonance-conflict heterogeneity.

Figure 3.26 Significant value hypotheses for diffusion maps 7

3.5.4.3 Cluster Explainability

Diffusion Maps 4 Aggregate SHAP value magnitudes for the 4-cluster outcome (Figure 3.27) are largest for resonances with the exception of conflicts with "the government should prohibit people from immigrating" and "it is not important to know about science in one's daily life." Gender-related value hypotheses are also prominent. However, we observe dominant representation by institution-related values.

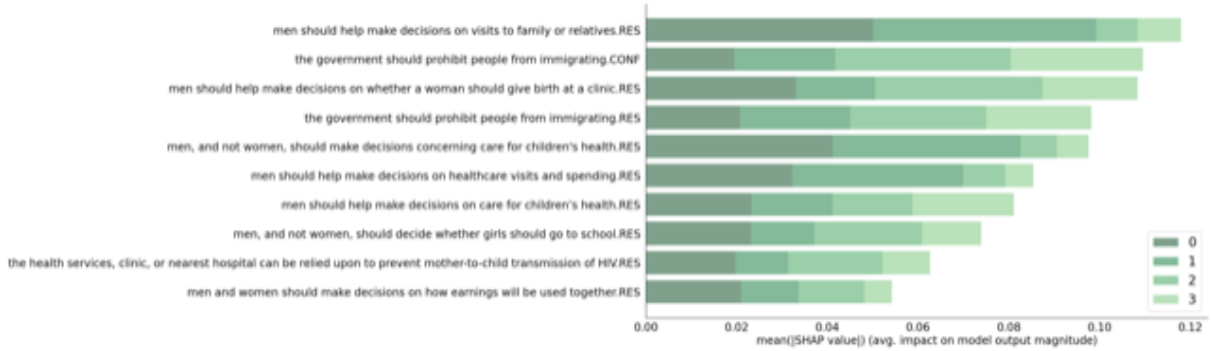

Figure 3.27 Ordered SHAP values for diffusion maps 4

SHAP values associated with clusters 0 and 1 fall along the lines of gender-related values hypotheses, specifically resonances regarding decision making, with smaller magnitudes associated with membership in cluster 0. This correlates with respondent sex. Cluster 2 SHAP values indicate cluster membership is associated with lower levels of resonance and conflict with "the government should prohibit people from immigrating" and more resonance with gender-related value hypotheses. This is in contrast to other clusters, particularly cluster 3.

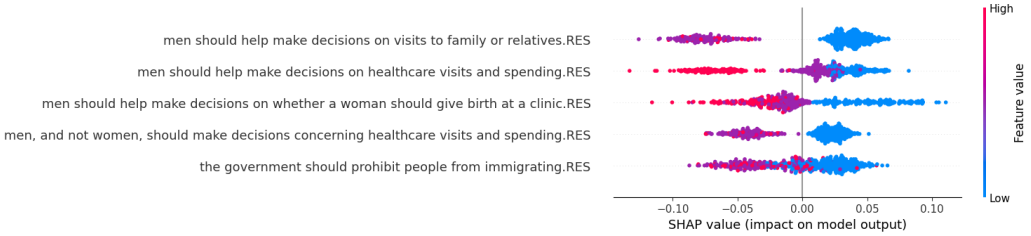

Figure 3.28 Top 5 SHAP values for diffusion maps 4, cluster 0

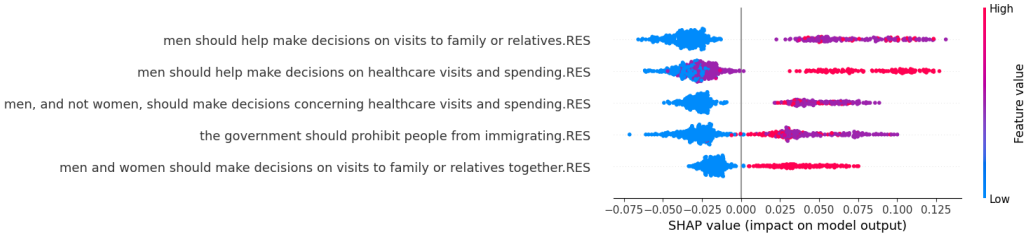

Figure 3.29 Top 5 SHAP values for diffusion maps 4, cluster 1

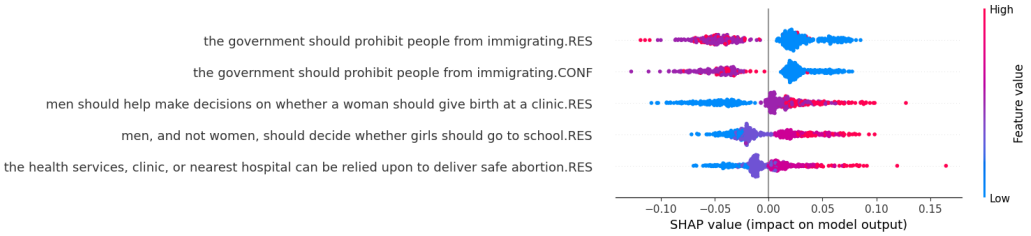

Figure 3.30 Top 5 SHAP values for diffusion maps 4, cluster 2

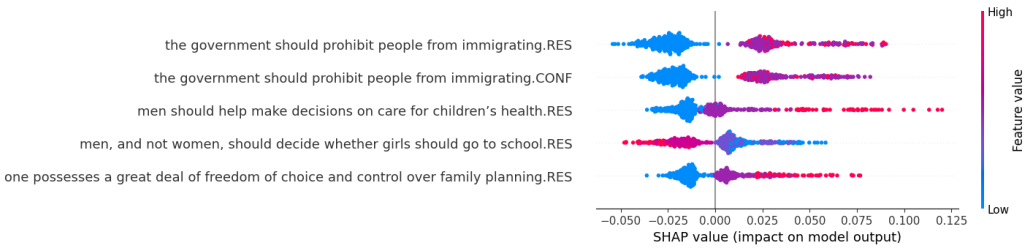

Figure 3.31 Top 5 SHAP values for diffusion maps 4, cluster 3

Diffusion Maps 7 There is a large representation by resonance with values hypotheses in the aggregate SHAP value magnitudes (Figure 3.32). However, we also see conflict with "the government should prohibit people from immigrating" and "it is not important to know about science in one's daily life" as well as "the health services, clinic, or nearest hospital can be relied upon to deliver safe abortion."

