Abstract

Feature Selection for Diffusion Methods Within a Supervised Context.

Minh-Tam Le

2014

We apply diffusion geometry to sociopolitical and public health datasets. Our specific goal is to reveal hidden trends and narratives behind UN voting records and alcohol questionnaire response patterns. Importantly, seeking those hidden variables in a supervised context, e.g. alcohol-*ab*use, can be problematic for diffusion geometry. We suggest two approaches to deal with these shortcomings. First, we develop a correlation-based hierarchical clustering algorithm that exposes sub-patterns in the feature (response) space; this works in the UN voting context. Second, we introduce a feature selection algorithm based on a second-order correlation measure to guide diffusion embeddings; this significantly improves the performance of diffusion methods in the alcohol context. Together they suggest how to structure embeddings when there exist strong correlations among features irrelevant to a given labeling function.

Feature Selection for Diffusion Methods Within a Supervised Context

A Dissertation

Presented to the Faculty of the Graduate School

of

Yale University

in Candidacy for the Degree of

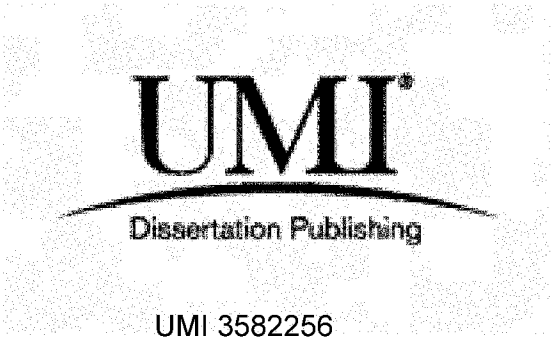
Doctor of Philosophy

by

Minh-Tam Le

Dissertation Advisor: Steven W. Zucker

December, 2014



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Acknowledgements

A few lines of thanks are not sufficient to contain the debt of gratitude I owe Professor Zucker. Perhaps the whole page is still too narrow to describe what he has done for me. He has always been a caring and patient teacher guiding me from unfamiliar ground, step by step, to the completion of my study. Time after time, he even went out of his way in order to keep me in good shape. Without his comments and suggestions, this dissertation would not have been possible.

I would like to thank my parents, sister and especially my wife, for their support during my study. This one special line of dedication could not match the endless patience and love given by my wife, Phuong, who has put up with me for the last six years, with an ocean physically separating us.

I am grateful to the wonderful professors and staff of the Computer Science Department at Yale for giving me a home away from home for all these years. My special thank to friends and colleagues in Computer Science and Applied Math programs Matt, Dan, Ben and Jerrod for their company and generous assistance.

Finally, I would like to thank Professors Ronald Coifman, Vladimir Rokhlin and Robert Zucker for agreeing to serve as readers of the dissertation.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Recent advances in technology have enabled acquisition and storage of information in massive amounts. Analyzing these datasets have become one of the top challenges across many fields, ranging from business [1, 2], to personality classification [3, 4], to healthcare [5–8] to governments [9–12]. Many datasets are bipartite, or in social science terminology, two-moded [13], whose data points are usually represented as a discrete function $A_{x,y}$, which provides measure of the relationship between observation $x \in X = \{x_1, x_2, \cdots, x_n\}$ to feature $y \in Y = \{y_1, y_2, \cdots, y_m\}$. The distinction between observations and features are interchangeable. Examples include surveys, questionnaires, vote records, ratings etc. An abridged version of one is presented in Table 1.1 In this work, we focus on datasets which come from social science domains.

These datasets are often (i) multidimensional (m and n is large) and (ii) sparse (missing values are the norm). The common interested tasks belong to one of the following categories:

- Unsupervised exploratory: describe the observations by a small amount of dimensions $\Psi^{(\kappa)} : x \in X \to (\psi_1(x), \psi_2(x), \ldots, \psi_\kappa(x))$, $\kappa \ll n = |X|$ (dimensionality reduction [14, 15]) or clusters $\Gamma_i \subset X, \Gamma_i \cap \Gamma_j = \varnothing, \forall i \neq j$, thus exposing existing *dominant pattern* of organization hidden in the datasets [16].

Table 1.1: *An excerpt from the UN voting data [9] of 5 countries (USA, UKG, RUS, POL, CHN) in 1990 on 5 issues, denoted by their roll call id's (RCID): #3508 (Dissemination of information on decolonization) #3510 (Observer status of national liberation movements recognized by the OAU and/or by the League of Arab States) #3515 (Cessation of all nuclear test explosions) #3538 (Calls upon Israel to become party to the Treaty on the Non-Proliferation of Nuclear Weapons) #3570 (Status of the International Convention on the Suppression and Punishment of the crime of Apartheid).*

	#3508	#3510	#3515	#3538	#3570	···
USA	No	No	No	No	No	
UKG	No	No	No	Abstain	Abstain	
RUS	Yes	Yes	Yes	Yes	Yes	
POL	Abstain	Abstain	Abstain	Abstain	Abstain	
CHN	Yes	Yes	Abstain	Yes	Yes	
⋮						

- Supervised classification, clustering or ranking: given a label function $f : x \in X \to \mathbb{R}$, organize the observations in $\Psi^{(\kappa)}(x)$ or clusters $\{\Gamma_i\}$ such that observations having similar value in f are mapped closer together in $\Psi^{(\kappa)}$ or grouped together in the same cluster [17].

