DETECTING COMMUNITY STRUCTURE IN DOMAIN-FEATURES DATA

BY

DIZHOU WU

A Thesis Submitted to the Graduate Faculty of

WAKE FOREST UNIVERSITY GRADUATE SCHOOL OF ARTS AND SCIENCES

in Partial Fulfillment of the Requirements

for the Degree of

MASTER OF SCIENCE

Statistics

May 2024

Winston-Salem, North Carolina

Approved By:

Kenneth S. Berenhaut, Ph.D., Advisor

Staci Hepler, Ph.D., Chair

Ciaran Evans, Ph.D.

Freddie Salsbury, Ph.D.

Acknowledgments

I am profoundly indebted to my advisor, Dr. Kenneth Berenhaut, for his invaluable guidance and support throughout my study and research in statistics. His exemplary dedication and lifestyle are a source of inspiration and aspiration for me. Similarly, I owe a great deal of gratitude to Dr. Freddie Salsbury, my PhD advisor, who introduced me to the rigors of independent research. Despite the challenges, his mentorship has been pivotal in shaping my research skills and intellectual growth.

My thanks extend to my colleagues and peers within the Department of Statistical Sciences and the Department of Physics at Wake Forest University. Their collaboration and camaraderie have significantly enriched my learning and research journey.

I am grateful for the financial support from the CMS, which has been crucial in facilitating my research endeavors with sufficient resources and focus.

Lastly, my heartfelt appreciation goes to my country and family, whose constant love, encouragement, and belief in my potential have been the bedrock of my strength and motivation.

Table of Contents

List of Figures

Abstract

This thesis investigates the enhancement of clustering methodologies on general domains, motivated by questions arising in molecular dynamics (MD) simulations. Traditional clustering methods face limitations such as parameter dependence and inability to reveal complex structures, which Partitioned Local Depth (PaLD) addresses by employing a parameter-free framework that emphasizes local relationships to discern nuanced community structures. The integration of domain-specific features into PaLD extends its applicability, enabling more contextually relevant analyses across various data-centric fields, including time-series, geospatial, and network data. The thesis advocates for future research to refine PaLD, particularly for large-scale MD datasets, underscoring the method's potential to significantly advance data science and biomolecular research.

Chapter 1

Introduction

Since their origin in 1977, molecular dynamics (MD) simulations have profoundly influenced computational biology, providing a detailed lens into the dynamic behavior of biomolecules [31]. These simulations produce vast datasets, necessitating sophisticated clustering methods to decipher complex patterns and conformational changes [25]. Traditional clustering techniques, however, struggle with parameter dependence and often overlook the nuanced internal structure and kinetic dynamics within clusters, calling for more advanced analytical methods [24, 39, 36].

This thesis begins in Chapter 2 with background on MD simulations and a discussion of relevant traditional clustering methods, emphasizing their limitations and the impact on biomolecular analysis. Chapter 3 introduces the PaLD method, an innovative clustering approach free from the constraints of predefined parameters, enhancing the understanding of cluster internals. Chapter 4 explores the integration of domain information into the PaLD framework, offering a more dynamic perspective on clustering data. Finally, Chapter 5 presents various applications of the method, demonstrating its effectiveness in extracting meaningful insights from complex datasets of varying types. Specific applications include revealing local community structure for time-, geographic- and network-based data.

Chapter 2

Motivation and Background

Much of the motivation for our work stems from considerations arising in clustering for molecular dynamics. In this chapter, we introduce the critical role of clustering in the analysis of MD simulations, particularly in understanding protein conformation and function. We delve into traditional clustering methods commonly employed in MD simulations, highlighting their strengths and limitations. This discussion sets the stage for our investigation into advanced clustering techniques, aiming to address the challenges inherent in conventional methods so that one may improve the accuracy and depth of MD simulation analysis.

2.1 Molecular Dynamics Simulations in Protein Conformation Analysis

MD simulations are a computational technique that models the motion of molecular systems over time, providing insights into the dynamic nature of biomolecules such as proteins. A protein's conformation, the three-dimensional arrangement of its atoms, is vital for understanding its biological function and interactions [31, 26, 42, 14, 20].

The initiation of MD simulations in 1977 by McCammon, Gelin, and Karplus [31] was a pivotal advancement in computational biology. These simulations enable the exploration of atomic-level interactions within biomolecules, which are difficult to

observe through conventional experimental methods.

Using experimental structural data, classical MD simulations apply Newton's equations of motion to predict the time-evolution of atomic positions and velocities within a molecular system. This generates a trajectory that elucidates the molecule's time-dependent behavior, offering a detailed view of its dynamic conformational changes [26, 42, 14, 20]. Through such simulations, researchers can observe the intricate dynamics and interactions of proteins, aiding in the deciphering of their functional mechanisms [19, 58, 57, 59, 60, 56, 55, 54].

Particularly, MD simulations at the microsecond scale and beyond have enabled the exploration of a wide range of protein conformations, facilitating a deeper understanding of their functional mechanisms [40, 56, 55, 53, 49, 29]. However, the extensive data generated by these simulations pose a significant challenge in terms of analysis and interpretation. The sheer volume of data, consisting of hundreds of thousands of frames, makes it impractical to manually inspect each frame for conformational changes.

To address the challenge of analyzing and interpreting the voluminous data from MD simulations, researchers employ clustering algorithms to systematically organize and scrutinize the generated datasets. These algorithms cluster similar conformations, thereby pinpointing representative structural states and their transitions, which are crucial for extracting meaningful biological insights [33, 57, 59, 58, 55, 54].

For example, Figure 2.1 demonstrates the representative structures of the 60s loop of thrombin, categorized into three clusters using HDBSCAN[1] based on the coordinates of the heavy atoms in the 60s loop, and Figure 2.2 depicts the time-evolved cluster membership in MD simulations, effectively outlining three biologically significant conformations, illustrating the dynamic unfolding of the 60s loop, and

[1]Hierarchical density-based spatial clustering of applications with noise. See Section 2.2.3, below, for a discussion of various density-based clustering methods.

identifying transition states as noise.

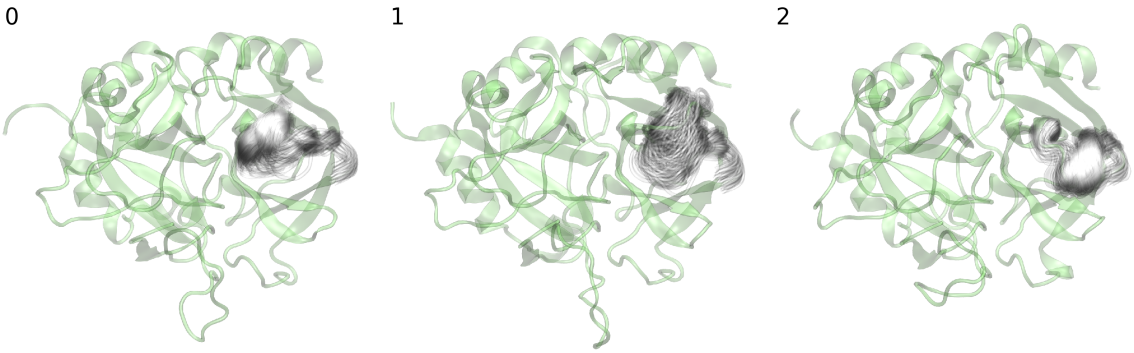


Figure 2.1: Representative structures of the 60s loop heavy atoms in Thrombin, TM56-bound, and TM456-bound Thrombin, clustered using HDBSCAN. The median thrombin structure is shown in transparent green, with the variance in the 60s loop across clusters illustrated by shadows, following the visual statistics method [34].

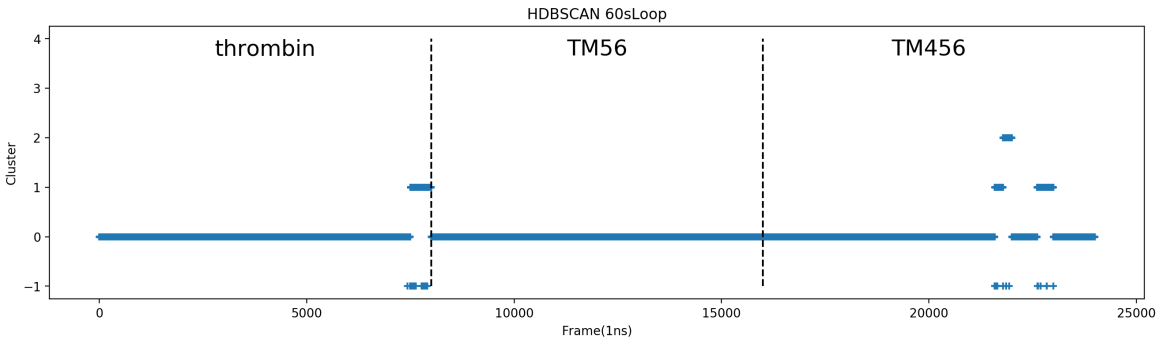


Figure 2.2: HDBSCAN cluster membership of frames throughout the MD simulation timeline, illustrating the transitions between clusters. This visualization effectively captures the dynamic behavior of the system, revealing the temporal progression and changes in cluster affiliation. The label '-1' denotes 'Noise', referring to data points not assigned to any cluster due to their low density, possibly representing transition structures.

However, the effectiveness of clustering can vary. When the Amorim-Hennig clustering method was applied to the same dataset, it inaccurately grouped structures into two clusters that were temporally mixed, as shown in Figure 2.3. Moreover, these clusters had nearly identical representative structures (see Figure 2.4), failing to reveal any significant biological activity. Similar issues can arise with HDBSCAN

when applied to different systems, underscoring the complex challenges and inherent limitations of conventional clustering methods. These topics will be further discussed in Section 2.2, "Traditional Clustering Methods in MD Simulations", and Section 2.3, "Limitations of Traditional Clustering Methods in MD Simulations".

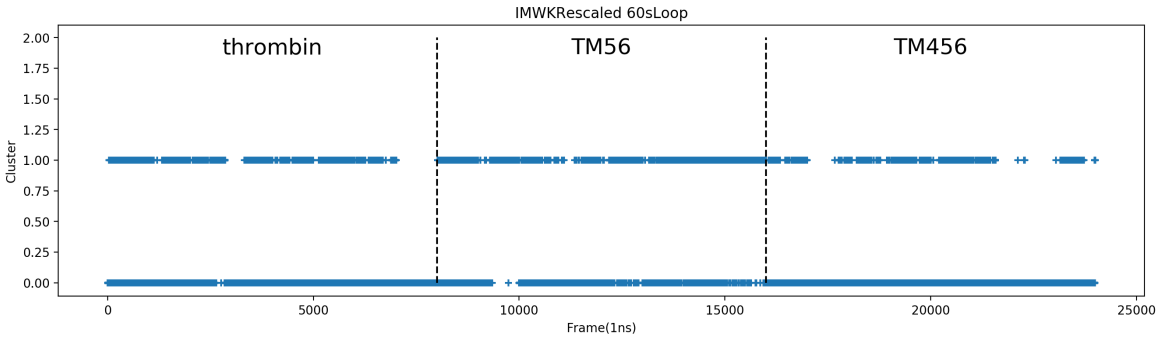


Figure 2.3: Cluster membership over time using the Amorim-Hennig method in MD simulations. The plot reveals a temporal overlap of frames among clusters, suggesting difficulties in distinguishing biologically relevant structures, which highlights the method's limitations in effectively separating distinct conformational states.

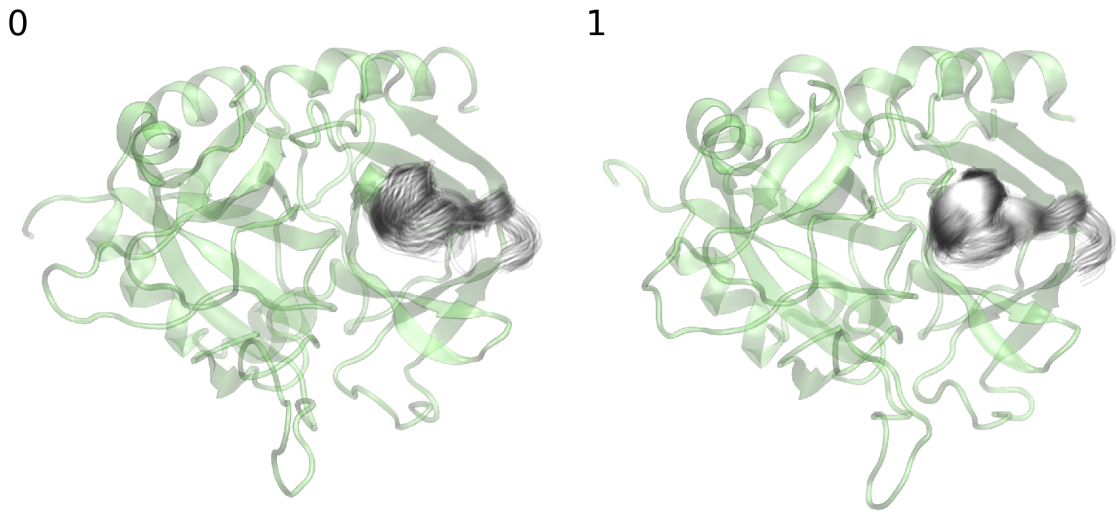


Figure 2.4: Representative structures of the 60s loop heavy atoms in Thrombin, TM56-bound, and TM456-bound Thrombin, clustered using Amorim-Hennig clustering. The median thrombin structure is shown in transparent green, with the variance in the 60s loop across clusters illustrated by shadows, following the visual statistics method [34].

2.2 Traditional Clustering Methods in MD Simulations

Clustering methods are essential in unsupervised machine learning, pivotal for pattern recognition within unlabeled datasets [25]. A wide range of clustering algorithms, primarily developed in computer science [30, 17, 25, 61, 5, 1], has been successfully adapted for MD datasets [33, 57, 59, 58, 55, 54]. In MD simulations, specific clustering algorithms are routinely utilized, including hierarchical clustering, center-based clustering, density-based clustering, and topological data analysis (TDA). Each of these algorithms addresses distinct facets of the data, facilitating a comprehensive analysis of the complex molecular structures and dynamics inherent in MD studies.

2.2.1 Hierarchical Clustering Algorithms

Hierarchical clustering in MD analysis is crucial for organizing conformations into a multilevel cluster hierarchy, revealing the structural nuances of biomolecular systems. This technique is bifurcated into two main strategies: divisive methods, which systematically divide a single cluster into smaller, more refined groups, and agglomerative methods, which incrementally merge individual conformations into comprehensive clusters.

Within these strategies, various linkage criteria that determine the clustering dynamics are employed:

1. **Single-Linkage Clustering:** Utilizes the shortest distance between clusters, favoring the formation of elongated clusters. While it's computationally efficient, its sensitivity to outliers can result in less cohesive cluster formation [46].

2. **Complete-Linkage Clustering:** Based on the furthest distance between clusters, this method yields more compact and clearly delineated clusters, effectively countering the effect of outliers [48].

3. **Average-Linkage Clustering:** Considers the average distance between clusters, striking a balance between compactness and separation, and is considered well-suited for a broad range of clustering scenarios [47].

4. **Centroid-Linkage Clustering:** Determines cluster proximity through the geometric mean, potentially leading to non-monotonic distance measures and inversions, which can affect the clustering hierarchy [47].

5. **Ward-Linkage Clustering:** Focuses on minimizing the variance increase within clusters, often resulting in well-defined and compact clusters, ideal for detailed MD analysis of structure-function relationships in biomolecules [52].

These hierarchical clustering algorithms significantly contribute to analytical depth in MD simulations, aiding in the systematic exploration of biomolecular dynamics and structural properties. For further details on application of hierarchical clustering in MD analysis, see references [41, 38, 18].

2.2.2 Center-Based Clustering Algorithms

Center-based clustering algorithms are pivotal in the analysis of MD simulations, as they identify representative conformations by selecting a predefined number of them as cluster centers and grouping others based on proximity. Notably, four center-based clustering algorithms are commonly employed:

1. **k-means:** This algorithm segregates MD conformations into k distinct clusters to minimize the within-cluster sum of squared distances (WCSS) between the conformations and their respective cluster centers. The centers are calculated as the geometric means of the conformations within each cluster. Iteratively, the algorithm selects initial cluster centers, assigns conformations to the nearest center, and recalculates the centers until there is minimal change in WCSS [30].

2. **k-medoids:** This algorithm, akin to k-means, partitions the dataset into k clusters but uses actual conformations (medoids) as the cluster centers, rather than geometric means. By minimizing the total dissimilarity between each conformation and its cluster center, k-medoids offers robustness against noise and outliers, yielding more interpretable results for MD simulations [27].

3. **k-centers:** This algorithm aims to establish k clusters by minimizing the maximum distance from any conformation to its nearest cluster center. The process initiates with a randomly selected conformation as the first center and subsequently chooses new centers as conformations most distant from the existing centers. This method is particularly adept at managing outliers and guarantees comprehensive coverage of the conformational space [21].

4. **Amorim-Hennig Algorithm (iMWk-means with explicit re-scaling followed by k-means):** An advanced adaptation of k-means, this algorithm starts with intelligent k-means (ik-means) [9] to autonomously determine the cluster count and initial centroids by detecting distinct patterns in the data. It incorporates the Minkowski Weighted k-means (MWk-means) algorithm [12], which modulates feature relevance through weight assignment based on their clustering contribution and employs the Minkowski distance to accommodate various cluster geometries beyond the spherical model intrinsic to k-means. The explicit re-scaling stage further tailors the feature scale in the dataset, enhancing alignment with its intrinsic structure. Subsequently, k-means is reapplied to the re-scaled data, utilizing refined centroids and feature weights for improved clustering accuracy. This elaborate method amalgamates ik-means' automatic cluster and centroid initialization with MWk-means' feature-sensitive clustering and k-means' operational simplicity, culminating in a robust, nuanced, and efficient clustering mechanism, especially effective in delineating stable biomolec-

ular systems [11, 33].