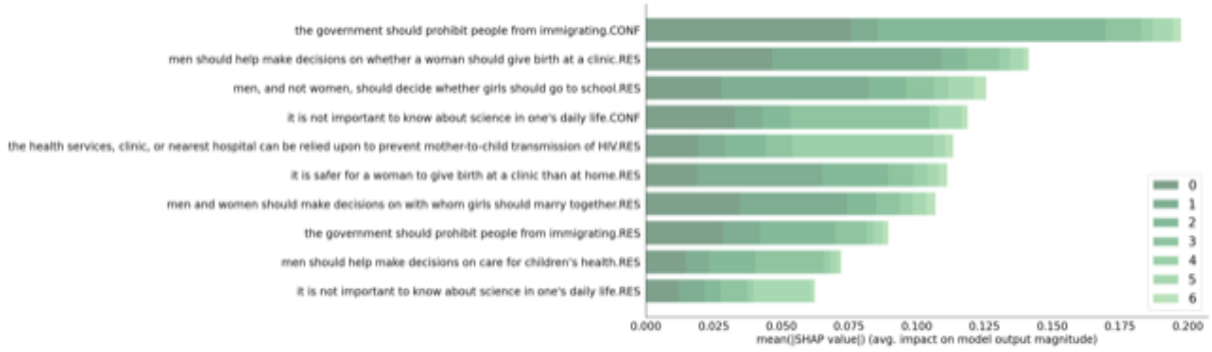

Figure 3.32 Ordered SHAP values for diffusion maps 7

Cluster 0 SHAP values indicate membership is associated with greater magnitude of gender-related decision making value hypotheses and lesser magnitudes of government-related. This is in contrast with cluster 1, where membership is associated with lower coefficient magnitudes of resonances. Cluster 2 SHAP values emphasize institutions and safety, whereas cluster 3 emphasizes knowledge, actors, and institutions. We observe that clusters 4 and 6 have large SHAP values associated with larger magnitude coefficients on single value hypotheses.

Cluster 0 SHAP values indicate membership is associated with greater magnitude of gender-related decision making value hypotheses and lesser magnitudes of government-related.

This is in contrast with cluster 1, where membership is associated with lower coefficient magnitudes of resonances. Cluster 2 SHAP values emphasize institutions and safety, whereas cluster 3 emphasizes knowledge, actors, and institutions. We observe that clusters 4 and 6 have large SHAP values associated with larger magnitude coefficients on single value hypotheses.

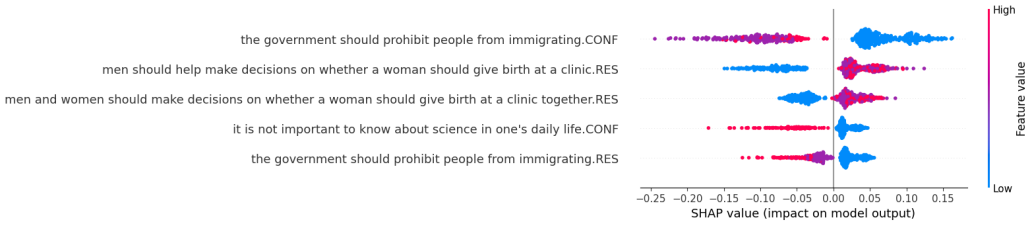

Figure 3.33 Top 5 SHAP values for diffusion maps 7, cluster 0

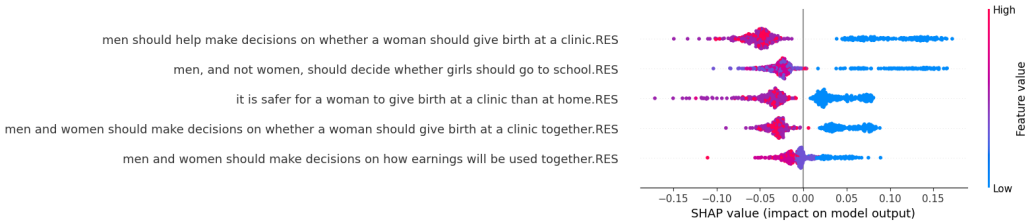

Figure 3.34 Top 5 SHAP values for diffusion maps 7, cluster 1

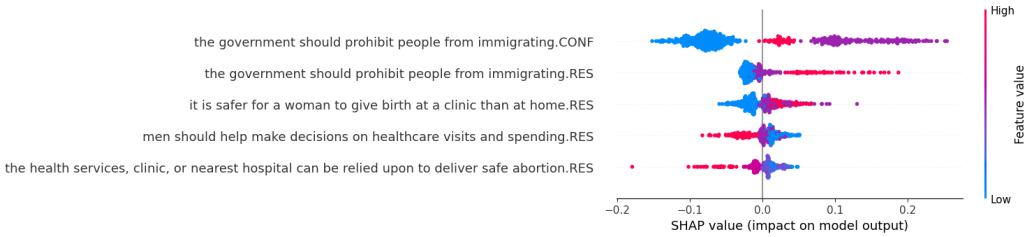

Figure 3.35 Top 5 SHAP values for diffusion maps 7, cluster 2

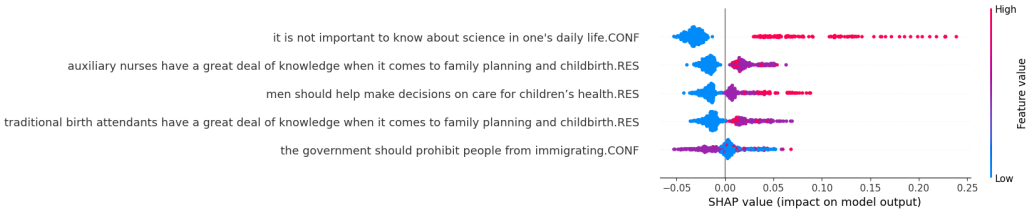

Figure 3.36 Top 5 SHAP values for diffusion maps 7, cluster 3

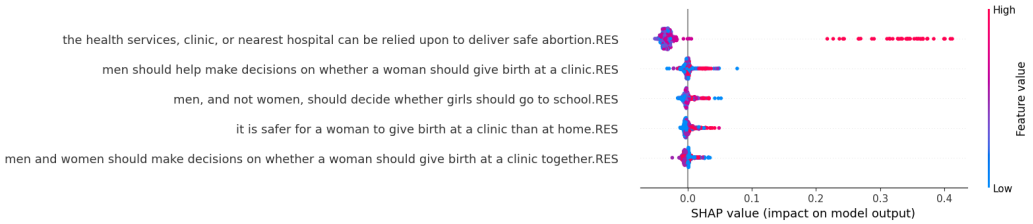

Figure 3.37 Top 5 SHAP values for diffusion maps 7, cluster 4

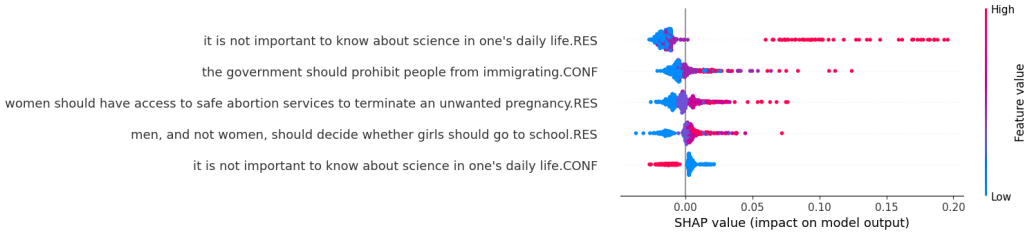

Figure 3.38 Top 5 SHAP values for diffusion maps 7, cluster 5

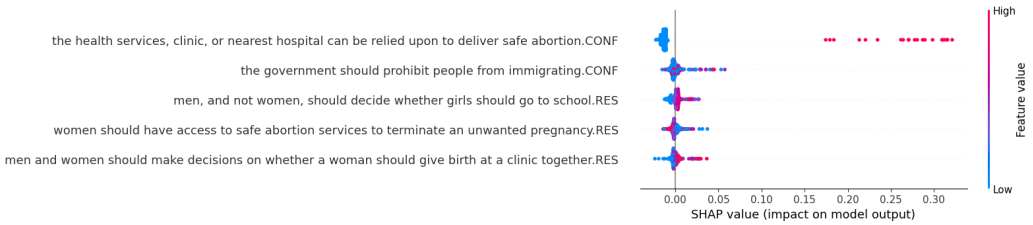

Figure 3.39 Top 5 SHAP values for diffusion maps 7, cluster 6

3.5.5 Comparative Analysis

Each of the three methods identifies a similar number of clusters, the size distributions of which appear in Figure 3.40. Cluster size distributions of diffusion maps 4 and k-means are remarkably similar. HDBSCAN partitions the respondent sample into two similarly sized large clusters while identifying 3 smaller groups. Diffusion maps 7 generates more uniformly sized clusters.

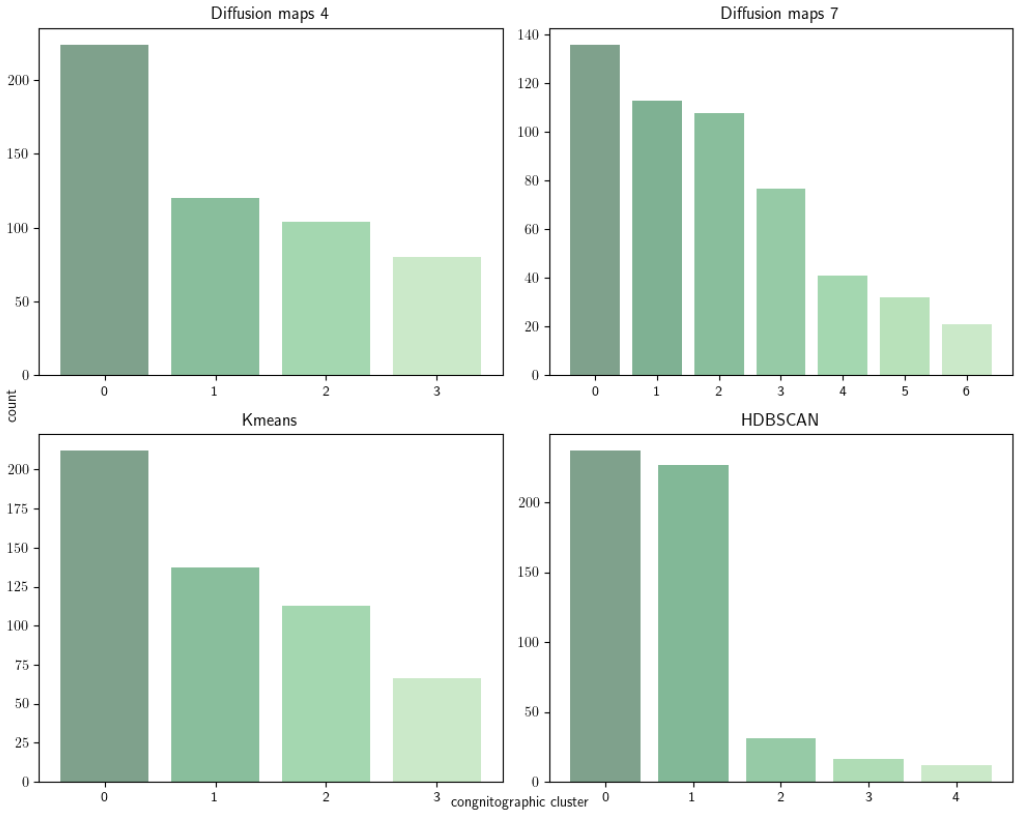

Figure 3.40 Cluster size distributions for each method

The total number of pairwise statistically significant cluster mean value hypothesis resonances and conflicts identified by each approach is given in Figure 3.41.

3.5.5.1 Jaccard Score

Jaccard similarity scores for each pair of methods are shown in Figures 3.42, 3.43, and 3.44. Sparser arrays with greater Jaccard scores signify greater agreement between methods. We observe the most sparsity in heatmaps between HDBSCAN and other methods. This

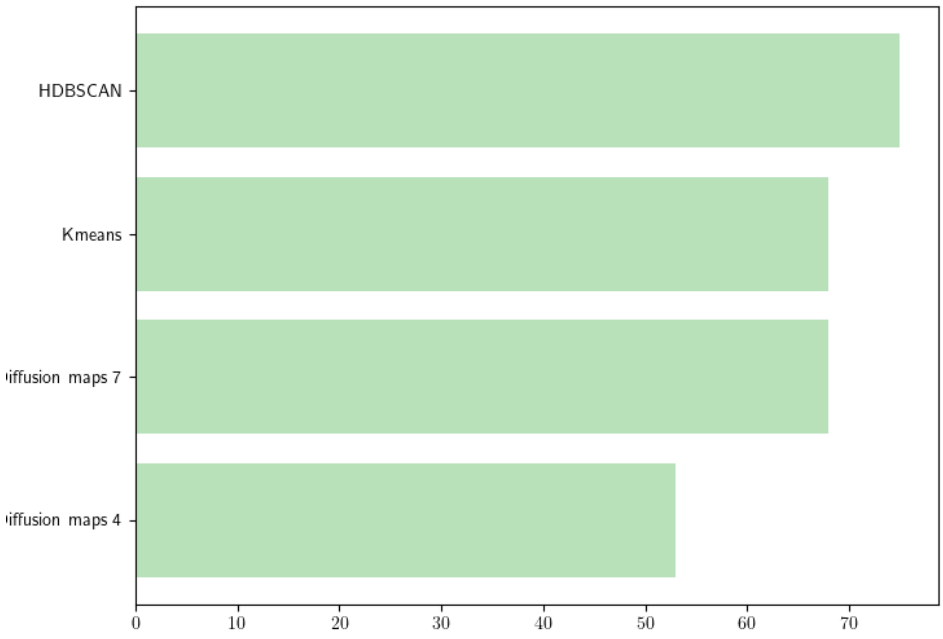

Figure 3.41 Number of significant pairwise cluster value hypotheses

may be partially attributed to the two large, primary clusters identified by HDBSCAN. The largest Jaccard scores appear between diffusion maps 7 and HDBSCAN, diffusion maps 7 and k-means, and diffusion maps 4 and k-means. Diffusion maps 7 and HDBSCAN are the most globally similar based upon sparsity and Jaccard score concentration.