Employing these algorithms significantly enhances the analysis and interpretation of large-scale MD simulations, facilitating the discernment of representative conformations and shedding light on the dynamic and structural attributes of the system.

2.2.3 Density-Based Clustering Algorithms

Density-based clustering algorithms are essential for analyzing MD conformations, clustering them based on density variations. These algorithms are particularly suitable for MD simulations of macromolecular systems, where high-density regions often correspond to free energy minima, making them intuitive choices for identifying stable conformations. Among various density-based algorithms, we focus on the neighbor-based algorithm, the density-peaks algorithm, the Robust-DB algorithm, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), and HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise).

1. **Neighbor-Based Algorithm:** Pioneered by Daura et al. [10], this algorithm employs a predetermined cut-off radius to calculate the density of conformations. The conformation with the highest number of neighbors within this radius is chosen as the initial cluster center. After assigning members to this cluster, they are removed from the pool, and the process repeats for the remaining conformations. This method has been effectively used to cluster MD trajectories and analyze the folding dynamics of a β-heptapeptide, aligning the predominant conformations with experimentally determined structures [10].

2. **Density-Peaks Algorithm:** This algorithm, introduced by Rodriguez and Laio [39], detects cluster centers as local density maxima separated by lower-density regions. It calculates each conformation's local density and distance

to the nearest higher density conformation, identifying cluster centers using a decision graph. The algorithm, with a runtime complexity of O(n^2), has been effectively used in MD studies to efficiently handle large datasets.

3. **Robust-DB Algorithm:** The Robust-DB algorithm [45] estimates local free energies of MD conformations to segregate them into clusters representing local free energy minima. Initially, dimensionality reduction via PCA is performed, followed by local population estimation within a hypersphere radius. Clustering at various free energy cutoffs yields microstates, aiding in the identification of metastable states. This algorithm has proven to be robust, deterministic, and computationally efficient in analyzing conformational dynamics of peptides.

4. **DBSCAN Algorithm:** Developed by Ester et al. [17], DBSCAN groups conformations in high-density areas into clusters while treating sparse ones as noise. It requires two parameters: the neighborhood distance ϵ and the minimum number of neighbors (minPts). Conformations are categorized as core, border, or outliers based on these parameters, and clusters are formed by linking core conformations within ϵ distance. DBSCAN is known for its efficiency, offering an average runtime complexity of O(nlogn) [17].

5. **HDBSCAN Algorithm:** HDBSCAN, an extension of DBSCAN, evolves into a hierarchical clustering algorithm that obviates the necessity to predetermine the number of clusters [6]. Its popularity surged following the development of its accelerated variant [32]. HDBSCAN enhances density-based clustering by evaluating core distances within ϵ neighborhoods and forming a hierarchy of clusters at different density levels. This approach determines clusters based on their stability across varying densities, facilitating the automatic estimation of clusters and improved management of noise and outliers. HDBSCAN is

adept at identifying clusters with diverse densities and sizes, rendering it an adaptable tool for analyzing complex MD datasets where conformational states exhibit substantial density and distribution variations. It efficiently identifies stable regions and captures significant structural transitions, offering a broad perspective on the structural dynamics within the system [33].

These density-based clustering algorithms provide powerful tools for dissecting large-scale MD simulations, offering insights into the stable and transient conformations within macromolecular systems.

2.2.4 Topological Data Analysis (TDA)

Topological Data Analysis (TDA) represents a novel methodology in data science, analyzing high-dimensional data sets through the principles of topology. It focuses on the intrinsic shape, connectivity, and structure of data, going beyond traditional clustering methods that depend on metrics and coordinates [8].

1. **Fundamentals of TDA:** TDA posits that data embody a topological space or shape, enabling the analysis of global features like loops, holes, and connected components. These features are stable under continuous transformations, offering a consistent perspective on the data's structure. The mathematical underpinning of TDA involves algebraic topology, with homology and persistent homology being key tools for quantifying topological features across various scales [7].

2. **Persistent Homology:** Persistent homology, a cornerstone of TDA, monitors the emergence and dissolution of topological features against a scaling parameter, typically associated with distance metrics. The outcome is a persistence

diagram or barcode that chronicles the lifespan of these features, thus delineating the data's topological structure across different scales [16].

3. **Mapper Algorithm:** The Mapper algorithm is an integral part of TDA, offering a simplified yet insightful view of the data's topological shape. It operates by filtering the data and organizing it into interlinked clusters, subsequently visualized in a graph known as the Mapper complex. This representation captures the high-dimensional topology of the data, making it easier to identify and understand complex patterns [44].

Applications of TDA include analyzing the structural dynamics of proteins, such as the SARS-CoV-2 spike protein's D614G mutation and its interaction effects, highlighting TDA's capability to discern critical structural and interactional variations [28]. Further, TDA has been instrumental in identifying mutation-induced conformational changes in β-lactamase and HIV-1 protease, thus contributing to the understanding of drug resistance mechanisms [50]. Integrating TDA with machine learning, the field has expanded to offer new analytical perspectives, enhancing the interpretability and robustness of clustering analysis across various domains [22].

2.3 Limitations of Traditional Clustering Methods in MD Simulations

Traditional clustering methods, while widely used in MD simulations, exhibit several limitations that can impact their effectiveness in analyzing complex biomolecular data:

1. **Parameter Dependence:** Many clustering algorithms require the pre-determination of specific parameters, which can significantly influence the outcome of the clustering process [24]. For example, the k parameter in k-means clustering dictates

the number of clusters to be formed, and the MinPts and ϵ parameters in DBSCAN are crucial for defining the density criteria for cluster formation. Selecting appropriate values for these parameters often necessitates prior knowledge or trial-and-error experimentation, which can be time-consuming and introduce human bias.

2. **Intra-cluster Structure Obscurity:** Traditional clustering approaches typically focus on grouping data points into distinct clusters based on similarity measures but do not provide insights into the internal structure or sub-groupings within each cluster [39]. This limitation can obscure detailed information about the conformational landscape of biomolecules, potentially overlooking important sub-states or transitional configurations.

3. **Exclusion of Kinetic Information:** While traditional clustering techniques excel in grouping data points based on static similarity measures, they largely overlook the kinetic information and temporal relationships inherent in MD simulations. This oversight means that such methods might fail to effectively capture and represent the dynamic processes, transition pathways, and temporal evolution of biomolecular systems. It is crucial to acknowledge that methodologies such as the Hidden Markov Model (HMM) [15, 37, 23] provide a robust framework for incorporating dynamics into the analysis of MD simulations without the need for clustering. However, the application of HMMs requires predefined emission distributions and a specified number of states. Establishing these inputs a priori can be a significant challenge, particularly in high dimensions. This highlights a gap in directly leveraging kinetic data for cluster formation and necessitates the development or integration of approaches that can intuitively account for the temporal aspect of MD simulations.

These limitations highlight the pressing need for the development of more advanced clustering techniques or the refinement of existing methods to better address these challenges. Such advancements could enable a more detailed and comprehensive analysis of MD simulation data, paving the way for deeper insights into biomolecular dynamics.

In the next chapter, we will delve into the Partitioned Local Depth (PaLD) methodology, discussing its practical applications and the technical intricacies of its implementation. We will explore how the PaLD framework facilitates the analysis of complex datasets by overcoming the limitations of traditional clustering methods, demonstrating the effectiveness of PaLD in real-world scenarios. Following this, we will present a recent generalization of the PaLD method which provides crucial insights that will play into developments and applications provided in later chapters.

Chapter 3

Partitioned Local Depth (PaLD) Methodology

The partitioned local depth (PaLD) methodology, introduced by Berenhaut et al. [4], offers a parameter-free approach to community detection.

3.1 Concept and operational mechanism

The PaLD approach operates on the base of 'local depth', assessing relationships among data points based on their relative positions. It involves defining a 'local focus' for each pair of points, and then evaluating the 'support' of other points relative to this pair, resulting in a measure of connection termed 'cohesion'.

3.1.1 Local focus

For any two points x, y in a finite set S of data points, endowed with a pairwise measure of dissimilarity, d, the local focus $U_{x,y}$ (illustrated in Figure 2 for Euclidean distance) is defined as:

$$U_{x,y} \stackrel{\text{def}}{=} \{z \in S \mid d(z,x) \leq d(y,x) \text{ or } d(z,y) \leq d(x,y)\}.$$

This local focus helps identify points that are "locally" relevant to a pair, facilitating a more nuanced understanding of community structure.

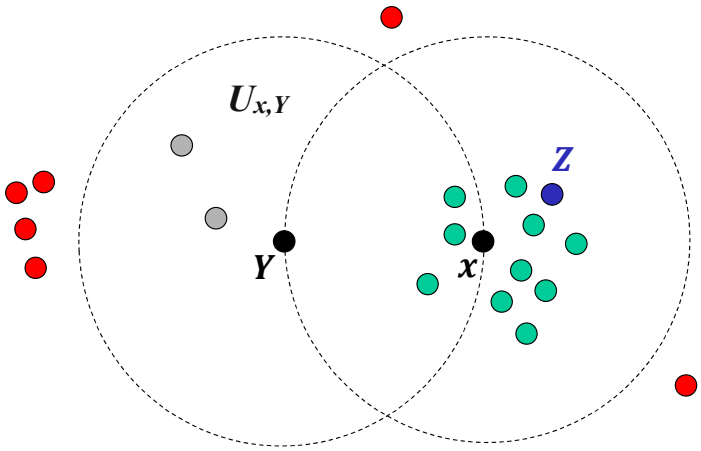


Figure 3.1: Depiction of the PaLD method showcasing the concept of local focus and support points. The dashed circles represent the local focus for points x and y. Points in red are outside the local focus of x and y. The cyan and blue points illustrate those supporting x, while the gray points indicate support for y. This visualization aids in understanding how PaLD discerns community structure based on proximity and support within the local focus. Figure included, with permission, from [3].

3.1.2 Cohesion and Strong Cohesion

The cohesion (partitioned local depth), $C_{x,w}$ of a point w to x is defined as the probability that a point Z selected uniformly at random from $U_{x,Y}$ (where Y is chosed uniformly from $S \setminus \{x\}$) is equal to w and is closer to x than it is to Y, i.e.

$$C_{x,w} \stackrel{\text{def}}{=} P(Z = w \text{ and } d(Z,x) < d(Z,Y)). \tag{3.1}$$

For the connection to "local depth", leading to the alternate name "partitioned local depth", see [4].

The threshold for strong cohesion is defined as:

$$T = P(Z = W \text{ and } d(Z,X) < d(Z,Y)), \tag{3.2}$$

where X, Y, Z, and W are selected uniformly at random from S, $S \setminus \{X\}$, $U_{X,Y}$, and $U_{X,Y}$ respectively.

Assuming that no other point is closer to x than x itself, a mild condition, we can calculate the threshold, T, as one half the average of the diagonal of the matrix of pairwise cohesion values (see [4]):

$$T \stackrel{\text{def}}{=} \frac{1}{2}\left(\frac{1}{n}\sum_{x=1}^{n} C_{x,x}\right). \tag{3.3}$$

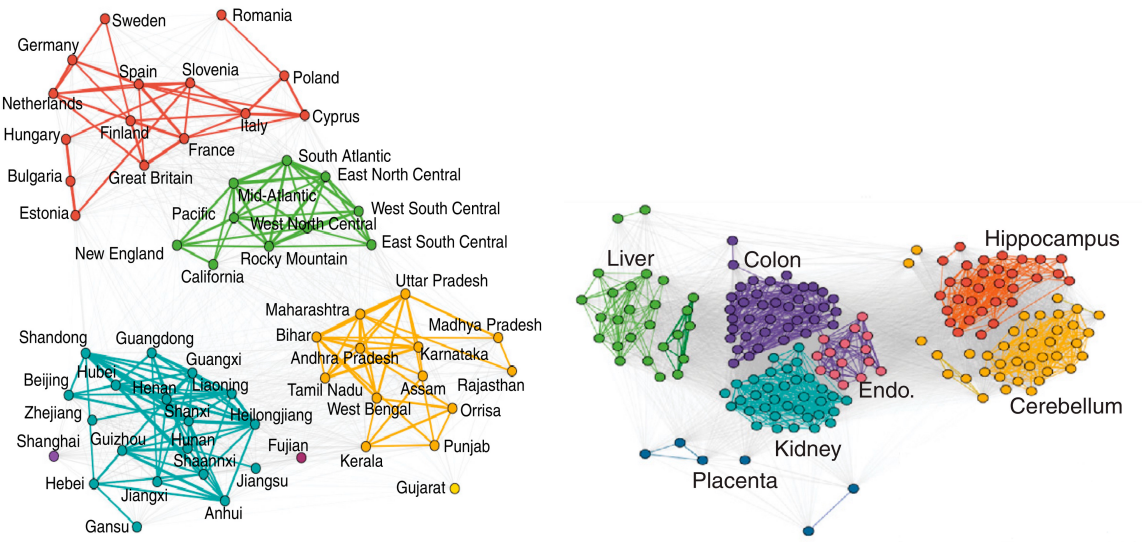
A strong relationship between x and w is established when:

$$\min\{C_{x,w}, C_{w,x}\} \geq T. \tag{3.4}$$

3.2 Advantages and applications

PaLD's adaptability to various data densities and structures (through local perspectives) is one of its key advantages. It effectively handles heterogeneous data, revealing intricate community structures that other methods might overlook.

Figures 3.2a and 3.2b demonstrate PaLD's effectiveness in various domains, including cultural and gene expression data analysis. These examples underscore PaLD's versatility and its ability to uncover deep community structures in diverse data sets. For details on the data considered, see [4]. The figures are replicated here with permission.



(a) Cultural data analysis using PaLD (b) Gene expression data analysis using PaLD

Figure 3.2: Applications of PaLD in different contexts. For details see [4]. Figures (a) and (b) are included, with permission, from [4].

3.3 Generalized Partitioned Local Depth

The Generalized Partitioned Local Depth (GPaLD) approach extends the traditional PaLD framework by integrating probabilistic aspects to better handle the uncertainties present in data relationships. This integration facilitates a more detailed analysis of data structures, particularly useful when dissimilarities are ambiguous or inherently uncertain, as explored in [3].

In this generalized setting, refined definitions of local focus, local depth, and cohesion are introduced, accounting for the uncertain nature of dissimilarities. Consider a pair $x, y \in S$; the local focus $U_{x,y}$ is expanded by defining the local relevance $R_{x,y,z}$ of an element z in relation to x and y. This relevance quantifies the probability of z being in close association with x and y, described by the equation:

$$R_{x,y,z} \stackrel{\text{def}}{=} P(z \in N(x, y)), \tag{3.5}$$

where $\mathcal{N} \stackrel{\text{def}}{=} \{N(x,y)|(x,y) \in S \times S\}$ denotes a stochastic neighborhood structure over the pairings in $S \times S$. The probabilistic measures are encapsulated within the matrix $\mathbf{R} \stackrel{\text{def}}{=} [R_{x,y,z}]$.

For establishing cohesion in this generalized framework, we outline a method to sample an element Z from S that is local to x, y. The selection probability for any $z \in S$ is given by:

$$P(Z = z) = P_{x,y}(Z = z) = \frac{R_{x,y,z}}{\sum_{w \in S} R_{x,y,w}}. \tag{3.6}$$

Additionally, for any distinct $x, y, z \in S$, we define the support division probability $Q_{x,y,z}$ as:

$$Q_{x,y,z} \stackrel{\text{def}}{=} P(\mathcal{C}_z(\{x,y\}) = x), \tag{3.7}$$

illustrating the probability of z supporting x over y. Here $\mathcal{C}_z$ is a random choice function.

The *cohesion* of w to x, $C_{x,w}$, can now be defined in a similar fashion to (3.1) via

$$C_{x,w} \stackrel{def}{=} P(Z = w, \mathcal{C}_Z(\{x,Y\}) = x). \tag{3.8}$$

where, Y is selected uniformly from $S \setminus \{x\}$, and Z is selected as in (3.6).

In the context of generalized PaLD, a new threshold $T_{S,\mathbf{R},\mathbf{Q}}$ is established to differentiate strong from weak cohesion, expressed as:

$$T_{S,\mathbf{R},\mathbf{Q}} \stackrel{\text{def}}{=} P(Z = W, \mathcal{C}_Z(\{X,Y\}) = X), \tag{3.9}$$

This threshold modification accounts for the complex interaction dynamics captured by the $\mathbf{R}$ and $\mathbf{Q}$ matrices, leading to an inequality:

$$T \leq \frac{1}{2n} \sum_{x \in S} C_{x,x}, \tag{3.10}$$

which underscores the probabilistic nature of cohesion in this enhanced model.

The generalized PaLD framework posits that the choice mechanism, represented by $\{\mathcal{C}_z : z \in S\}$, is not necessarily connected to the neighborhood structure $\mathcal{N}$, thus allowing for a more flexible interpretation of relative relationships [3]. This implies that the metrics governing local focus $U_{x,y}$ and cohesion $C_{x,w}$ can be distinct, accommodating the variability and complexity of the underlying data space. We will leverage this fact in the next chapter to integrate domains (such as time, geometric locality and network structure) into the PaLD framework.

Chapter 4

Integrating domains into PaLD

While PaLD excels as a parameter-free approach, providing deep insights into the intrinsic structure of datasets, it may not fully capture the nuances introduced by inherent domain characteristics that significantly shape data relationships and classifications. This domain can manifest in various forms, such as time sequences (as is the case in molecular dynamics as considered in Chapter 2), geographic spaces, or graph structures. A notable example is observed in datasets with temporal domains, where the same set of points can yield differing community structures under different time orders. As a simple illustration in Figure 4.1, the visible community configuration alters dramatically with the change in time order; one scenario presents a distinct singular community, while another reveals a bifurcated community structure.