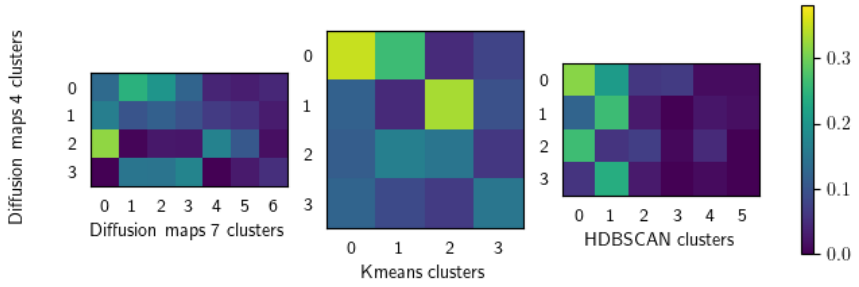

Figure 3.42 Jaccard similarity between diffusion maps 4 and other methods

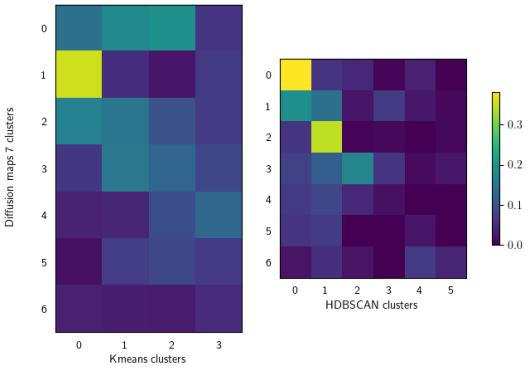

Figure 3.43 Jaccard similarity between diffusion maps 7 and other methods

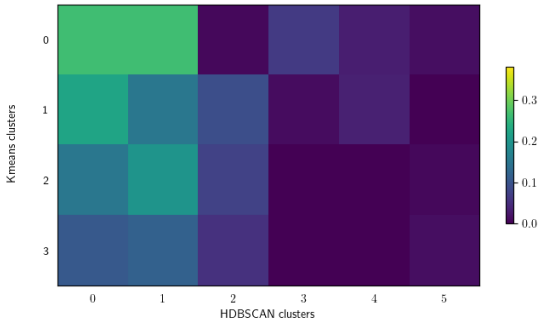

Figure 3.44 Jaccard similarity between k-means and HDBSCAN

3.5.5.2 Rank Biased Overlap

Figure 3.45 presents the pairwise Rank Biased Overlap (RBO) similarity scores for each pair of clustering algorithms. HDBSCAN appears in the most similar methodology pairing alongside diffusion maps 7, and in the least similar when compared to both k-means and diffusion maps 4. The 7 cluster diffusion maps algorithm showed the greatest levels of global similarity to the other methods.

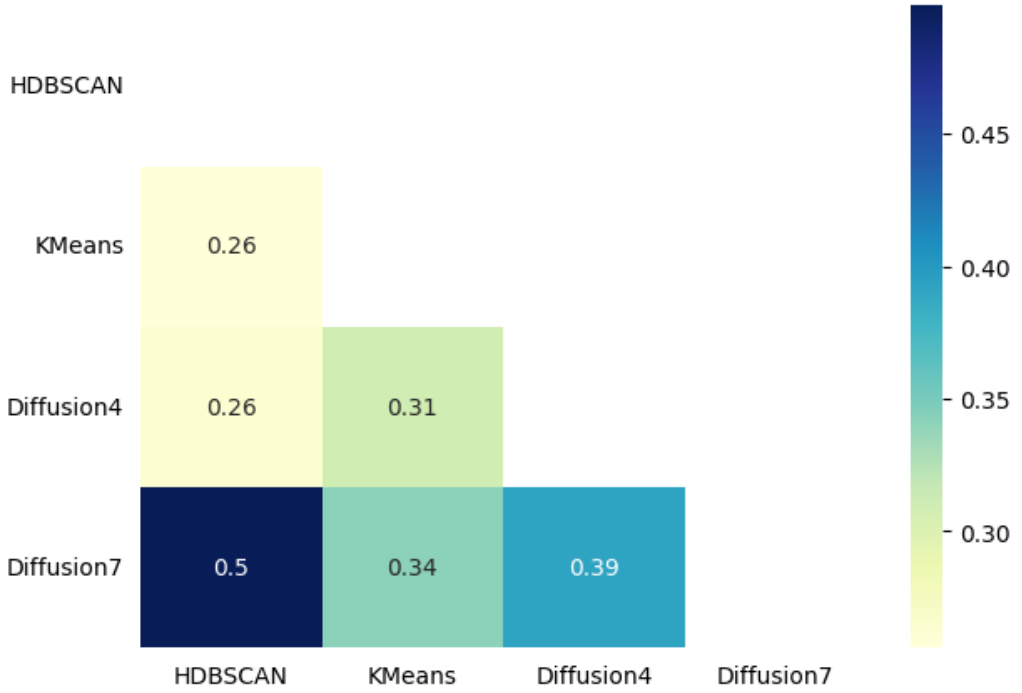

Figure 3.45 Rank Biased Overlap (RBO) similarity scores for each pair

3.6 Discussion

We observe the emergence of cognitographic clusters that express resonance and conflict with values related to gender, decision making, institutions, social actors, and knowledge. The methods can be arranged along a spectrum, with k-means and HDBSCAN falling at the extremes, corresponding to greater emphasis on similarities and differences, respectively. Diffusion maps may be viewed as a qualitative interpolation between them. This is supported not only by the demographic analyses, RBO and Jaccard scores, but also visual inspection of SHAP plots, cluster counts, and cluster size distributions.

At coarser scales, k-means finds clusters that are more readily differentiated by demographics and value hypotheses. The method also appears to be biased towards discriminating largely on the most frequently embedded values and teasing out the differences in the degree to which these commonly espoused values are present across clusters. This would also explain why the values with the highest SHAP scores are exclusively resonances, as the data set as a whole is heavily biased towards resonance, as opposed to conflicts. Because a large subset of

the most common values relate to gender roles, it is perhaps not surprising that differences in alignment with such values would be most pronounced along the gender dimension, as demonstrated by the demographic breakdown of the k-means clusters.

At the greater granularities of diffusion maps 7 and HDBSCAN, we observe the emergence of more culturally indicated features. The smaller clusters produced by these methods exhibit large SHAP values for a few value hypotheses, suggesting greater heterogeneity that coalesces around smaller numbers of beliefs. SHAP analysis of the diffusion maps methods as well as HDBSCAN places a large weight on "the government should prohibit people from immigrating." This is in contrast to the k-means SHAP analysis where this hypothesis does not appear. The HDBSCAN method apparently zeroes in on values with the largest variance of resonance/conflict among the population (i.e. the most polarizing values). This would align with "the government should prohibit people from immigrating," and "it is not important to know about science in one's daily life" being the two values assigned an overwhelming degree of importance in determining cluster membership compared to the remaining set of values. Because the maximally polarized values may not necessarily be well-represented throughout the population (as is the case here), the resulting clusters may be skewed in size, and the demographic differences may be more subtle and less interpretable. Finally, the diffusion maps method strikes a balance between the others by bestowing roughly equal importance to values both commonly espoused and polarizing, as one may observe from the cluster explainability diagrams. There is notable agreement between SHAP and statistical analyses across the methods.