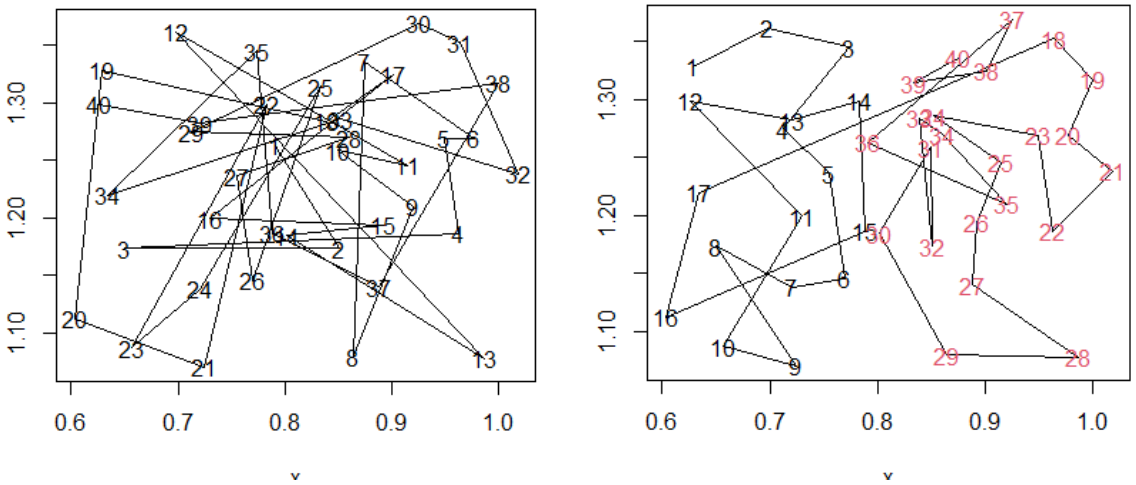


Figure 4.1: Comparison of two community structures under differing time orders, illustrating the impact of the temporal domain on community detection. Label coloring reflects community membership according to the method proposed below. Note that points here are labeled by time order.

This phenomenon underscores the potential for integrating domain considerations into community detection methodologies. In this section, we aim to augment the PaLD approach by incorporating domain aspects, thereby enhancing its efficacy in deciphering complex data structures. This integration is not limited to temporal domains but extends to other fundamental structures such as geographic locations and graph networks, broadening the scope and applicability of our method in diverse analytical scenarios.

4.1 Adapting local focus and cohesion for domains

In this extended framework of PaLD, we consider two distinct types of dissimilarity measures: *domain distance* (d_D) and *feature distance* (d_F), both defined over the set S.

- **Local focus for domain distance:** The local focus, $U_{x,y}$, for a pair of points

$x, y \in S$ is now defined with respect to the domain distance d_D:

$$U_{x,y} = \{z \in S \mid d_D(z, x) \leq d_D(y, x) \text{ or } d_D(z, y) \leq d_D(x, y)\}. \tag{4.1}$$

This redefinition allows the incorporation of domain-specific characteristics into the analysis with regard to locality.

- **Cohesion (support) for feature distance:** The cohesion $C_{x,w}$ of a point w to x is then computed using the feature distance d_F:

$$C_{x,w} = P(Z = w \text{ and } d_F(Z, x) < d_F(Z, Y)), \tag{4.2}$$

where Y and Z are selected uniformly at random from $S \setminus \{x\}$ and $U_{x,Y}$, as given in Equation (4.1), respectively.

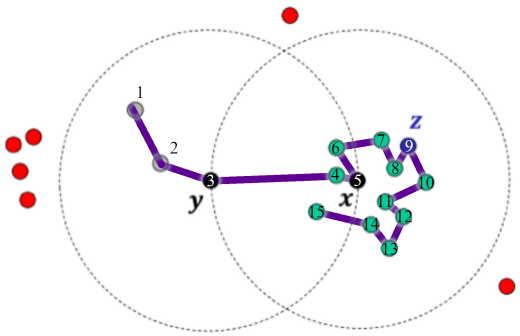
Note that the computational form of the threshold (i.e. half the average of the diagonal of the matrix of pairwise cohesion values) caries over upon applying results from [3]. As mentioned therein, there is actually equality in (3.10) whenever $\mathbf{R}$ is a (0,1) array, (allowing for complete flexibility in the choice of support division). This is the case in (3.3).

4.2 Impact of domain integration on PaLD

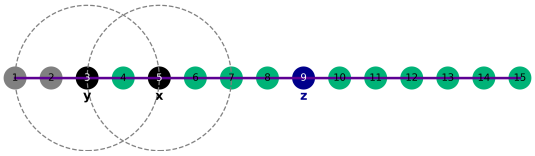
The integration of domains into PaLD can have a significant impact on results. This is quite evident when using time as an underlying domain for an ambient feature space.

In traditional PaLD, any close points in S could be included in the local focus. However, when time is used as a domain, only points within a specific time range are considered, leading to potentially different interpretations of the data structure. This change in perspective is illustrated in Figure 4.2, where the local focus is defined using

distinct metrics across the subfigures. Specifically, Part (a) employs 2-D Euclidean distance between points, emphasizing spatial proximity. In contrast, Part (b) utilizes time distance, wherein distance is quantified by the number of intervening time steps between points, such as those at Steps 3 and 5, resulting in a time distance of 2.



(a) Traditional PaLD



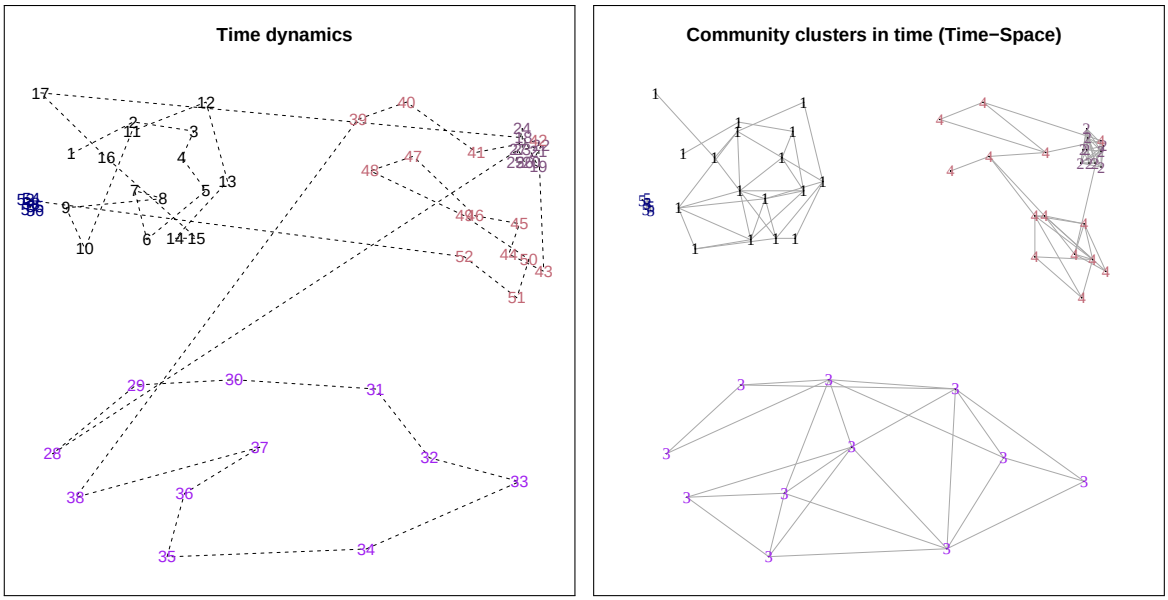
(b) Time-domain-adjusted PaLD

Figure 4.2: Comparison of traditional PaLD and time-domain-adjusted PaLD. This figure demonstrates the influence of temporal integration on the local focus and cohesion among data points. The numbers 1-15 indicate the time sequence of the points. In traditional PaLD (a), the local focus of points x and y is shown without considering time, while in the time-domain-adjusted PaLD (b), the local focus is redefined to incorporate time order, leading to a different interpretation of the structure of the data.

4.3 2-D time-space example

This subsection presents a simple example of the time-domain-adjusted PaLD methodology in a 2-D time-space analysis, wherein, time is the domain and the features are in two-dimensional Euclidean space. The approach allows us to explore how temporal sequences affect community structures and relationships within data. Notably, the

application of time-domain considerations alters not only the composition of the communities but also their structural organization within the same community (Figure 4.3).



(a) Time dynamic of the points (b) Time-domain-adjusted PaLD result

Figure 4.3: A comparative analysis of PaLD results with and without time-domain adjustments. The plot in (a) illustrates the time dynamics of the data points, while (b) shows the community structure as identified by time-domain-adjusted PaLD. This comparison underscores the significant impact of temporal information on community detection and analysis.

Figure 4.4 illustrates the application of our method to geographic latitude-longitude data across the United States, demonstrating its efficacy in spatial analysis compared to the traditional PaLD approach depicted in Figure 4.5. In Figure 4.4, the corresponding feature space is based on four scaled education-based variables crossing decades - illiteracy and high school graduate proportions in 1977, and high school graduate and bachelor's degree graduates in 2019 (each scaled into standard units). The domain is the centroids' latitude-longitude coordinates under Haversine distance, while the features are considered in 4-dimensional Euclidean space. In contrast, in Fig-

ure 4.5, the feature and domain spaces are both based strictly on latitude-longitude data (and the education variables are disregarded). Further exploration of similar datasets will be undertaken in the subsequent chapter.

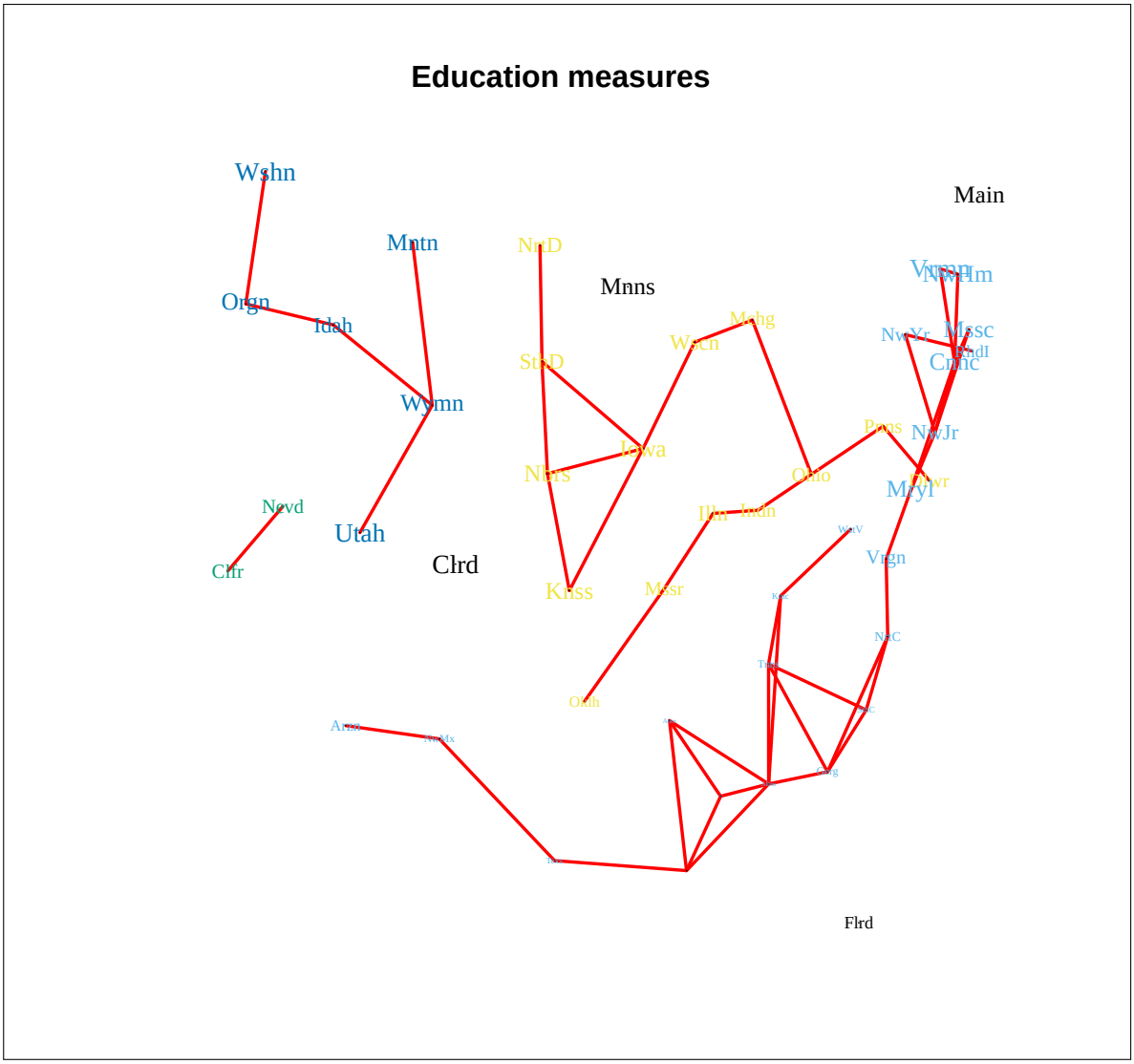


Figure 4.4: Community structure for latitude-longitude domain data. The corresponding feature space is based on four scaled education-based variables, crossing decades – illiteracy and high school graduate proportions in 1977, and high school graduate and bachelor's degree graduates in 2019. The domain is the centroids' latitude-longitude coordinates under Haversine distance, while the features are k-dimensional Euclidean space. The font size is scaled to reflect the average value of these four educational variables. Note the regional clusters, displaying the Midwest, Pacific Northwest and the SouthEast and Eastern seaboard.

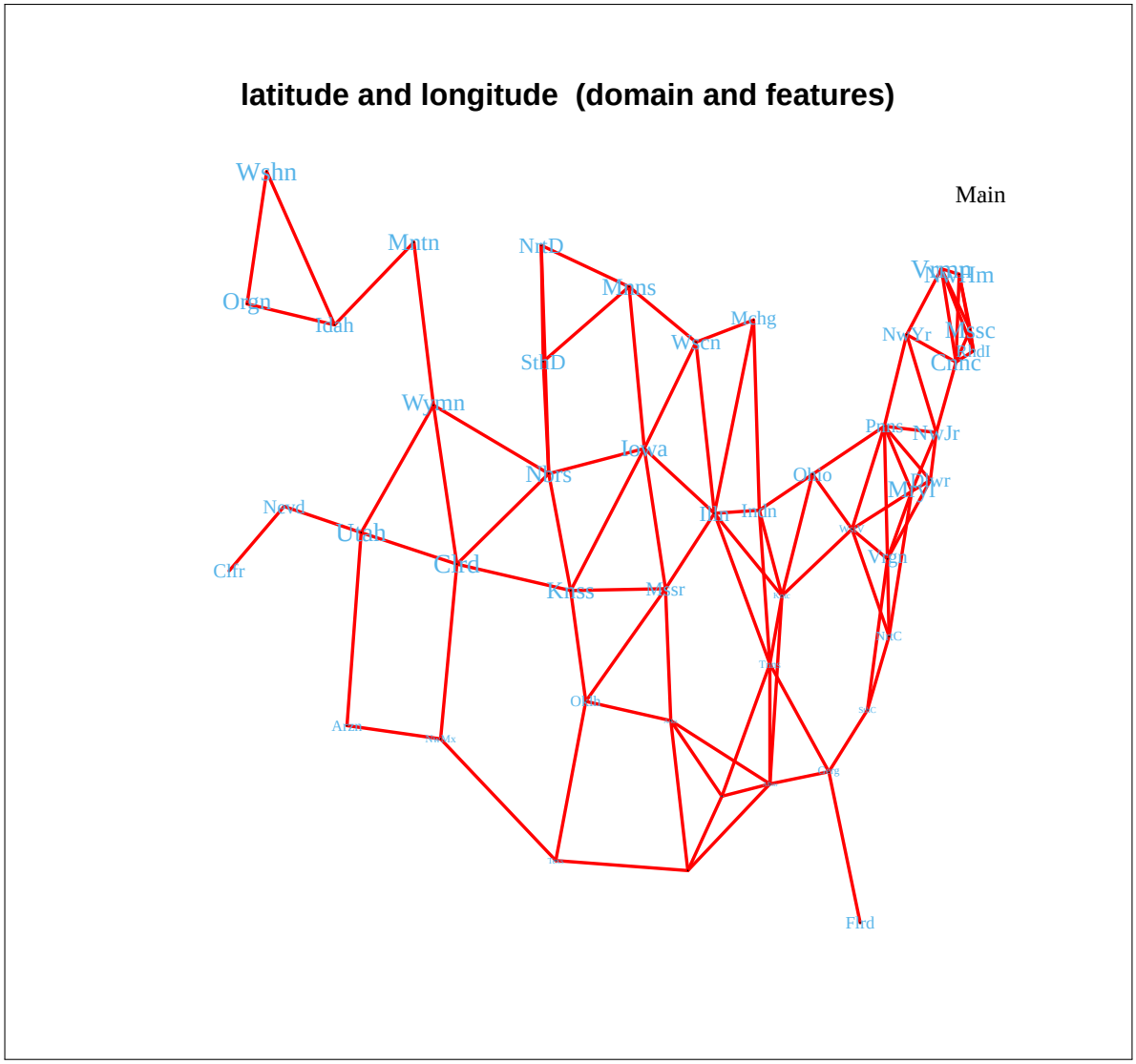


Figure 4.5: Community structure for latitude-longitude. Provided for comparison when the feature and domain spaces are both based strictly on latitude-longitude data (and education is disregarded).

We now turn to the consideration of general applications of the method, with a focus on time-, geographic-, and network-type data.