We remark that, although intended as soft validations, care is required for interpretation of statistical analyses of demographic features and value resonance magnitudes when considered in aggregate. The emergence of patterns among cluster demographics suggests that they may be used as proxies for cluster identification; however the main interest is values-based clustering. By contrast, the population could be clustered solely by demographics, but this would miss heterogeneous values among demographic groups. Remarkably, while analyses

of demographics are based upon 90% confidence intervals (CI), with an 85% CI, all methods identify a cluster that can be differentiated from others based upon Arabic education. Diffusion maps 7 captures this at the more stringent level.

These results indicate that each method is potentially well-suited for a different use case. If the objective is to tease out differences in how different groups relate to values that are among the most salient across the whole population then k-means would be an appropriate choice. If one, however, wishes to elicit how the population breaks down with respect to highly polarizing (albeit perhaps fringe or niche) values, HDBSCAN would work well. To capture a mixture of both phenomena, one could choose diffusion maps applied to optimal transport distances.

3.7 Conclusion

Inferences about values-based subpopulations are method-dependent. This should be interpreted not as a bug but as a feature. Consider a scenario in which 3 social scientists (say, from different fields) are asked to manually cluster respondents based upon values expressed in narratives. Surely one would observe some inter-researcher agreement. Yet without further constraint, each would attend to different aspects of the data and draw conclusions thereon. Furthermore, in light of intersectionality[7], participants would likely self-organize into clusters in context-specific ways. Perhaps this would occur, for instance, along the lines of sex or gender if reproductive rights were invoked in a cultural context favoring self-direction. Yet in a context favoring tradition, emergent structures might coalesce around religious and political affiliation or some other points of consensus and divergence.

Translating these insights into wisdom of crowd estimates is delicate. Is a greater emphasis on value system consensus or divergence relevant to the task at hand? Relevant to whom? And to what ends? While we cannot answer these questions generally, we can consider them relative to sustainable development operations. In these contexts, one could imagine operationalizing an ensemble of models formed at varying granularities, perhaps utilizing multiple

[7] *Intersectionality* refers to the interconnected nature of social categorizations such as race, gender, and religion as they apply to a particular individual or group

methods. Suppose the task is predictive modeling of hyperlocal MCH utilization on one-month time lags. Interpreted within the framework of deep uncertainty Walker et al. (2012), one would hope for an ensemble that yields lower variance estimates or, perhaps, a family of competing estimates that are associated with different subpopulations. Accounting for these heterogeneous sources of uncertainty should simultaneously enable practitioners to explore richer intervention scenarios and achieve more locally equitable, desirable, and meaningful outcomes.

3.8 Acknowledgements

This material is based upon work supported by the Army Contracting Command, DARPA, and ARO under Contract No. W911NF-21-C-0007. The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

BIBLIOGRAPHY

Allaoui, M., Kherfi, M. L., and Cheriet, A. (2020). Considerably improving clustering algorithms using umap dimensionality reduction technique: a comparative study. In *Image and Signal Processing: 9th International Conference, ICISP 2020, Marrakesh, Morocco, June 4–6, 2020, Proceedings 9*, pages 317–325. Springer.

Aminpour, P., Gray, S. A., Jetter, A. J., Introne, J. E., Singer, A., and Arlinghaus, R. (2020). Wisdom of stakeholder crowds in complex social–ecological systems. *Nature Sustainability*, 3(3):191–199.

Aminpour, P., Gray, S. A., Singer, A., Scyphers, S. B., Jetter, A. J., Jordan, R., Murphy Jr, R., and Grabowski, J. H. (2021). The diversity bonus in pooling local knowledge about complex problems. *Proceedings of the National Academy of Sciences*, 118(5):e2016887118.

Asyaky, M. S. and Mandala, R. (2021). Improving the performance of hdbscan on short text clustering by using word embedding and umap. In *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–6. IEEE.

Benkler, N., Friedman, S., Schmer-Galunder, S., Mosaphir, D., Sarathy, V., Kantharaju, P., McLure, M. D., and Goldman, R. P. (2022). Cultural value resonance in folktales: A transformer-based analysis with the world value corpus. In *Social, Cultural, and Behavioral Modeling: 15th International Conference, SBP-BRiMS 2022, Pittsburgh, PA, USA, September 20–23, 2022, Proceedings*, pages 209–218. Springer.

Bernau, C., Riester, M., Boulesteix, A.-L., Parmigiani, G., Huttenhower, C., Waldron, L., and Trippa, L. (2014). Cross-study validation for the assessment of prediction algorithms. *Bioinformatics*, 30(12):i105–i112.

Blanco-Portals, J., Peiró, F., and Estradé, S. (2022). Strategies for eels data analysis. introducing umap and hdbscan for dimensionality reduction and clustering. *Microscopy and Microanalysis*, 28(1):109–122.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.

Coifman, R. R. and Lafon, S. (2006). Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30.

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., and Sen, P. (2020). A survey of the state of explainable ai for natural language processing. *arXiv preprint arXiv:2010.00711*.

De Plaen, H., Fanuel, M., and Suykens, J. A. (2020). Wasserstein exponential kernels. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766.

Galton, F. (1907). Vox populi.

Graeber, D. (2001). *Toward an anthropological theory of value: The false coin of our own dreams*. Springer.

Gray, S. A., Gray, S., De Kok, J. L., Helfgott, A. E., O'Dwyer, B., Jordan, R., and Nyaki, A. (2015). Using fuzzy cognitive mapping as a participatory approach to analyze change, preferred states, and perceived resilience of social-ecological systems. *Ecology and Society*, 20(2).

Green, L. W. (1974). Toward cost-benefit evaluations of health education: some concepts, methods, and examples. *Health Education Monographs*, 2(1_suppl):34–64.

Green, L. W. and Kreuter, M. W. (1991). *Health education planning*. Mayfield Pub. Co.

GSMoH, G. S. M. o. H. (2010). Gombe state government strategic health development plan.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Inglehart, R. (2020). The inglehart-welzel world cultural map–world values survey 7 [provisional version].

Inglehart, R., Basanez, M., Diez-Medrano, J., Halman, L., and Luijkx, R. (2000). World values surveys and european values surveys, 1981-1984, 1990-1993, and 1995-1997. *Ann Arbor-Michigan, Institute for Social Research, ICPSR version*.

Inglehart, R. and Welzel, C. (2010). The wvs cultural map of the world. *World Values Survey*.

Kosko, B. (1986). Fuzzy cognitive maps. *International journal of man-machine studies*, 24(1):65–75.

Levin, P. S., Gray, S. A., Möllmann, C., and Stier, A. C. (2021). Perception and conflict in conservation: The rashomon effect. *BioScience*, 71(1):64–72.

Little, A. V., Maggioni, M., and Murphy, J. M. (2020). Path-based spectral clustering: Guarantees, robustness to outliers, and fast algorithms. *Journal of machine learning research*, 21.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Madigan, D., Stang, P. E., Berlin, J. A., Schuemie, M., Overhage, J. M., Suchard, M. A., Dumouchel, B., Hartzema, A. G., and Ryan, P. B. (2014). A systematic statistical approach to evaluating evidence from observational studies. *Annual Review of Statistics and Its Application*, 1:11–39.