Chapter 5

Applications and Examples of Community Structure in Domain-Features Data

The integration of domain-specific features into the PaLD methodology enhances its versatility and applicability across various fields. This chapter delves into the multifaceted applications of the approach, demonstrating its adaptability and efficacy in addressing unique challenges in data analysis.

5.1 Expanding the Scope of PaLD

PaLD's adaptability to domain-specific features allows for the nuanced analysis of complex data sets. By incorporating different domains alongside feature spaces, PaLD becomes a powerful tool for uncovering intricate community structures, facilitating a deeper understanding of the underlying data.

5.1.1 Time Series Analysis

In time series analysis, the domain-integrated PaLD method proves invaluable for detecting change points and identifying significant temporal shifts or anomalies. This application is critical in fields such as finance, where market trends can be analyzed; weather forecasting, where it can help predict climatic changes; and process monitor-

ing in industrial settings, ensuring operational efficiency and safety.

5.1.2 Geospatial Analysis

Geospatial analysis can benefit from an ability to integrate geographic space as a domain, enhancing the detection of local outliers and regional structures. This feature can be particularly beneficial for environmental monitoring, urban planning, and resource management, where spatial relationships and distributions are crucial.

5.1.3 Network Analysis

In network analysis, domain-integrated PaLD assists in identifying nodes with significant deviations from their surroundings, can enhance the understanding of network community stability. This is invaluable in analyzing social networks, biological networks, and in neuroscientific studies where brain networks are examined.

5.2 Illustrative Examples and Applications

To showcase the practical implications of the domain-integrated PaLD approach, this section will present varied examples, encompassing both artificial and real-world scenarios, to demonstrate its robust ability to identify and analyze community structures within different datasets.

Initially, we delve into artificial and simulated datasets where the domain is temporal, and the feature space exhibits low dimensionality. This exploration will help illustrate how PaLD effectively uncovers patterns and transitions over time in a controlled environment.

Following this, our attention will shift to geographic data, utilizing latitude and longitude coordinates in domain considerations. In this context, we explore how PaLD

can unravel complex community structures based on various attributes, including educational, financial, and cultural metrics. This analysis will underscore the method's potential in revealing spatially driven insights and patterns.

The discussion then advances to the realm of network science, specifically examining brain networks. Here, we leverage PaLD to discern local connectivity patterns, demonstrating its utility in furthering the understanding of intricate organization and function in neural systems.

To round off the chapter, we revisit the initial motivation for employing PaLD, contemplating its prospective applications in MD simulations. This section will showcase the potential benefits of the PaLD method's domain-integrated approach over traditional clustering techniques and explore potential directions for future research. For additional information and references pertinent to the data discussed in this chapter, see Appendix C.

5.2.1 Artificially constructed data

In this section, we present two illustrative examples, encompassing both one-dimensional (1-D) and two-dimensional (2-D) datasets, where the domain is temporal, and the feature space is characterized by a limited set of real-valued variables, assessed under Euclidean distance. The one-dimensional example, depicted in Figure 5.1, exemplifies the proficiency of the domain-integrated PaLD in identifying change points. Additionally, the community structure within the 1-D dataset is illustrated in Figure 5.2.

As an example for the two-dimensional data, Figure 5.3 demonstrates that the domain-integrated PaLD adeptly detects change points, even when these points are spatially intermingled. A more complex community structure is revealed in Figure 5.4. Additionally, Figure 5.5 presents the results of applying HDBSCAN to the 2-D

temporal data, with the analysis executed over a range of minimum point (minPts) thresholds. These results underscore the efficacy of the domain-integrated PaLD in accurately identifying change points.

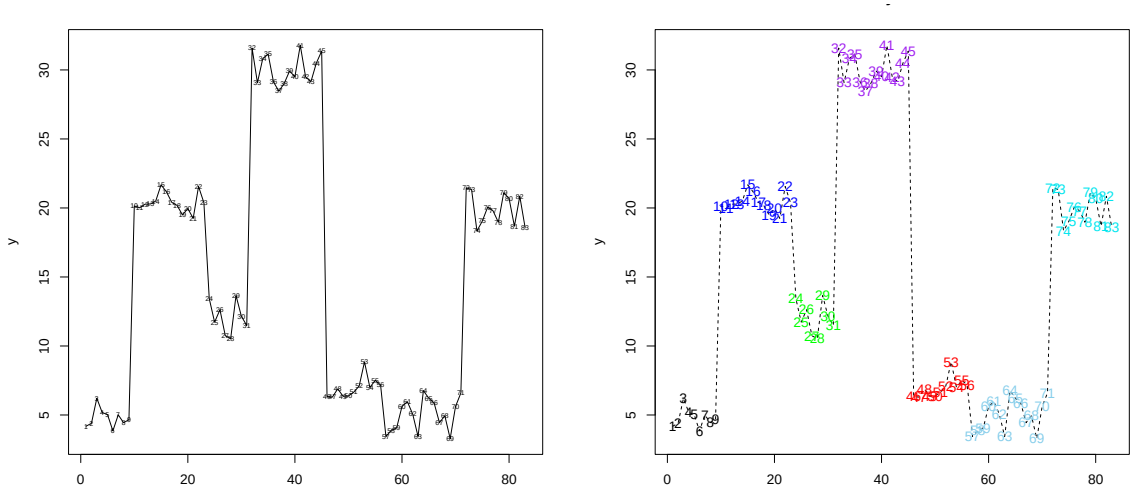


Figure 5.1: On the left panel, 1-D Euclidean data is plotted with variable y against time x. On the right panel, the clustering results are illustrated, with each community distinctly colored based on the domain-feature analysis.

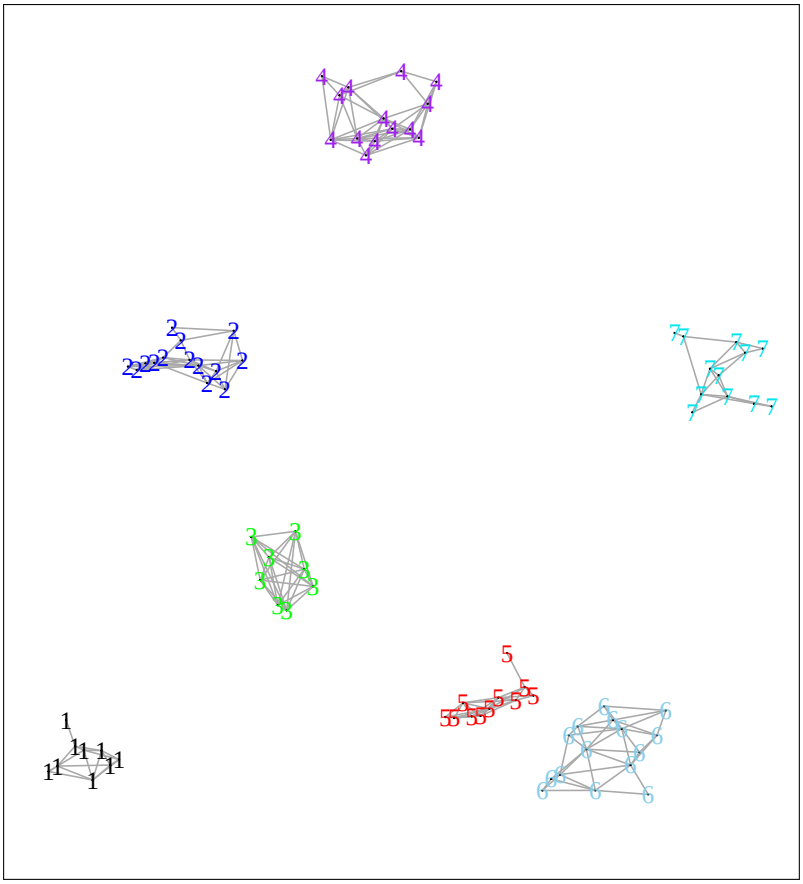


Figure 5.2: Community structure for 1-D data in time.

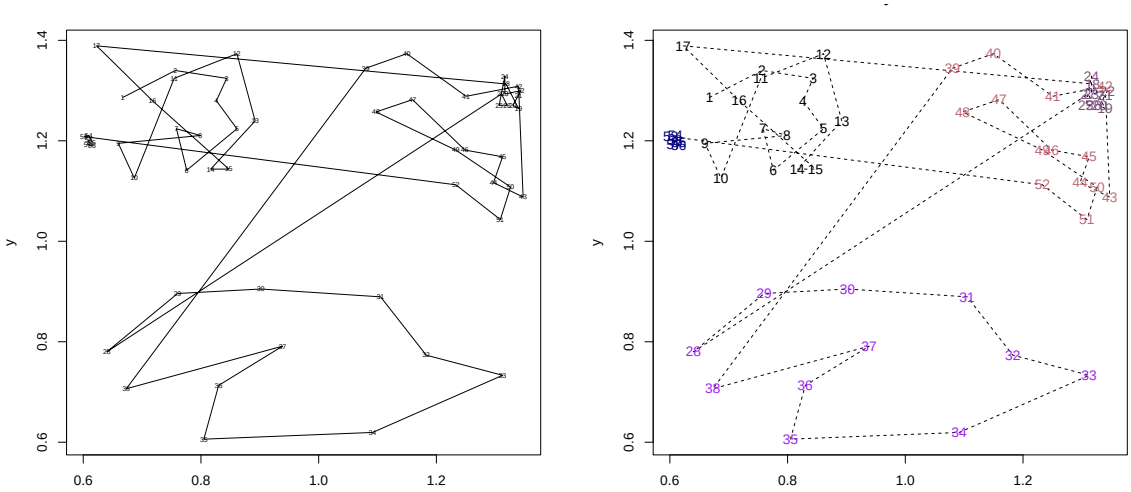


Figure 5.3: On the left panel, 2-D Euclidean data is plotted, with each point labeled according to its temporal order and connected to adjacent points in time by lines. On the right panel, clustering results are displayed, with each community distinctly colored according to the domain-feature analysis.

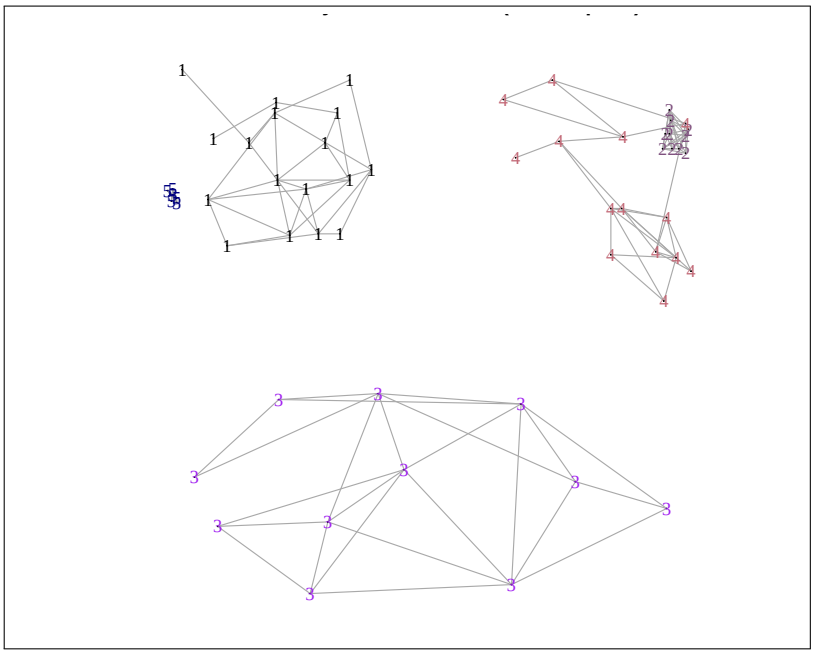


Figure 5.4: Community structure for 2-D data in time.

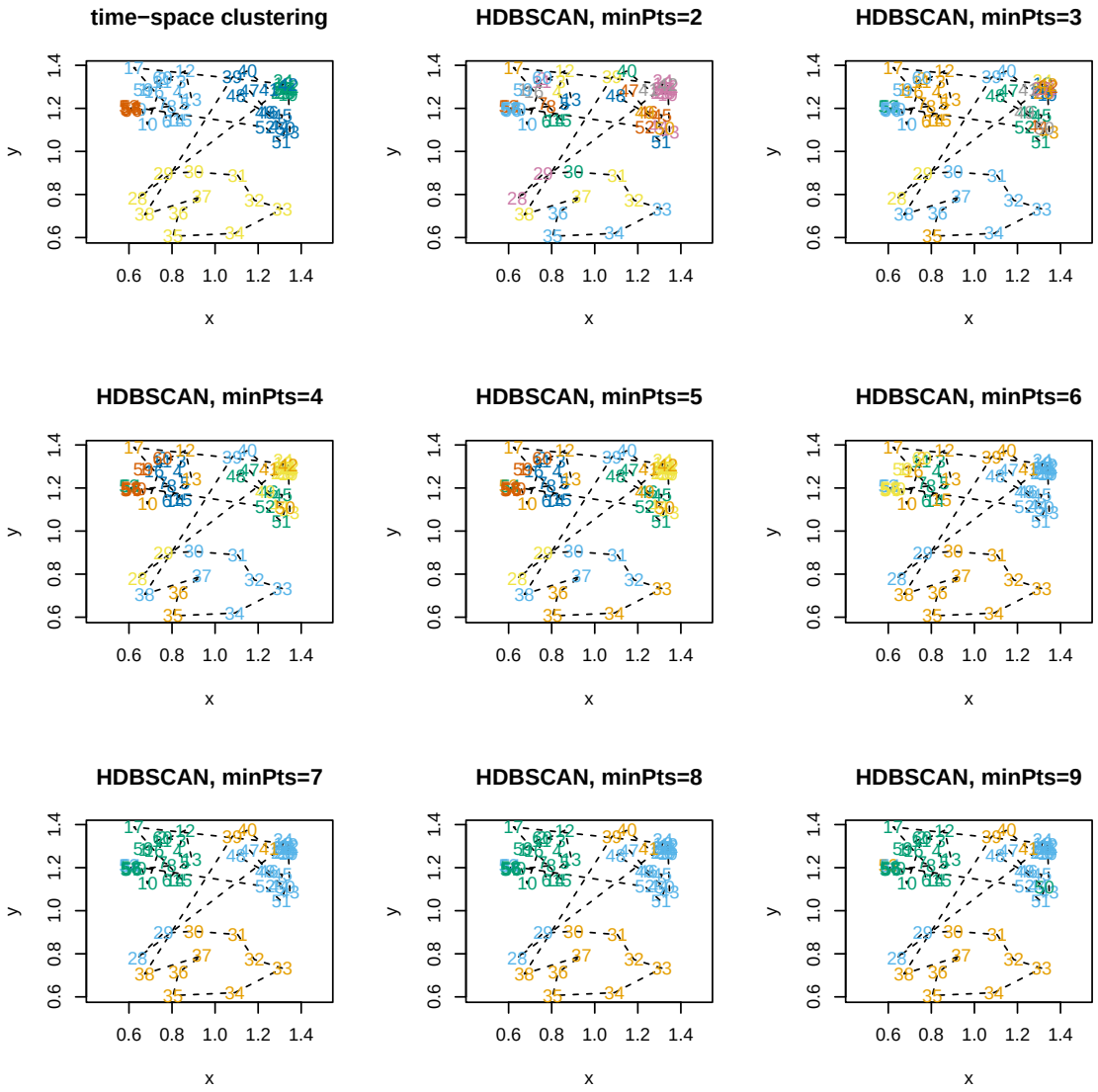


Figure 5.5: Results of HDBSCAN for 2-D data in time, under varying values of minPts.

5.2.2 Geographic data

In this section, we consider some examples of community structure wherein the domain is determined through geographic latitude-longitude data, with Haversine (great-circle) distance. A variety of point sets (for cities, states and countries) are considered for the domain, and distance in the feature space is determined for cultural,

financial, and educational data.

Figure 5.6 and Figure 5.7 present the community structure for population growth rates and housing prices across cities in China. The pattern of population growth does not exhibit a clear community structure, although small clusters with similar growth rates emerge, exemplified by cities like Chengdu, Chongqing, and Kunming. Housing prices, a perennially significant issue in China due to their persistent rise, show a more defined community structure, forming five main clusters: cities in the North China Plain, the Yangtze River Delta, the Pearl River Delta, the Sichuan Basin, and the mountainous regions around Guizhou Province. Major cities such as Beijing, Fuzhou, Xiamen, Shenzhen, Chengdu, and Chongqing are notably distinguished by their substantially higher housing prices. It is important to mention that the data for Shanghai's housing prices is absent; however, it is anticipated that Shanghai would similarly be distinguished if included.

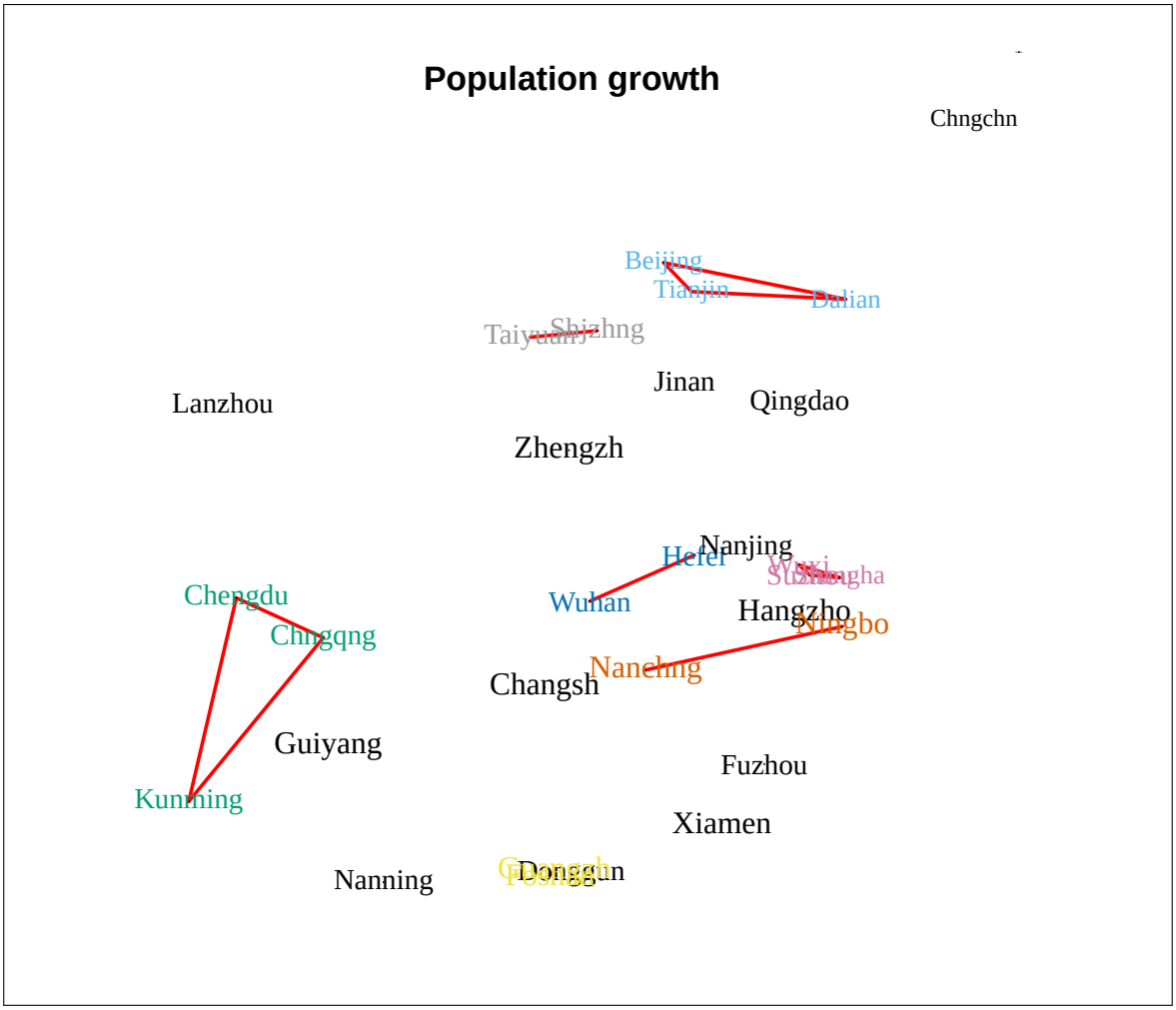


Figure 5.6: Community structure for population growth rates for selected cities in China.

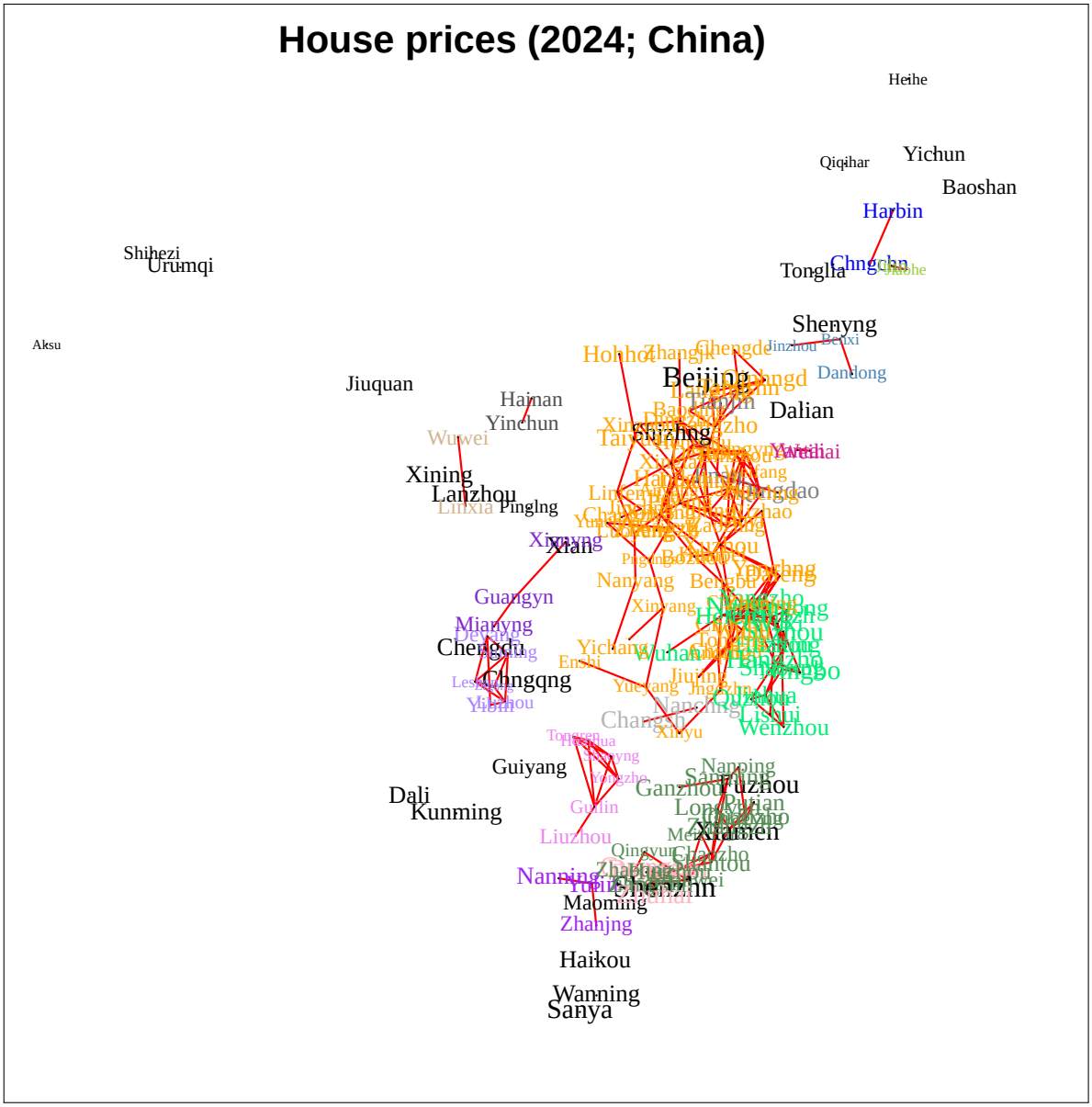


Figure 5.7: Community structure for house prices for cities in China.

Figure 5.8 and Figure 5.9 illustrate the community structure derived from cultural distances across multiple dimensions, encompassing a selection of countries and regions as well as an exhaustive list as considered at `culturaldistance.com` (see also [35]). Within the subset of European nations, the domain-integrated PaLD effectively segregates them into four primary clusters: Northern Europe, Central Eu-

rope, Western Europe, and South-Eastern Europe. This categorization is consistent with established geographical and cultural delineations. On a global scale, the algorithm articulates connections between Europe and the Middle East, and it clearly demarcates Western Africa, Central America, South America, and Southeast Asia as separate clusters, reflecting their distinct geographical and cultural identities. Remarkably, countries such as Canada, Russia, China, India, Indonesia, and Australia are identified as standalone entities, a phenomenon possibly linked to their large territorial expanse. The domain distance between these countries and others may be influenced by the methodological choice to compute distances using the centroid coordinates of each country. Other options are possible, potentially incorporating population density or transportation distances.

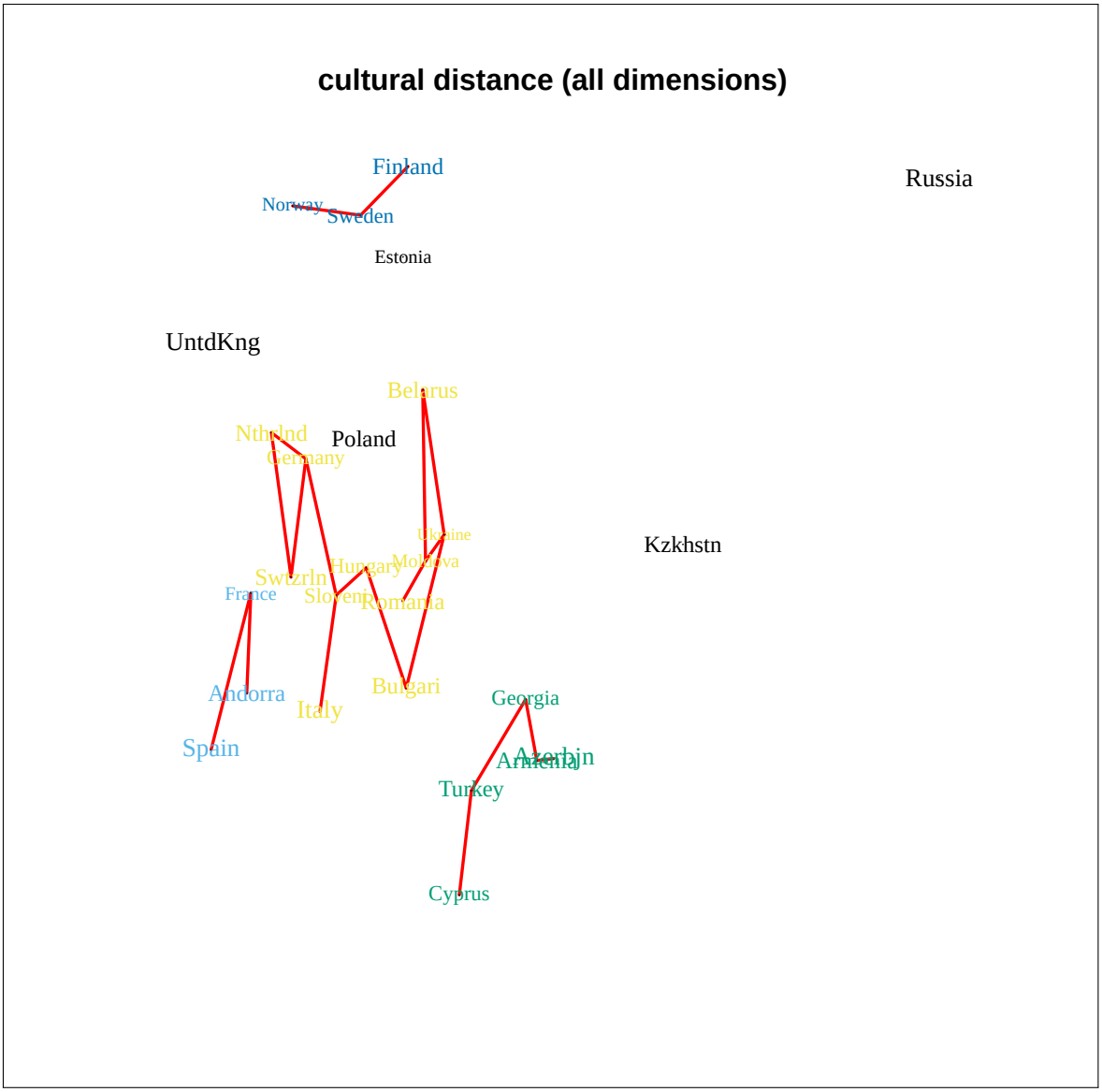


Figure 5.8: Community structure for cultural distance over all dimensions, for the selected countries or regions as given at http://culturaldistance.com/; see also [35].

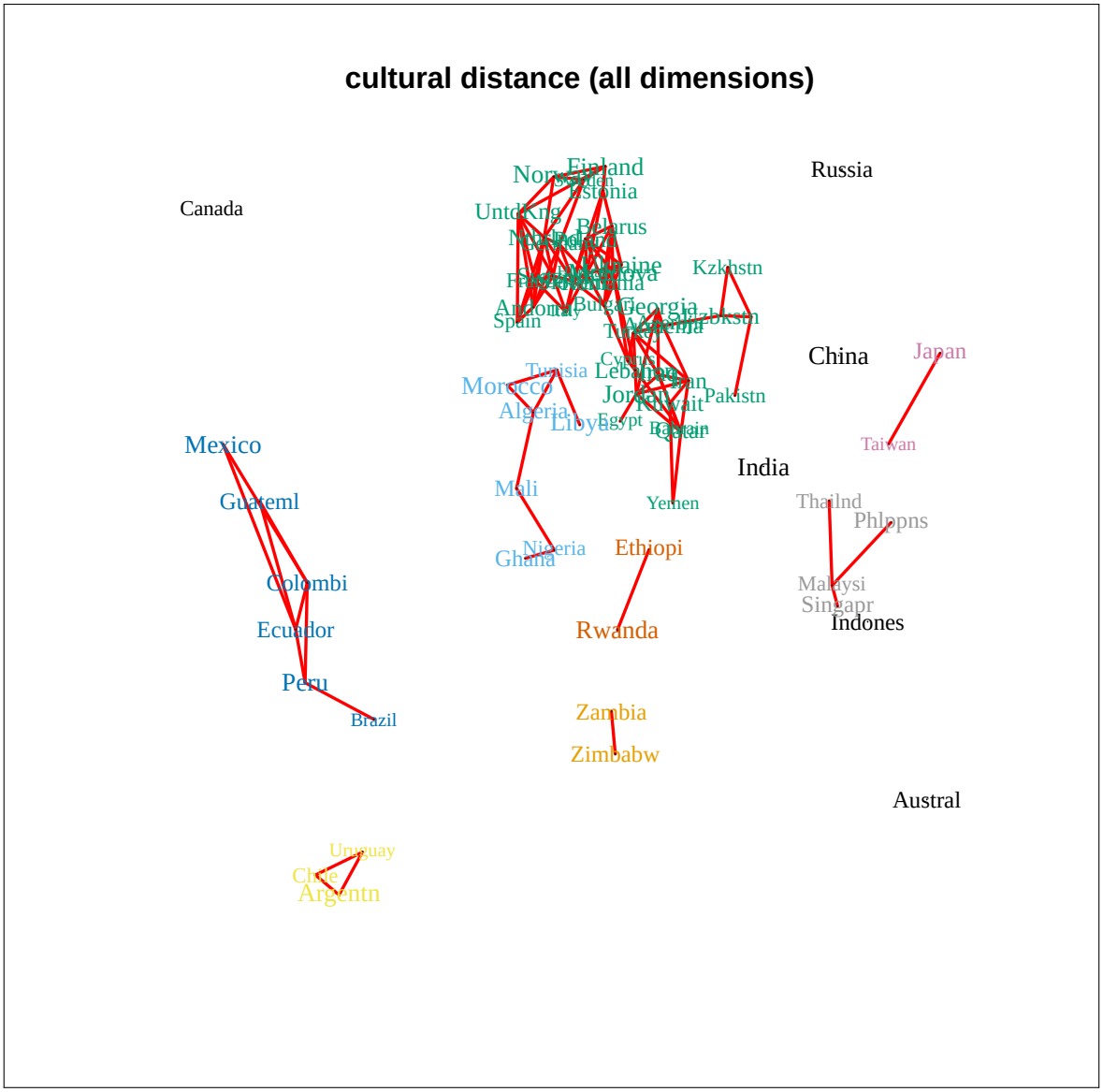


Figure 5.9: Community structure for cultural distance over all dimensions for the full list of countries or regions as given at http://culturaldistance.com/; see also [35].

Figures 5.10, 5.11, and 5.12 depict community structure based on state-level data in the United States, reflecting the nation's cultural and geographical nuances. The features data, here, is from a variety of U.S. government agencies in the 1970s and 2010s. This data is available in the `state.x77` data frame in `R` and the `facts` data frame from the `usa` package. Specifically, Figure 5.10, and 5.11 illustrate the rela-

tionship between per capita income and life expectancy, organizing states primarily into three regions: Northeast, Southeast, and Central. This distribution indicates a probable association between the economic and health metrics within these distinct geographical zones.

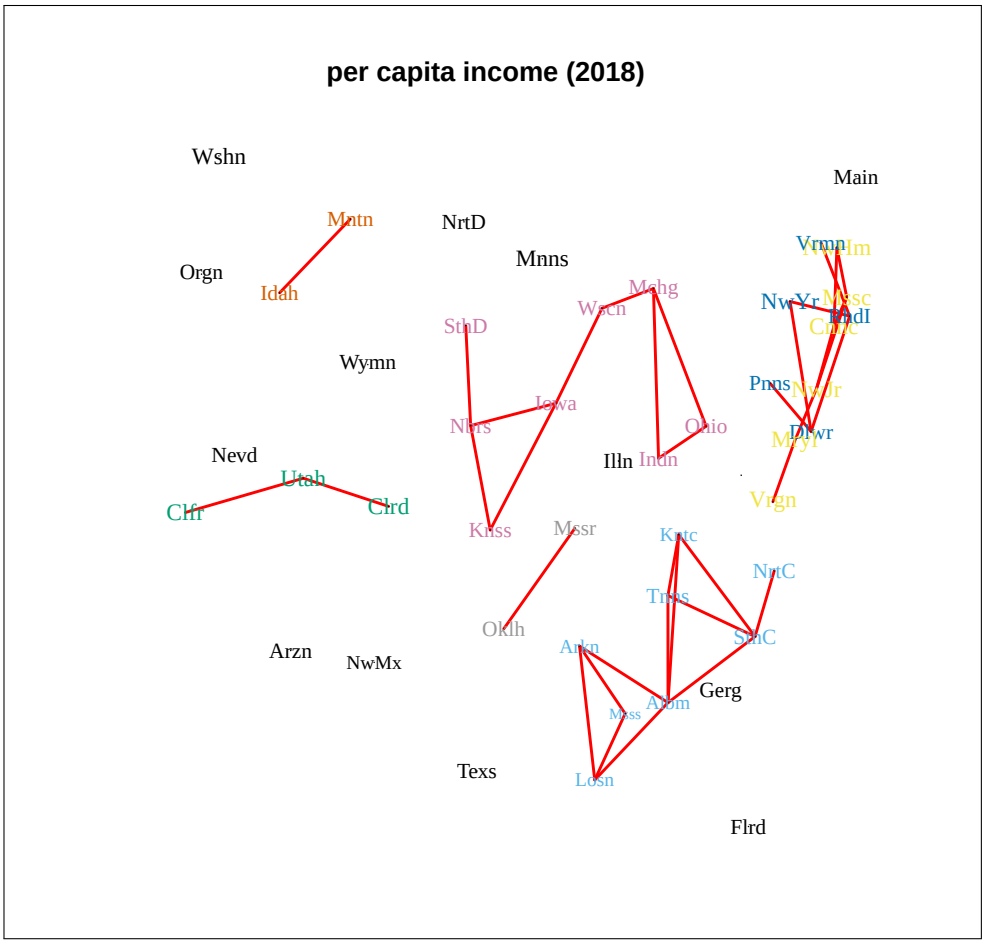


Figure 5.10: Community structure for per capita income for states in the U.S.