Maxmen, A. (2015). How the fight against ebola tested a culture's traditions. *National Geographic*, 30.

McInnes, L., Healy, J., and Astels, S. (2017). hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Moon, K. R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., Yim, K., Elzen, A. v. d., Hirn, M. J., Coifman, R. R., et al. (2019). Visualizing structure and transitions in high-dimensional biological data. *Nature biotechnology*, 37(12):1482–1492.

Papageorgiou, E. I. and Salmeron, J. L. (2012). A review of fuzzy cognitive maps research during the last decade. *IEEE transactions on fuzzy systems*, 21(1):66–79.

Pealat, C., Bouleux, G., and Cheutet, V. (2021). Improved time-series clustering with umap dimension reduction method. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5658–5665. IEEE.

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Rittel, H. W. and Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy sciences*, 4(2):155–169.

Schwartz, S. (2006). A theory of cultural value orientations: Explication and applications. *Comparative sociology*, 5(2-3):137–182.

Sinai, I., Anyanti, J., Khan, M., Daroda, R., and Oguntunde, O. (2017). Demand for women's health services in northern nigeria: a review of the literature. *African Journal of Reproductive Health*, 21(2):96–108.

Smola, A. J. and Schölkopf, B. (1998). *Learning with kernels*, volume 4. Citeseer.

Team, I. (2023). Infer and uk professional head of intelligence assessment launch collaboration. https://www.infer-pub.com/the-pub/uk-collaboration/.

Uzochukwu, B. S. (2017). Primary health care systems (primasys): case study from nigeria. *Geneva: World Health Organization*.

Voinov, A., Jenni, K., Gray, S., Kolagani, N., Glynn, P. D., Bommel, P., Prell, C., Zellner, M., Paolisso, M., Jordan, R., et al. (2018). Tools and methods in participatory modeling: Selecting the right tool for the job. *Environmental Modelling & Software*, 109:232–255.

Walker, W. E., Lempert, R. J., and Kwakkel, J. H. (2012). Deep uncertainty. *Delft University of Technology*, 1(2).

Webber, W., Moffat, A., and Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38.

WHO, U. (2023). Unfpa, world bank group and the united nations population division. trends in maternal mortality: 2000 to 2020: estimates by who, unicef.

Yi, S. K. M., Steyvers, M., Lee, M. D., and Dry, M. J. (2012). The wisdom of the crowd in combinatorial problems. *Cognitive science*, 36(3):452–470.

Özesmi, U. and Özesmi, S. L. (2004). Ecological models based on people's knowledge: a multi-step fuzzy cognitive mapping approach. *Ecological Modelling*, 176(1):43–64.

APPENDIX

VALUE HYPOTHESES & DEMOGRAPHICS

The full list of values considered by the recognizing value resonance (RVR) model follows.

1. family are important in life.
2. friends are important in life.
3. leisure time is important in life.
4. politics are important in life.
5. work is important in life.
6. religion is important in life.
7. it is important for children to have good manners.
8. it is important for children to be independent.
9. it is important for children to be hard workers.
10. it is important for children to have a sense of responsibility.
11. it is important for children to be imaginative.
12. it is important for children to be tolerant and respect others.
13. it is important for children to be thrifty in saving money and other economic pursuits.
14. determination and perseverance are important qualities in children.
15. it is important for children to possess religious faith.
16. it is important for children to be unselfish.
17. it is important for children to be obedient.
18. drug addicts make bad neighbors.
19. people of a different race make bad neighbors.
20. people with AIDS make bad neighbors.
21. immigrants and foreign workers make bad neighbors.
22. homosexuals make bad neighbors.
23. people who practice a different religion make bad neighbors.
24. heavy drinkers make bad neighbors.

25. unmarried couples living together make bad neighbors.
26. people who speak a different language make bad neighbors.
27. making one's parent's proud is of central importance in life.
28. preschool children suffer from having a working mother.
29. men make better political leaders than women do.
30. university is more important for a boy than it is for a girl.
31. men make better business executives than women do.
32. being a housewife is just as fulfilling as working.
33. under employment scarcity, men should have more right to a job than women.
34. under employment scarcity, employers should give priority to the nation's people over immigrants.
35. it is wrong if women have more income than their husbands.
36. homosexual couples make equally good parents as other couples.
37. it is an individual's duty towards society to have children.
38. it is a child's duty to take care of their ill parents.
39. people who don't work turn lazy.
40. work is one's duty towards society.
41. work should always come first even if it means less spare time.
42. the entire way society is organized must be radically changed by revolutionary action.
43. society must be gradually improved by reforms.
44. present society must be valiantly defended against all subversive forces.
45. in the future, less importance should be placed on work.
46. in the future, greater emphasis should be placed on technological advancement.
47. in the future, people should show greater respect for authority.
48. one generally feels happy.
49. one generally feels healthy.
50. there is a great deal of freedom of choice and control in one's life.

51. the current state of one's life is satisfactory.
52. the financial situation of one's household is satisfactory,
53. there has been a recurring lack of sufficient food to eat.
54. one cannot feel safe from criminal activity, even in one's own home.
55. proper medicine or medical treatment is frequently inaccessible.
56. a cash income is frequently unobtainable.
57. safe shelter is frequently inaccessible.
58. possessing a much higher standard of living than one's parents is a familiar personal experience.
59. most people can be trusted.
60. one's family is trustworthy.
61. one's neighborhood is trustworthy.
62. the people one knows personally are trustworthy.
63. one should trust people upon first meeting them.
64. people of a different religion are trustworthy.
65. people of a different nationality are trustworthy.
66. one can have confidence in churches.
67. one can have confidence in the armed forces.
68. one can have confidence in the press.
69. one can have confidence in television.
70. one can have confidence in labor unions.
71. one can have confidence in the police.
72. one can have confidence in the justice system.
73. one can have confidence in the government.
74. one can have confidence in political parties.
75. one can have confidence in parliament.
76. one can have confidence in the civil services.

77. one can have confidence in universities.
78. one can have confidence in elections.
79. one can have confidence in major companies.
80. one can have confidence in banks.
81. one can have confidence in the environmental protection movement.
82. one can have confidence in charitable and humanitarian organizations.
83. one can have confidence in major regional organizations.
84. it is more important for international organizations to be effective than it is for them to be democratic.
85. one can have confidence in their religious community.
86. incomes should be made more equal.
87. private ownership of business and industry should be increased.
88. the government should take more responsibility to ensure that everyone is provided for.
89. competition is good. it stimulates people to work hard and develop new ideas.
90. in the long run, hard work usually brings a better life.
91. protecting the environment should be given priority, even if it causes slower economic growth and some loss of jobs.
92. there exists a tremendous amount of corruption.
93. most state authorities are involved in corruption.
94. most business executives are involved in corruption.
95. most local authorities are involved in corruption.
96. most civil service providers are involved in corruption.
97. most journalists and media personnel are involved in corruption.
98. ordinary people have to pay bribes, give gifts, and do favors for local officials and service providers all the time.
99. on the whole, women are less corrupt than men.