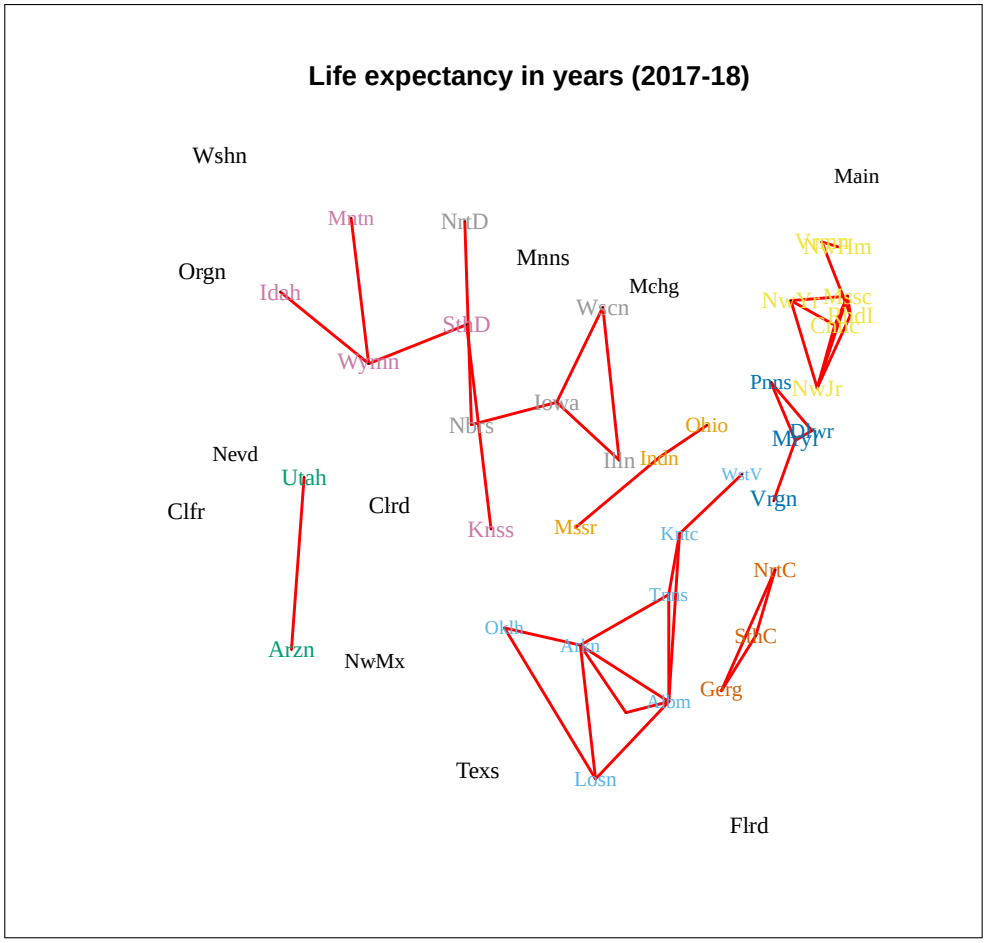


Figure 5.11: Community structure for life expectancy for states in the U.S.

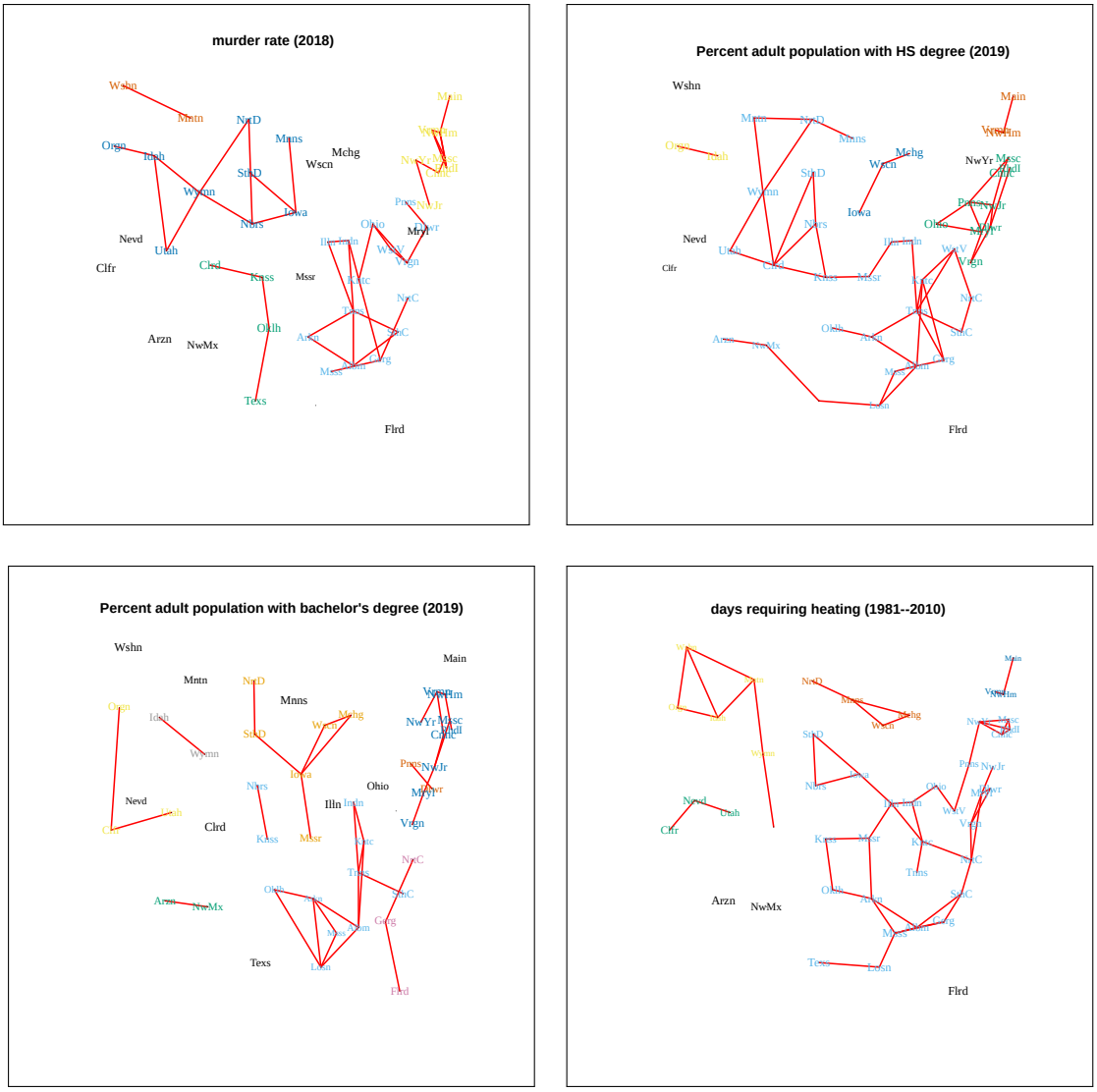


Figure 5.12: Community structure for US geographic (state) features data

In Appendix A, we provide several additional examples of community structure as it arises in geographic data.

5.2.3 Brain networks and local connectivity

In this section, we explore the use of a domain-features approach for analyzing local connectivity in f-MRI correlation brain network data (employing data from [51]).

The analysis is based on time series data from 268 voxels, aligned with the Shen parcellation scheme [43]. The spatial domain is defined using Euclidean distance on the voxel coordinates (x, y, z), whereas the feature space distance is ascertained through the shortest-path distance in the network. This network, here, is derived by thresholding to retain only the highest 1444 pairwise correlations.

Figures 5.13 and 5.15 demonstrate application of a domain-integrated approach can elucidate distinct and significant structures in the complex network, underscoring local connectivity patterns. The exploration of these networks' value and their implications for understanding brain connectivity is earmarked for future work. It is noteworthy that this method is flexible and can be adapted for a wide array of network data types where the nodes have positioning in a physical setting.

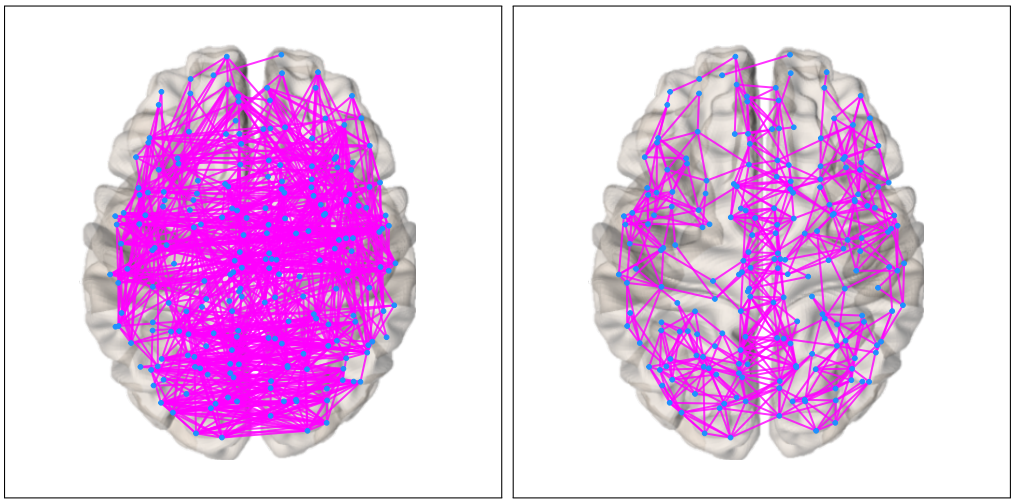


Figure 5.13: Example of brain network structure obtained directly via thresholded cohesion (left) and via a domain features approach integrating voxel locations (right). Note the enhanced local structure.

For comparison, the results of applying standard PaLD to the voxel distances are given in Figure 5.14.

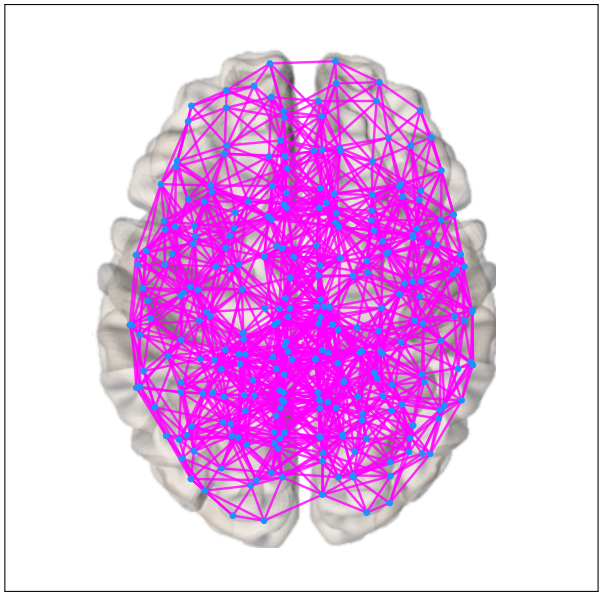


Figure 5.14: Example of brain network structure obtained via thresholded cohesion based on voxel locations.

Corresponding results for the same individual (this time at rest) are given below.

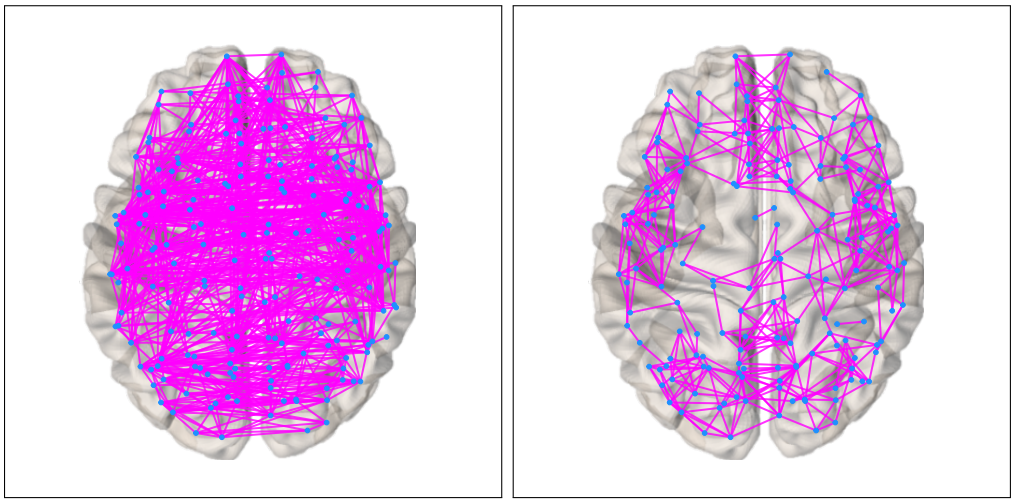


Figure 5.15: Example of brain network structure obtained directly via thresholded cohesion (left) and via a domain-features approach integrating voxel locations (right). Note the enhanced local structure.

Note the sharp difference in local connectivity across hemispheres, for the two

resulting domain-features networks, which is not readily apparent when examing the correlation networks, directly.

Chapter 6

Conclusions and future directions

In this thesis, we first reviewed advancements and applications of clustering techniques in the analysis of MD simulations, underscoring their limitations; these include parameter dependence and the obscuring of inter- and intra-cluster structures. The PaLD methodology introduced in [4] provides a significant shift towards a more nuanced and parameter-independent analysis of complex datasets. By prioritizing local relationships and cohesion among data points, PaLD offers a flexible and insightful framework for uncovering community structures within diverse datasets.

Here we extend PaLD to incorporate domain-specific information. This opens new avenues for its application, allowing for more contextually relevant analyses. The adaptation is particularly effective in fields such as time-series analysis, geospatial data interpretation, and network analysis, where the integration of domain characteristics can have significant impact on the accuracy and meaningfulness of data clustering.

The approach suggested here provides significant opportunities to analysis in MD simulations and elsewhere. The capacity of the approach to identify subtle community structures autonomously makes it a valuable preliminary tool, even in the absence of prior knowledge. The resulting nuanced understanding of both inter- and intra-cluster dynamics promises to deepen insights into the dynamic behaviors and functional

mechanisms of biomolecules and other entities.

Future research can focus on refining the algorithm to enhance its performance and accuracy with large-scale datasets, a crucial step. Recent advancements, such as the acceleration of PaLD introduced in [13] and approximations such as those in [2], indicate promising progress in this direction. Continued development, extension and application of the general PaLD methodology are poised to make substantial contributions to data science and related fields. Keeping in mind that the PaLD methodology is free of parameters, iteration, and distributional assumptions, the progress here suggests continued work to identify other potential extensions in the study of community structure in data.

Bibliography

[1] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[2] Jacob D Baron, RWR Darling, J Laylon Davis, and R Pettit. Partitioned k-nearest neighbor local depth for scalable comparison-based learning. *arXiv preprint arXiv:2108.08864*, 2021.

[3] Kenneth S Berenhaut, John D Foley, and Liangdongsheng Lyu. Generalized partitioned local depth. *Journal of Statistical Theory and Practice*, 18(1):1–17, 2024.

[4] Kenneth S Berenhaut, Katherine E Moore, and Ryan L Melvin. A social perspective on perceived distances reveals deep community structure. *Proceedings of the National Academy of Sciences*, 119(4):e2003634119, 2022.

[5] Pavel Berkhin. A survey of clustering data mining techniques. In *Grouping multidimensional data*, pages 25–71. Springer, 2006.

[6] Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. *Proceedings of the 17th Pacific-Asia conference on knowledge discovery and data mining (PAKDD 2013)*, pages

160–172, 2013.

[7] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[8] Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *arXiv preprint arXiv:1710.04019*, 2017.

[9] Mark Ming-Tso Chiang and Boris Mirkin. Intelligent choice of the number of clusters in k-means clustering: an experimental study with different cluster spreads. *Journal of classification*, 27:3–40, 2010.

[10] Xavier Daura, Karl Gademann, Bernhard Jaun, Dieter Seebach, Wilfred F van Gunsteren, and Alan E Mark. Peptide folding: When simulation meets experiment. *Angewandte Chemie International Edition*, 38(1-2):236–240, 1999.

[11] Renato Cordeiro De Amorim and Christian Hennig. Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Information sciences*, 324:126–145, 2015.

[12] Renato Cordeiro De Amorim and Boris Mirkin. Minkowski metric, feature weighting and anomalous cluster initializing in k-means clustering. *Pattern Recognition*, 45(3):1061–1075, 2012.

[13] Aditya Devarakonda and Grey Ballard. Sequential and shared-memory parallel algorithms for partitioned local depths. *arXiv preprint arXiv:2307.16652*, 2023.

[14] Ron O Dror, Robert M Dirks, J P Grossman, Huafeng Xu, and David E Shaw. Biomolecular simulation: a computational microscope for molecular biology. *Annual review of biophysics*, 41:429–452, 2012.

[15] Sean R Eddy. Hidden markov models. *Current opinion in structural biology*, 6(3):361–365, 1996.