100. there exists a high risk of being held accountable for being involved in bribery.
101. immigrants have a positive impact on national development.
102. immigrants help fill job vacancies.
103. immigrants strengthen cultural diversity.
104. immigrants increase crime rates.
105. immigration gives asylum to political refugees who are persecuted elsewhere.
106. immigration increases the risk of terrorism.
107. immigration helps poor people establish new lives.
108. immigrants increase unemployment rates.
109. immigration leads to social conflict.
110. regarding immigration, the government should let anyone come who wants to.
111. regarding immigration, the government should let people come as long as there are jobs available.
112. the government should place strict limits on the number of foreigners who can immigrate.
113. the government should prohibit people from immigrating.
114. locally, one can generally feel secure.
115. local robberies occur frequently.
116. locally, people frequently drink alcohol in the streets.
117. locally, the police or the military frequently interfere with people's private lives.
118. locally, racist behavior happens frequently.
119. locally, people sell drugs on the streets all the time.
120. local violence and street fights happen frequently.
121. locally, there is a high rate of sexual harassment.
122. It is unsafe to carry too much money on one's person.
123. It is unsafe to go out at night.
124. one should carry a weapon on their person for safety.

125. losing one's job or being unable to find employment is of genuine concern.
126. being unable to give one's children a good education is of genuine concern.
127. being personally victimized by crime is a familiar experience.
128. one's family being victimized by crime is a familiar experience.
129. national involvement in war is of genuine concern.
130. terrorist attacks are of genuine concern.
131. civil war is of genuine concern.
132. freedom is more important than equality.
133. freedom is more important than security.
134. one must be willing to fight for one's country.
135. science and technology make life healthier, easier, and more comfortable.
136. because of science and technology, there will be more opportunities for the next generation.
137. society depends too much on science and not enough on faith.
138. one of the bad effects of science is that it breaks down people's ideas of right and wrong.
139. it is not important to know about science in one's daily life.
140. the world is better off because of science and technology.
141. God is incredibly important in life.
142. God exists.
143. there exists life after death.
144. hell exists.
145. heaven exists.
146. whenever science and religion conflict, religion is always right.
147. the only acceptable religion is one's own religion.
148. frequently attending religious services is routine.
149. praying often is routine.

150. religiosity is an inseparable part of one's identity.
151. religion is more about following religious norms and ceremonies than it is about doing good to other people.
152. religion is more about making sense of life after death than it is about making sense of life in this world.
153. these days, one often has trouble deciding which moral rules are the right ones to follow.
154. claiming government benefits to which you are not entitled is perfectly justifiable.
155. avoiding a fare on public transportation is eminently justifiable.
156. stealing property is justifiable.
157. cheating on one's taxes is justifiable.
158. accepting a bribe in the course of one's duties is perfectly justifiable.
159. homosexuality is completely justifiable.
160. prostitution is justifiable.
161. abortion is easily justifiable.
162. divorce is justifiable.
163. sex before marriage is justifiable.
164. suicide is justifiable.
165. euthanasia is justifiable.
166. it is justifiable for a man to beat his wife.
167. it is justifiable for parents to beat their children.
168. violence against other people is justifiable.
169. terrorism as a political, ideological, or religious mean is justifiable.
170. having casual sex is justifiable.
171. political violence is justifiable.
172. the death penalty is justifiable.
173. the government has the right to surveil people in public areas.

174. the government has the right to monitor all emails and other information exchanged online.
175. the government has the right to collect information about its residents without their knowledge.
176. politics are of interest.
177. discussions of political matters with one's friends are standard.
178. the daily newspaper is a reliable source of information.
179. tv news is a reliable source of information.
180. news radio stations are reliable sources of information.
181. one's mobile phone is a reliable source of information.
182. one's email is a reliable source of information.
183. the internet is a reliable source of information.
184. social media is a reliable source of information.
185. conversations with friends or colleagues are reliable sources of information.
186. signing a petition as a political action is both viable and justifiable.
187. boycotts are a viable and justifiable political action.
188. peaceful political demonstrations are viable and justifiable political actions.
189. going on strike is a viable and justifiable political action.
190. donating to a campaign fund or group one believes in is a viable and justifiable political action.
191. contacting a government official for a cause one believes in is a viable and justifiable political action.
192. encouraging others to take action about political issues is a viable and justifiable political action.
193. encouraging people to vote during elections is a viable and justifiable political action.
194. searching for information about politics and political events online is a viable and justifiable political action.

195. signing an electronic petition is a viable and justifiable political action.
196. encouraging others to take political action using the internet is a viable and justifiable political action.
197. organizing political activities, events, and protests using the internet is a viable and justifiable political action.
198. voting in local elections is standard.
199. voting in national elections is standard.
200. votes are always counted fairly during national elections.
201. opposition candidates are always prevented from running during national elections.
202. TV news always favors the governing party during national elections.
203. voters are always offered bribes during national elections.
204. journalists always provide fair coverage of national elections.
205. election officials are always fair during national elections.
206. rich people always buy the national elections.
207. during national elections, voters are always threatened with violence at the polls.
208. voters are always offered a genuine choice during national elections.
209. women always have equal opportunities to run the office during national elections.
210. having honest elections is important.
211. people have a great deal of say in what the government does under the current political system.
212. politically, it is good to have a strong leader who does not have to bother with parliament and elections.
213. politically, it is good to have experts, not the government, make decisions according to what they think is best for the country.
214. politically, it is good to have the army rule.
215. it is good to have a democratic political system.
216. it is good to have a system governed by religious law in which there are no political

parties or elections.

217. one's political views should align with the political left.
218. one's political views should align with the political right.
219. governments taxing the rich and subsidizing the poor is an essential characteristic of democracy.
220. religious authorities interpreting the laws is an essential characteristic of democracy.
221. people choosing their leaders in free elections is an essential characteristic of democracy.
222. people receiving state aid for unemployment is an essential characteristic of democracy.
223. military leadership under governmental incompetence is an essential characteristic of democracy.
224. civil rights, designed to protect people's liberty against oppression are essential characteristics of democracy.
225. state-ensured income equality is an essential characteristic of democracy.
226. obedience to the governing body is an essential characteristic of democracy.
227. equal rights for women is an essential characteristic of democracy.
228. it is important to live in a democratically governed country.
229. one's nation is completely democratically governed.
230. the current national political system is functioning satisfactorily. nationally, there is a great deal of respect for individual human rights.
231. one possesses national pride.
232. one experiences fellowship with one's town, one's village, or one's city.
233. one experiences fellowship with one's district or one's region.
234. one experiences fellowship with one's country.
235. one experiences fellowship with one's continent.
236. one experiences fellowship with the world.
237. girls and women should themselves decide when, if, and with whom they should marry.
238. a girl should wait to marry until she has completed secondary school.

239. a boy should wait to marry until he has completed secondary school.
240. marrying girls young can help provide them security.
241. a boy should wait to have children until he is at least 18 years old.
242. it is important for a woman to have children as soon as possible after she has married.
243. it is important for a man to have children as soon as possible after he has married.
244. a woman should be in love with someone before having sex with that person.
245. a man should be in love with someone before having sex with that person.
246. women who carry condoms on them are easy.
247. men should be outraged if their wife or partner asks them to use a condom.
248. a real man produces a male child.
249. a couple should decide together if they want to have children.
250. a man and a woman should decide together whether to use contraceptives.
251. women in the community are motivated to use modern contraceptives.
252. men in the community are motivated to use modern contraceptives, including supporting female partners.
253. it is easy for women in the community to access and use modern contraceptives.
254. it is acceptable for women in the community and neighborhood to use contraceptives.
255. unplanned pregnancies are not a familiar personal experience.
256. doctors have a great deal of knowledge when it comes to family planning and childbirth.
257. nurses have a great deal of knowledge when it comes to family planning and childbirth.
258. auxiliary nurses have a great deal of knowledge when it comes to family planning and childbirth.
259. the midwives at the clinic have a great deal of knowledge when it comes to family planning and childbirth.
260. family planning counselors have a great deal of knowledge when it comes to family planning and childbirth.
261. community health workers have a great deal of knowledge when it comes to family

planning and childbirth.

262. traditional birth attendants have a great deal of knowledge when it comes to family planning and childbirth.
263. traditional healers have a great deal of knowledge when it comes to family planning and childbirth.
264. religious leaders have a great deal of knowledge when it comes to family planning and childbirth.
265. the youth clinic nearby has a great deal of knowledge when it comes to family planning and childbirth.
266. one's family has a great deal of knowledge when it comes to family planning and childbirth.
267. the health services, clinic, or nearest hospital can be relied upon to deliver safe contraceptives.
268. the health services, clinic, or nearest hospital can be relied upon to deliver family planning counseling.
269. the health services, clinic, or nearest hospital can be relied upon to deliver safe child delivery.
270. the health services, clinic, or nearest hospital can be relied upon to deliver good antenatal care.
271. the health services, clinic, or nearest hospital can be relied upon to deliver good postnatal care.
272. the health services, clinic, or nearest hospital can be relied upon to deliver safe abortion.
273. the health services, clinic, or nearest hospital can be relied upon to deliver HIV testing and counseling.
274. the health services, clinic, or nearest hospital can be relied upon to deliver antiretroviral therapy.
275. the health services, clinic, or nearest hospital can be relied upon to prevent mother-to-

child transmission of HIV.

276. the health services, clinic, or nearest hospital can be relied upon to deliver support for gender-based violence.
277. one possesses a great deal of freedom of choice and control over family planning.
278. it is important for girls to continue their schooling even if they become pregnant and have children.
279. a girl is ready for marriage once she starts menstruating.
280. a girl should honor the decisions/wishes of her family even if she does not want to marry.
281. a boy should honor the decisions/wishes of his family even if he does not want to marry.
282. a girl should wait to have children until she is at least 18 years old, even if she is married.
283. it is safer for a woman to give birth at a clinic than at home.
284. women should have access to safe abortion services to terminate an unwanted pregnancy.
285. is a woman's responsibility to avoid getting pregnant.
286. only when a woman has a child is she a real woman.
287. having a son is always better than having a daughter.
288. contraceptives should be available for everyone, whether or not one is married.
289. sexual education promotes sexual activity among young people.
290. men, and not women, should decide how earnings will be used.
291. men, and not women, should make decisions on major household purchases.
292. men, and not women, should make decisions concerning healthcare visits and spending.
293. men, and not women, should decide whether a woman should give birth at a clinic.
294. men, and not women, should make decisions concerning care for children's health.
295. men, and not women, should make decisions concerning visiting family or relatives.
296. men, and not women, should decide whether girls should go to school.

297. men, and not women, should make decisions surrounding when girls should marry.
298. men, and not women, should decide with whom girls should marry.
299. men, and not women, should decide if and when to have children.
300. men, and not women, should decide on the number of children.
301. men, and not women, should decide if and when to have sex.
302. men, and not women, should decide whether to use condoms.
303. men, and not women, should decide whether to use modern contraceptives other than condoms.
304. men, and not women, should decide if girls should be circumcised.
305. men, and not women, should decide if boys should be circumcised.
306. men should help decide how earnings will be used.
307. men and women should make decisions on how earnings will be used together.
308. men should help make decisions on major household purchases.
309. men and women should make decisions on major household purchases together.
310. men should help make decisions on healthcare visits and spending.
311. men and women should make decisions on healthcare visits and spending together.
312. men should help make decisions on whether a woman should give birth at a clinic.
313. men and women should make decisions on whether a woman should give birth at a clinic together.
314. men should help make decisions on care for children's health.
315. men and women should make decisions on care for children's health together.
316. men should help make decisions on visits to family or relatives.
317. men and women should make decisions on visits to family or relatives together.
318. men should help make decisions on whether girls should go to school.
319. men and women should make decisions on whether girls should go to school together.
320. men should help make decisions on when girls should marry.
321. men and women should make decisions on when girls should marry together.

322. men should help make decisions on with whom girls should marry.
323. men and women should make decisions on with whom girls should marry together.
324. men should help make decisions on if and when to have children.
325. men and women should make decisions on if and when to have children together.
326. men should help make decisions on the number of children.
327. men and women should make decisions on the number of children together.
328. men should play a role in deciding if and when to have sex.
329. men and women should decide together if and when to have sex.
330. men should play a role in deciding whether to use condoms.
331. men and women should decide whether to use condoms together.
332. men should play a role in deciding whether to use modern contraceptives other than condoms.
333. men and women should decide together whether to use modern contraceptives other than condoms.
334. men should play a role in deciding if girls should be circumcised.
335. men and women should decide together if girls should be circumcised.
336. men should play a role in deciding if boys should be circumcised.
337. men and women should decide together if boys should be circumcised.
338. saving money is realistically feasible.
339. one must spend savings or borrow money to get by.
340. over the coming years, the government should emphasize a high level of economic growth.
341. over the coming years, the government should prioritize ensuring the country has strong defense forces.
342. over the coming years, the government should focus on ensuring that people have more say about how things are done at their jobs and in their communities.
343. over the coming years, the government should prioritize work to make the nation's

cities and countryside more beautiful.

344. maintaining order in the nation is of utmost importance.
345. giving people more say in important government decisions is of utmost importance.
346. fighting rising prices is of utmost importance.
347. protecting freedom of speech is of utmost importance.
348. having a stable economy is of utmost importance.
349. progress toward a less impersonal and more humane society is of utmost importance.
350. progress toward a society in which ideas count more than money is of utmost importance.
351. fighting crime is of utmost importance.

Demographic features that we considered include:

1. Respondent sex
    - Male
    - Female
2. Income
    - Monthly income
3. Religion
    - Christianity
    - Islam
4. Residence
    - Urban
    - Rural
5. Occupation
    - Agriculture
    - Business
    - Teacher
    - Trade

- Other
- None

6. Education
    - Arabic education
    - Primary
    - Secondary
    - Post-secondary
    - Unclear
7. Age
    - Respondent age
    - Age as of household head
8. Children
    - Number of children going to school
    - Number of children (under 18) in the household
    - Number of children ever born
    - Number of children alive
    - Number of children dead
9. Cowives
    - Rank among wives