[16] Herbert Edelsbrunner and John Harer. Persistent homology—a survey. *Contemporary mathematics*, 453:257–282, 2008.

[17] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231. AAAI Press, 1996.

[18] Aldo Glielmo, Brooke E Husic, Alex Rodriguez, Cecilia Clementi, Frank Noé, and Alessandro Laio. Unsupervised learning methods for molecular simulation data. *Chemical Reviews*, 121(16):9722–9758, 2021.

[19] Ryan Godwin, William Gmeiner, and Freddie R Salsbury Jr. Importance of long-time simulations for rare event sampling in zinc finger proteins. *J. Biomol. Struct. Dyn.*, 34(1):125–134, 2016.

[20] Ryan C Godwin, Ryan Melvin, and Freddie R Salsbury. Molecular dynamics simulations and computer-aided drug discovery. In *Computer-aided drug discovery*, pages 1–30. Springer, 2015.

[21] Teofilo F Gonzalez. Clustering to minimize the maximum intercluster distance. In *Proceedings of the seventeenth annual ACM symposium on Theory of computing*, pages 293–306. ACM, 1985.

[22] Felix Hensel, Michael Moor, and Bastian Rieck. A survey of topological machine learning methods. *Frontiers in Artificial Intelligence*, 4:681108, 2021.

[23] Brooke E Husic and Vijay S Pande. Markov state models: From an art to a science. *Journal of the American Chemical Society*, 140(7):2386–2396, 2018.

[24] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

[25] Anil K. Jain, M. Narasimha Murty, and Patrick J. Flynn. Data clustering: A review. *ACM Computing Surveys (CSUR)*, 31(3):264–323, 1999.

[26] Martin Karplus and J Andrew McCammon. Molecular dynamics simulations of biomolecules. *Nature structural biology*, 9(9):646–652, 2002.

[27] Leonard Kaufman and Peter J Rousseeuw. Clustering by means of medoids. *Statistical Data Analysis Based on the L1 Norm and Related Methods*, pages 405–416, 1987.

[28] Jun Koseki, Shuto Hayashi, Yasuhiro Kojima, Haruka Hirose, and Teppei Shimamura. Topological data analysis of protein structure and inter/intra-molecular interaction changes attributable to amino acid mutations. *Computational and Structural Biotechnology Journal*, 21:2950–2959, 2023.

[29] Thomas J Lane, Diwakar Shukla, Kyle A Beauchamp, and Vijay S Pande. Milliseconds of molecular dynamics simulations on anton. *Proceedings of the National Academy of Sciences*, 110(32):12716–12720, 2013.

[30] James B MacQueen. Some methods for classification and analysis of multivariate observations. 1(14):281–297, 1967.

[31] J Andrew McCammon, Bruce R Gelin, and Martin Karplus. Dynamics of folded proteins. *nature*, 267(5612):585–590, 1977.

[32] Leland McInnes and John Healy. Accelerated hierarchical density based clustering. In *2017 IEEE international conference on data mining workshops (ICDMW)*, pages 33–42. IEEE, 2017.

[33] Ryan L Melvin, Ryan C Godwin, Jiajie Xiao, William G Thompson, Kenneth S Berenhaut, and Freddie R Salsbury Jr. Uncovering large-scale conformational change in molecular dynamics without prior knowledge. *J. Chem. Theory Comput.*, 12(12):6130–6146, 2016.

[34] Ryan L Melvin and Freddie R Salsbury Jr. Visualizing ensembles in structural biology. *J. Mol. Graphics Modell.*, 67:44–53, 2016.

[35] M. Muthukrishna, A.V. Bell, J. Henrich, C.M. Curtin, A. Gedranovich, J. McInerney, and B. Thue. Beyond western, educated, industrial, rich, and democratic (weird) psychology: Measuring and mapping scales of cultural and psychological distance. *Psychol. Sci.*, 31:678—-701, 2020.

[36] Frank Noé and Feliks Nüske. A variational approach to modeling slow processes in stochastic dynamical systems. *Multiscale Modeling & Simulation*, 11(2):635–655, 2013.

[37] Frank Noé, Hao Wu, Jan-Hendrik Prinz, and Nuria Plattner. Projected and hidden markov models for calculating kinetics and metastable states of complex molecules. *The Journal of chemical physics*, 139(18), 2013.

[38] Jun-hui Peng, Wei Wang, Ye-qing Yu, Han-lin Gu, and Xuhui Huang. Clustering algorithms to analyze molecular dynamics simulation trajectories for complex chemical and biological systems. *Chinese Journal of Chemical Physics*, 31(4):404–420, 2018.

[39] Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.

[40] Chandrabose Selvaraj, Umesh Panwar, Dhurvas Chandrasekaran Dinesh, Evzen Boura, Poonam Singh, Vikash Kumar Dubey, and Sanjeev Kumar Singh. Mi-

crosecond md simulation and multiple-conformation virtual screening to identify potential anti-covid-19 inhibitors against sars-cov-2 main protease. *Frontiers in Chemistry*, 8:595273, 2021.

[41] Jie Shao, Steven W Tanner, Natalie Thompson, and Thomas E Cheatham. Clustering molecular dynamics trajectories: 1. characterizing the performance of different clustering algorithms. *Journal of Chemical Theory and Computation*, 3(6):2312–2334, 2007.

[42] David E Shaw, Paul Maragakis, Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, Michael P Eastwood, Joseph A Bank, John M Jumper, John K Salmon, Yibing Shan, et al. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346, 2010.

[43] Xilin Shen, Fuyuze Tokoglu, Xenios Papademetris, and R Todd Constable. Groupwise whole-brain parcellation from resting-state fmri data for network node identification. *Neuroimage*, 82:403–415, 2013.

[44] Gurjeet Singh, Facundo Mémoli, and Gunnar Carlsson. Topological methods for the analysis of high dimensional data sets and 3d object recognition. In *SPBG*, pages 91–100, 2007.

[45] Florian Sittel, Aashish Jain, and Gerhard Stock. Robust density-based clustering to identify metastable conformational states of proteins. *The Journal of Chemical Physics*, 141(24):244101, 2014.

[46] Peter HA Sneath. The application of computers to taxonomy. *Microbiology*, 17(1):201–226, 1957.

[47] Robert R Sokal and Charles D Michener. A statistical method for evaluating systematic relationships. 1958.

[48] Thorvald Sorensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske skrifter*, 5:1–34, 1948.

[49] Terra Sztain, Surl-Hee Ahn, Anthony T Bogetti, Lorenzo Casalino, Jory A Goldsmith, Evan Seitz, Ryan S McCool, Fiona L Kearns, Francisco Acosta-Reyes, Suvrajit Maji, et al. A glycan gate controls opening of the sars-cov-2 spike protein. *Nature chemistry*, 13(10):963–968, 2021.

[50] Wai Shing Tang, Gabriel Monteiro da Silva, Henry Kirveslahti, Erin Skeens, Bibo Feng, Timothy Sudijono, Kevin K Yang, Sayan Mukherjee, Brenda Rubenstein, and Lorin Crawford. A topological data analytic approach for discovering biophysical signatures in protein dynamics. *PLoS computational biology*, 18(5):e1010045, 2022.

[51] David C Van Essen, Kamil Ugurbil, Edward Auerbach, Deanna Barch, Timothy EJ Behrens, Richard Bucholz, Acer Chang, Liyong Chen, Maurizio Corbetta, Sandra W Curtiss, et al. The human connectome project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231, 2012.

[52] Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

[53] Dizhou Wu and Freddie R. Salsbury. Unraveling the role of hydrogen bonds in thrombin via two machine learning methods. *Journal of Chemical Information and Modeling*, June 2023.

[54] Dizhou Wu and Freddie Salsbury Jr. Unraveling the impact of w215a/e217a mutations on thrombin dynamics and thrombomodulin binding through molecular dynamics simulations. *bioRxiv*, pages 2023–12, 2023.

[55] Dizhou Wu and Freddie R Salsbury Jr. Simulations suggest double sodium binding induces unexpected conformational changes in thrombin. *Journal of Molecular Modeling*, 28(5):120, 2022.

[56] Dizhou Wu, Jiajie Xiao, and Freddie R Salsbury Jr. Light chain mutation effects on the dynamics of thrombin. *Journal of chemical information and modeling*, 61(2):950–965, 2021.

[57] Jiajie Xiao, Ryan L Melvin, and Freddie R Salsbury. Mechanistic insights into thrombin's switch between "slow" and "fast" forms. *Phys. Chem. Chem. Phys.*, 19(36):24522–24533, 2017.

[58] Jiajie Xiao, Ryan L Melvin, and Freddie R Salsbury Jr. Probing light chain mutation effects on thrombin via molecular dynamics simulations and machine learning. *J. Biomol. Struct. Dyn.*, 37(4):982–999, 2019.

[59] Jiajie Xiao and Freddie R Salsbury. Molecular dynamics simulations of aptamer-binding reveal generalized allostery in thrombin. *J. Biomol. Struct. Dyn.*, 35(15):3354–3369, 2017.

[60] Jiajie Xiao and Freddie R Salsbury. Na+-binding modes involved in thrombin's allosteric response as revealed by molecular dynamics simulations, correlation networks and markov modeling. *Phys. Chem. Chem. Phys.*, 21(8):4320–4330, 2019.

[61] Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.

Appendix A

Additional applications of community structure in domain-features data

In this appendix, we provide some additional examples of community structure in domain features data. For details and references see Appendix C.

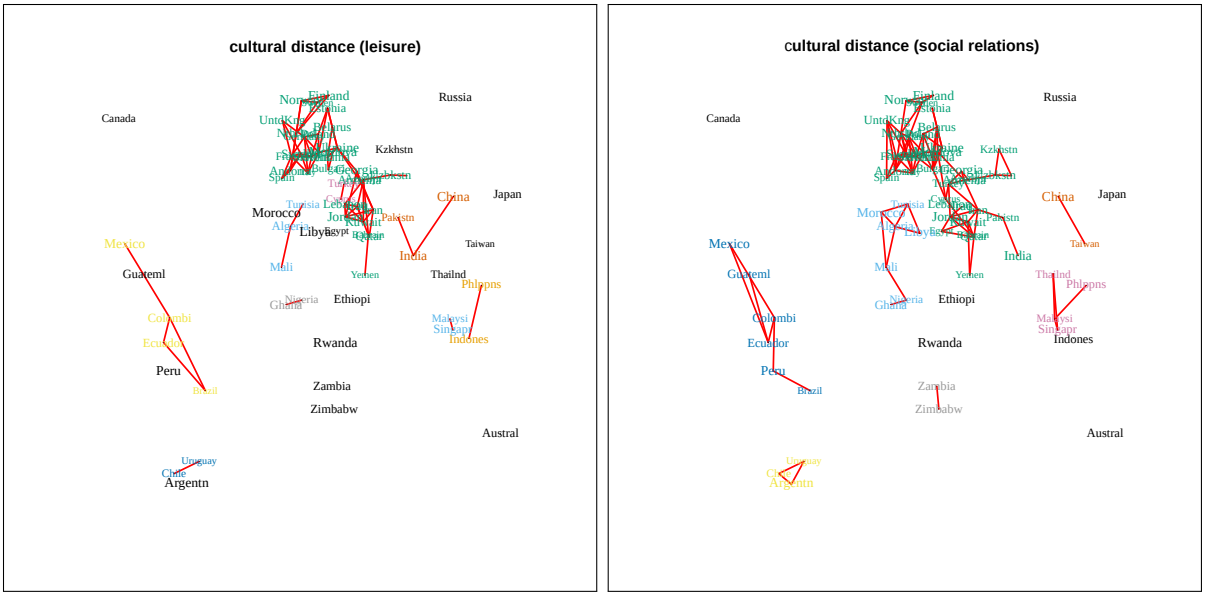


Figure A.1: Community structure for cultural distance over dimensions, for the full list of countries as given at culturaldistance.com

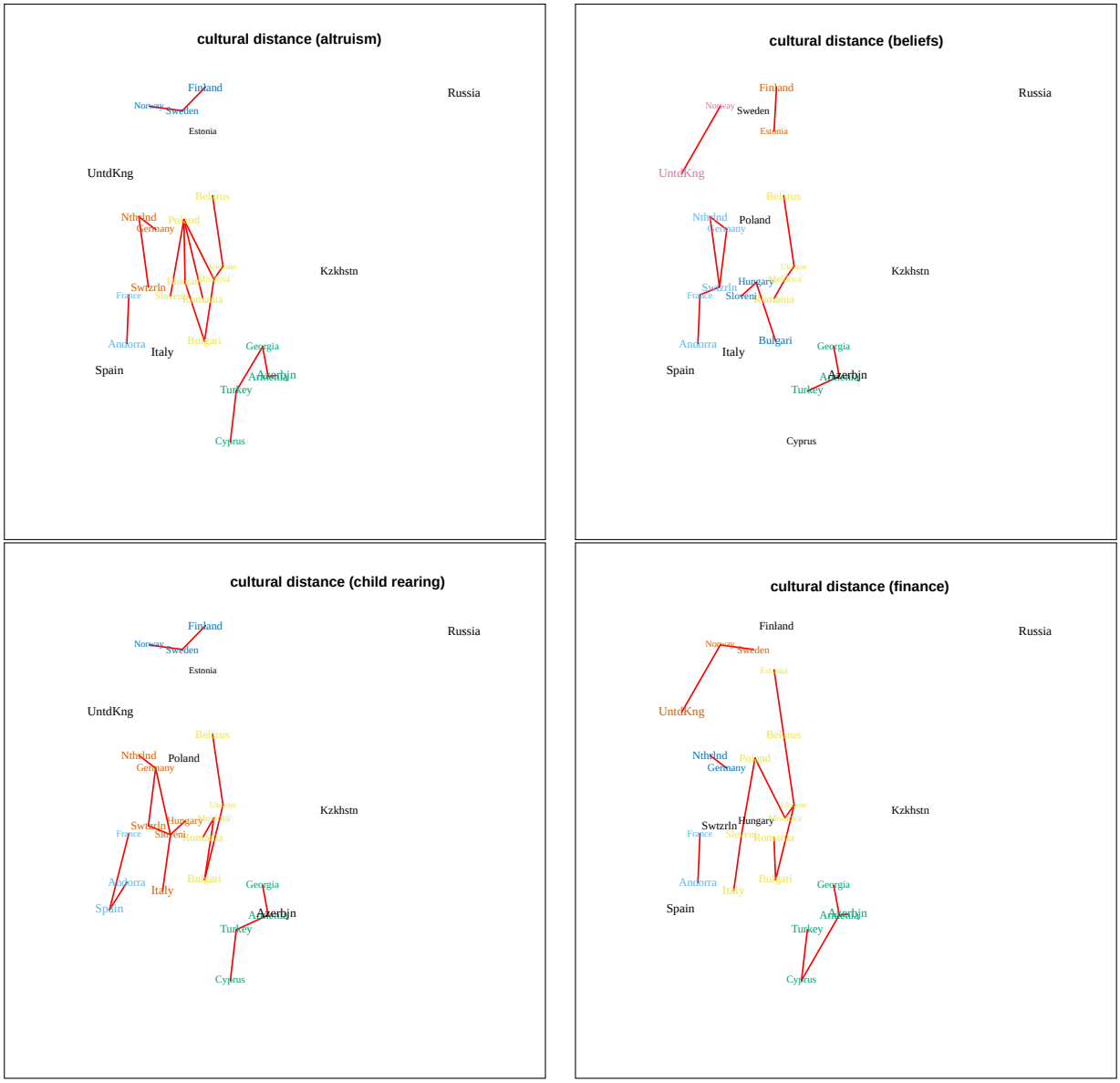


Figure A.2: Community structure for cultural distance over all dimensions, for selected countries as given at culturaldistance.com

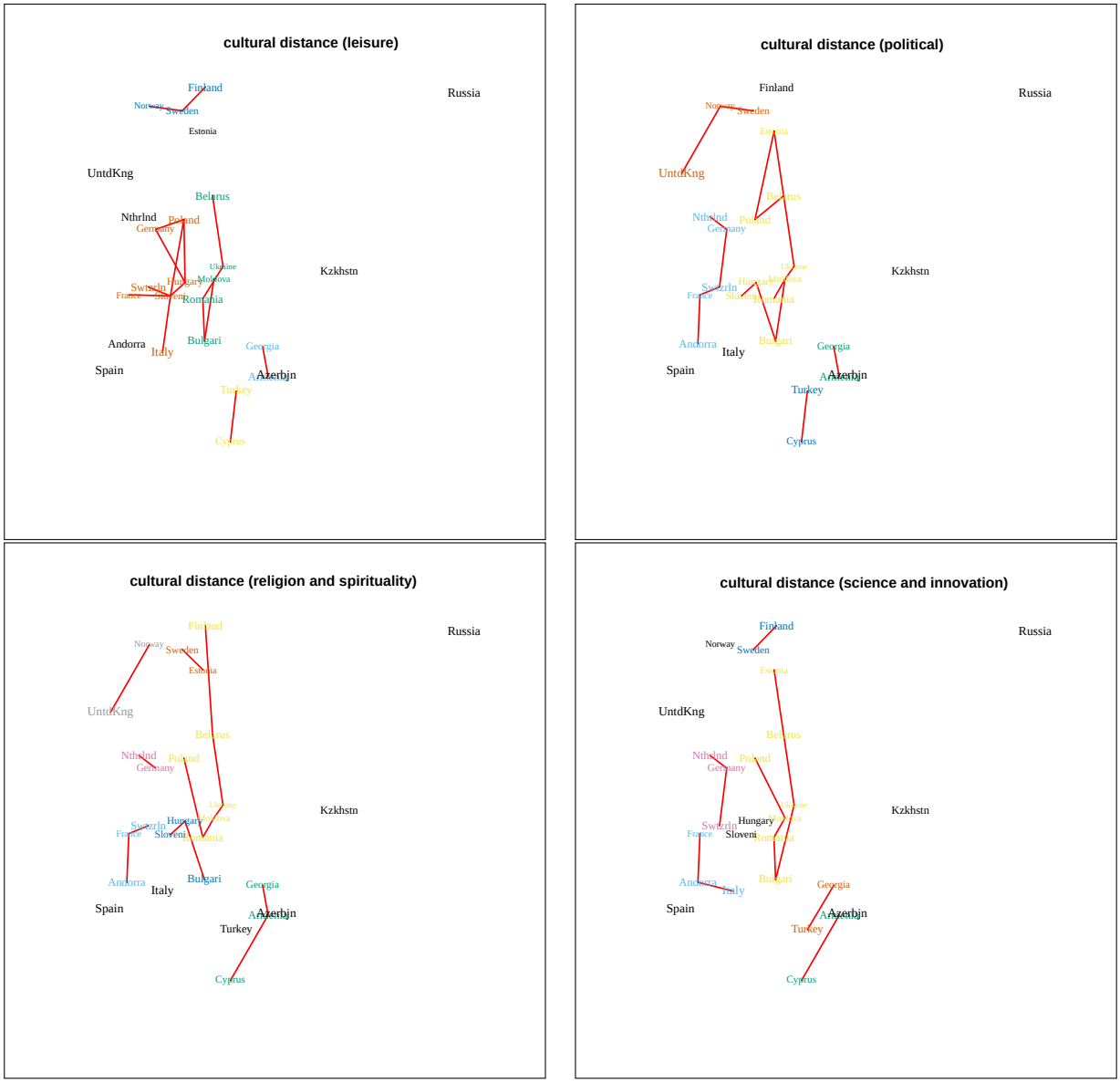


Figure A.3: Community structure for cultural distance over all dimensions, for selected countries as given at culturaldistance.com

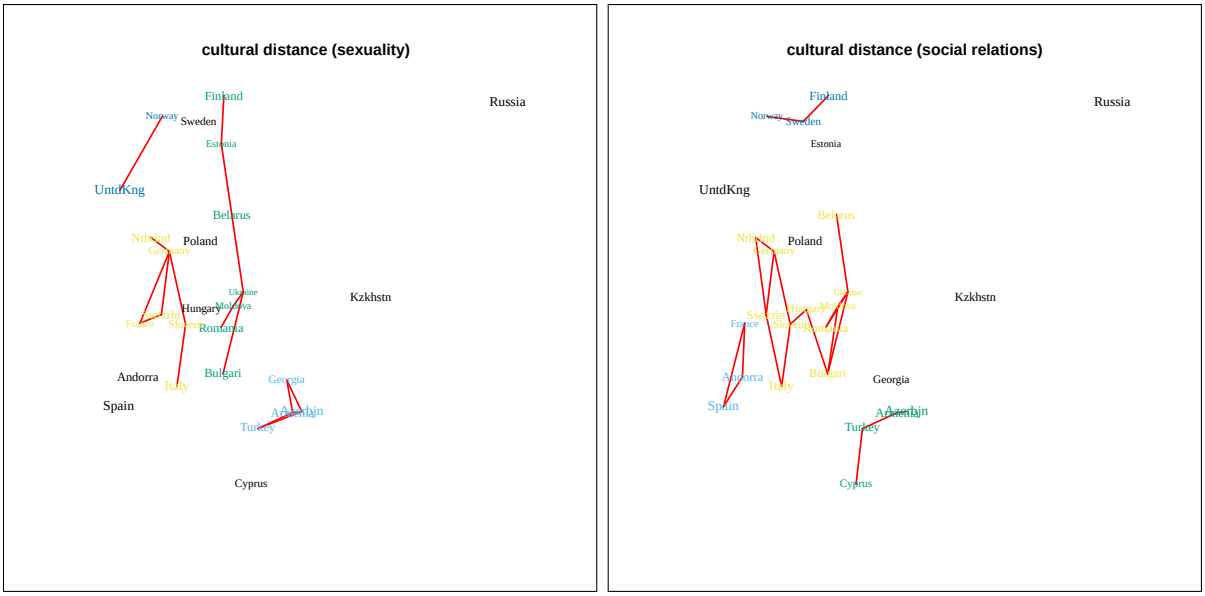


Figure A.4: Community structure for cultural distance over all dimensions, for selected countries as given at culturaldistance.com

Appendix B

R code

```
# Takes in a distance (or dissimilarity) matrix and
   outputs the local relevance and support division arrays
getRQ<-
function(D,r=1){
  n = nrow(D)
  R = array(0,dim=c(n,n,n))
  Q = array(0,dim=c(n,n,n))
  for(i in 1:n){
    for(j in 1:n){
      for(z in 1:n){
        if(D[z,i]<=r*D[j,i] | D[z,j]<=r*D[i,j]) {R[i,j,z]
           = 1}else{R[i,j,z] = 0}
        if(D[z,i]<D[j,z]) {
          Q[i,j,z] = 1
        }else{ if(D[z,i]==D[j,z]){
          Q[i,j,z] = .5 }else{Q[i,j,z] = 0}}
      }
    }
  }
  list(R=R,Q=Q)
}

# Takes in local relevance and support division arrays and
    outputs the resulting cohesion matrix
paldrq<-
function(R,Q){
  n = dim(Q)[1]
  A = matrix(0,n,n)
  for(i in 1:n){
```

```
    for(j in 1:n){
      if(j==i){next};
      Rij = sum(R[i,j,])
      for(z in 1:n){
        A[i,z] = A[i,z] + (R[i,j,z] / Rij) * Q[i,j,z]
      }
    }
  }
  C = A /(n-1)
  C
}

# Takes in a distance (or dissimilarity) matrix and
   outputs the resulting (domain-features) strong cohesion
    # network along with intermediary information. This
   code also provides plotting capabilities.

timeclust2dists<-
function(D, D2,T=cbind(1:nrow(D)),C_plot=FALSE,ord=1,show_
   plot=TRUE,self=FALSE)
  {
  n2<-nrow(D);n2
  q<-1:n2

  if (C_plot) {dev.new();myImagePlot(D);title("Distance")}

#  if(show_plot){
#    dev.new();plot(N2,type="l",xlab="x",ylab="y")
#    text(N2,labels=1:nrow(N2),cex=.5)
#  }
  if(ord==1){
    RQD<-getRQ(D)
    RQD2<-getRQ(D2)} else {
      RQD<-getRQ(D2)
      RQD2<-getRQ(D)}

  C1<-cohesion_matrix(D)
  if (C_plot) {dev.new();myImagePlot(C1);title("cohesion␣(
     usual)")}

  C1b<-pmin(C1,t(C1))
  C1b[C1b<mean(diag(C1b)/2)]<-0
```

```
if (C_plot) {dev.new();myImagePlot(C1b);title("cohesion␣
   (usual;␣thresholded)")}

C<-paldrq(RQD2$R,RQD$Q)
rownames(C)<-q
if (C_plot) {dev.new();myImagePlot(C);title("cohesion␣
   (2-1)")}

Cb<-pmin(C,t(C))
if (!self)  {Cb[Cb<mean(diag(Cb)/2)]<-0} else {
  Cb[Cb<0.5*matrix(diag(Cb),nrow(Cb),nrow(C))]<-0}
if (C_plot) {dev.new();myImagePlot(Cb);title("cohesion␣
   (2-1;␣thresholded)")}

Cb2<-Cb
#  Cb2[abs(row(Cb2)-col(Cb2))>1]<-0

gb<-graph.adjacency(Cb2,mode="undirected",weight=TRUE)
gb<-simplify(gb)


cl<-clusters(gb)$membership;cl
clp<-rev(cummin(rev(cl)))

### Just in case inputs are higher dimensional (not used
    in this 2D version)

#PN<-isoMDS(D,k=1)$points
#PT<-isoMDS(D2,k=1)$points

### Depths (not implemented here)

#dep<-apply(C,1,sum)
#dev.new();plot(T,dep,type="l");abline(h=.5)

#lay<-cbind(PT,PN)

###

cols<-clp

u <- max(cols)
```

```
  #Some potential colours

  pal <- sample(c("pink1", "violet", "mediumpurple1", "
     slateblue1", "purple", "purple3",
                  "turquoise2", "skyblue", "steelblue", "
                     blue2", "navyblue",
                  "orange", "tomato", "coral2", "
                     palevioletred", "violetred", "red2",
                  "springgreen2", "yellowgreen", "
                     palegreen4",
                  "wheat2", "tan", "tan2", "tan3", "brown"
                     ,
                  "grey70", "grey50", "grey30"))

  palette<-colorRampPalette(c("black","blue","orange","
     purple","red","skyblue","navyblue"))(u)


#  dev.new();pm(3,3)
#  plot(N2,type="l",main="time-space clustering",asp=1,lty
   =2,xlab="x",ylab="y")
#  text(N2,labels=1:nrow(N),cex=1,col=clp+1)
#  for(a in c(2:9)){
#    plot(N2,type="l",xlab="x",ylab="y",main=paste(c("
   HDBSCAN, minPts=",a),collapse=""),asp=1,lty=2)
#    text(N2,labels=1:nrow(N),cex=1,col=hdbscan(N,minPts=a
   )$cluster+1)}


#  dev.new();pm(1,1)
#  plot(N2,type="l",lty=2,xlab="x",ylab="y",main="Time
   dynamics")
#  text(N2,labels=1:nrow(N2),col=palette[cols])
#  if (ncol(N2)==1) {Npl<-cbind(T,N2)} else {Npl<-N2}
#  plot(gb,vertex.size=.2,vertex.label=clp,vertex.label.
   color=palette[cols])
#  if (ord==1){title("Community clusters in time (Time-
   Space)")} else{
#    title("Community clusters in time (Space-Time;
   reversed)")}
```

```
#  for (a in 1){
#    a1<-(row(C)-col(C)==a)
#    a2<-(row(C)-col(C)==(-a))
#    dev.new();pm(1,1)
#    plot(C[a1],type="l",ylim=range(C),main=paste(c("
   distance",a,collapse=", ")),ylab="cohesion")
#    text(C[a1],labels=1:sum(a1))
#    lines(C[a2],col="red")
#    text(C[a2],labels=1:sum(a2),col="red")
#    abline(h=mean(diag(C))/2,col="forestgreen")
#  }

  #dev.new();plot(N,type="l",lty=2,xlab="x",ylab="y",main
     ="Time dynamics",
  #                 xlim=c(.95,1.1),ylim=c(1.1,1.25))
  #  text(N,labels=1:nrow(N[-61,]),col=(1:7)[cols])
  rownames(Cb)<-1:n2;  rownames(C)<-1:n2
  colnames(Cb)<-1:n2;  colnames(C)<-1:n2

    print(table(clp))
  list(g=gb,cl=cl,clp=clp,C=C,Cb=Cb)
}
```

Appendix C

Background for data sets used in applications

In this Appendix, we some background and references from data sets employed in applications throughout the thesis.

- *Cultural distance data* See Figures 5.8, and 5.9. This is distance information available at the website `http://culturaldistance.com/`. These are distances compiled via the theory of fixation indices (see [35]*** added reference in bibtex***) from responses on the World Values Survey from 2005 to 2014. A list of questions considered for different dimensions is available at `https://www.michael.muthukrishna.com/cultural-distance-methods/`

- *European education data*. See Figures 4.4, and 4.5. This is data, from the National Center for Education Statistics, of highest level of educational attainment for 25- to 64-year-olds. Data was for the year 2015 (`https://nces.ed.gov/programs/digest/`).

- *United States state data* See Figures 5.10, 5.11, and 5.12. State facts based on data from a variety of U.S. government agencies in the 1970s and 2010s. This data is available in the `state.x77` data frame in `R` and the `facts` data frame from the `usa` package.

- *China Data*: Refer to Figures 5.6 and 5.7. The first dataset explores the annual

population growth rates in various Chinese cities in 2017, accessible at `https://www.statista.com/statistics/992720/china-annual-population-growth-in-first-and-second-tier-cities-by-city/`. The second dataset investigates the housing prices across different cities in China, using the 2024 data available at `https://mobile.anjuke.com/fangjia/quanguo2024/`.

- *Brain network data* This is from f-MRI data from [51]. The times series correlations have been thresholded to only include an upper percentage of correlated voxel pairs (1444 edges). The parcellation is employing that from [43] with 268 voxels.

Dizhou Wu

Curriculum Vitae

1022 Carolina AVE
Winston-Salem, NC 27101
(336) 666-7669
wud18@wfu.edu

Education

2023– **M.S. Statistics**, *Wake Forest University*, Winston-Salem, NC

2018– **Ph.D. Physics**, *Wake Forest University*, Winston-Salem, NC

2015–2018 **B.S. Mathematics**, *Shanghai Jiao Tong University*, Shanghai, China

2014–2018 **B.S. Physics**, *Shanghai Jiao Tong University*, Shanghai, China

Relevant coursework

Physics: Quantum Mechanics, Statistical Mechanics, Classical Mechanics, Electromagnetism, and General Relativity (audit). **STA:** Advanced Statistical Inference, Generalized Linear Models, Stochastic Processes and Applications, Applied Survival Analysis, Time Series and Forecasting, Networks: Models and Analysis, Intro to Statistical Learning, Multivariate Statistics, and Linear Models. **CS:** Bioinformatics, and Multidimensional Data Analysis.

Professional experience

Research

2023– **MS Thesis: Detecting Local Change in Domain-Features Data**, *Advisor: Kenneth Berenhaut*, Email: berenhks@wfu.edu

Focused on clustering data points based on domain features, specifically within time series contexts. Developed an innovative approach by integrating temporal sequencing into the non-parametric clustering method, PaLD, to enhance the analysis of molecular dynamics (MD) data. This approach emphasizes the significance of temporal ordering in the clustering process, allowing for a more accurate characterization of dynamic changes in MD simulations.

2018– **Ph.D. Dissertation: Molecular Dynamics Insights into Thrombin: Allosteric Modulations, Mutational Impacts, and Therapeutic Implications**, *Advisor: Freddie Salsbury Jr*, Email: salsbufr@wfu.edu

Conducted comprehensive research on thrombin using MD simulations and machine learning. Investigated light chain mutations, dual sodium ion binding sites, and the role of hydrogen bonds in biomacromolecules. Explored allosteric modulation by thrombomodulin and the impact of mutations like W215A/E217A on thrombin's dynamics, contributing valuable insights into its allostery and potential therapeutic applications in biochemistry and pharmacology.

Teaching

2018–2023 Teaching assistant for PHY 307 Biophysics, PHY 337 Analytical Mechanics, PHY 113/114 General Physics I/II, PHY 110 Introductory Physics, Department of Physics, Wake Forest University

2018–2023 Lab instructor for PHY 114L General Physics II Lab, PHY 120A Physics & Chemistry of the Environment, Department of Physics, Wake Forest University

Other experience

2023– **Route Setter**, *Climbing Wall, Outdoor Pursuits*, Wake Forest University
Responsible for designing and setting climbing routes for various skill levels.

Awards

2024 Travel Award for BPS 2024 Annual Meeting (International conference), awarded by Wake Forest Univ.

2023 Center of Molecular Signaling Fellowship, awarded at Wake Forest Univ.

2023 Travel Award for APS 2023 March Meeting (International conference), awarded by Wake Forest Univ.

Publications

2024 **Wu, D.**, & Salsbury Jr, F. R. (2024). Allosteric Modulation of Thrombin by Thrombomodulin: Insights from Machine Learning and Statistical Analysis of Molecular Dynamics Simulations. J Chem Inf Model. (In review) [Available upon request]

2023 **Wu, D.**, & Salsbury Jr, F. R. (2023). Unraveling the Impact of W215A/E217A Mutations on Thrombin's Dynamics and Thrombomodulin Binding through Molecular Dynamics Simulations. (Preprint/Submitted)

2023 **Wu, D.**, Prem, A., Xiao, J., & Salsbury, F. R. (2023). Thrombin-A Molecular Dynamics Perspective. Mini Rev Med Chem. 2023 Aug 21. Online ahead of print.

2023 **Wu, D.**, & Salsbury Jr, F. R. (2023). Unraveling the role of hydrogen bonds in protein via two machine learning methods. J Chem Inf Model. 2023 Jun 26;63(12):3705-3718. Epub 2023 Jun 7.

2022 **Wu, D.**, & Salsbury Jr, F. R. (2022). Simulations suggest double sodium binding occurs in thrombin with possible functional significance. J Mol Model. 2022 Apr 13;28(5), Article number: 120.

2021 **Wu, D.**, Xiao, J., & Salsbury Jr, F. R. (2021). Light Chain Mutation Effects on the Dynamics of Thrombin. J Chem Inf Model. 2021 Feb 22;61(2):950-965.

Presentations

2024 **Wu, D.** & Salsbury, F. (2024). Allosteric Modulation of Thrombin by Thrombomodulin: Insights from Machine Learning and Statistical Analysis of Molecular Dynamics Simulations. Annual Biophysical Society Meeting 2024.

2024 Salsbury, F., **Wu, D.**, & Xiao, J. (2024). Dynamic allostery and thrombin: Insights from modeling mutants, ions and binding using molecular dynamics, statistical analysis and machine learning. Annual Biophysical Society Meeting 2024.

2023 **Wu, D.** & Salsbury, F. (2023). Unraveling the role of Hydrogen bonds via two machine learning methods. 2023 APS March Meeting, Bulletin of the American Physical Society.

2023 Salsbury, F., & **Wu, D.** (2023). Examining dynamic allostery and communication in proteins via machine learning and statistical analysis. 2023 APS March Meeting, Bulletin of the American Physical Society.

2022 Salsbury, F., & **Wu, D.** (2022). Simulations suggest double sodium binding induces unexpected conformational changes in thrombin. 2022 APS March Meeting, Bulletin of the American Physical Society.

2021 **Wu, D.**, & Salsbury, F. (2021, October). Simulations of Thrombin and the Role of Sodium Binding. In PROTEIN SCIENCE (Vol. 30, pp. 23-23).