- Collaborative filtering: Using the existing values in $A_{x,y}$ to infer the missing data points, e.g. infer person x_i's answer to question y_j in the survey even though he skipped it; or infer a representative x_i's vote on an issue y_j in case he abstained or was absent from voting [1, 2, 18].

These categories are related to each other, and considerations are due when the goal transits from one category to another.

Surveys and voting records have been studied extensively in multiple social fields such as political science [19, 20], psychology and sociology [21, 22]. Existing studies usually assume (i) a fixed form of relationships between observations and features which is inflexible and hard to extend to new data points [23–26] and/or (ii) an egocentric, dyadic approach toward observations, i.e. observations are only affected

by other observations directly linked to them [13, 21, 22, 27]. We argue that there is a transitionary nature in social data: goods *diffuse* through the network of trading countries; political goodwills or influences *diffuse* through the network of countries, signaled by votes. In IR (International Relations), it is not uncommon that the pressure to resolve political issues come from third or even fourth-party outsiders. In surveys and questionnaires, similarity measures between pairs of participants are often approximated by considering other intermediate participants, because different people may answer different sets of questions. Similarity among survey participants can thus be considered as *diffusing* through the network of participants. Therefore, the "distance" or "closeness" measure between any pair of observations is not simply the given direct relationship between them, but has to take into account all indirect paths of diffusion through intermediate observations. We apply this notion of diffusion distance [28] to these social datasets and show that it reveals hidden global trends behind the numbers.

Investigating these social datasets in supervised contexts, we found that in many cases, a significant amount of the features provided have nothing to do with the label function f. This may come from the design of the survey itself, i.e. the features, questions are usually designed to cover a wide range of areas in order to explore novel patterns, even though the majority of these well-intentioned coverage may turn out to be irrelevant. Similarly, voting and rating records may cover a wide range of topics, subjects which are unrelated to the interested classification. In addition, these "extra" features are likely to be correlated with each other, forming consistent patterns which may interfere with the discovery of any other structure in the data that is relevant to f. We show this phenomenon in synthesized data experiment, and introduce a filtering algorithm to select features relevant to f. We report several examples on real datasets where this algorithm significantly improves mapping, clustering results. In the collaborative filtering context, for each feature y_i, we can use the same filtering

algorithm to find a set of other relevant features which may help inferring the missing values of y_i.

1.1 Contributions and Organization of the Thesis

Chapter 2 reviews the social network model and classical dimensionality reduction techniques commonly used in social sciences and public health literature. Chapter 3 discusses diffusion distance and diffusion maps in sociopolitical applications.

The main contribution of this thesis is in two parts: First, in Chapter 4, we applied diffusion geometry to sociopolitical datasets, mainly International Relations (IR) data, uncovering hidden spatial-temporal patterns and underlying political alignments in the network of nations. Examination of the non-linear embeddings of these data across time reveals interesting historical narratives, suggesting the results may serve as a proxy for the analysis of security-related datasets. To our knowledge, this has never been done before in these fields. Second, in Chapter 5, we investigated the performance of diffusion maps on correlated datasets such as voting records, surveys, questionnaires, *under supervised context.* We empirically showed, with synthetic and real datasets, that, in supervised learning tasks, exploratory diffusion maps may be ineffective against these datasets in which the number of correlated, irrelevant features could be overwhelming. In 5.4, we proposed a feature selection scheme using second-order correlation to pick out only features relevant to the labeling function, thereby enhancing diffusion methods' performance.

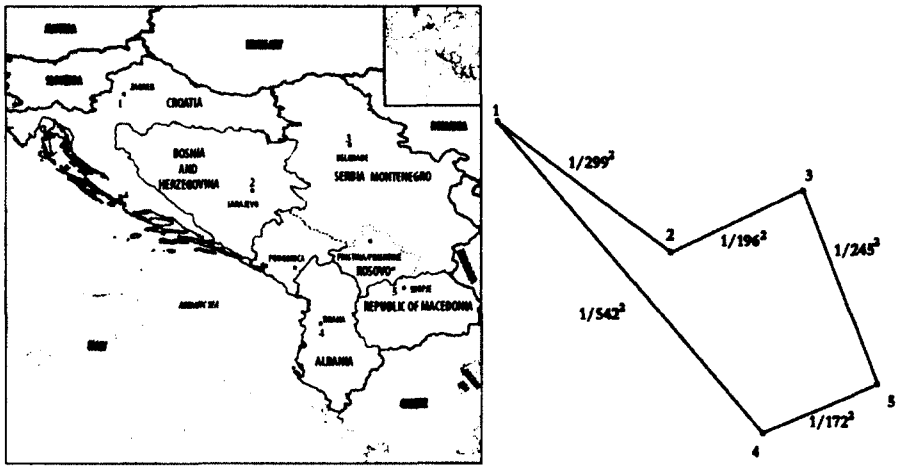
Chapter 2

Social network analysis and Dimensionality Reduction

2.1 Social networks, graphs and kernels

We begin by reviewing the social network model which, given a pairwise (symmetric) relation $W_{ij} = W(x_i, x_j)$ between a set of entities or social actors $X = \{x_1, \ldots, x_n\}$, represents the data as a graph $G(V, W)$. Each observation x_i is modeled as a node or vertex $v_i \in V$ in the graph. Fig. 2.1 considers the small network of five Balkan countries in the year 2000. Each country is represented as a [numbered] vertex in the network. The pairwise relation W_{ij} is defined as a set of *weighed edges* between pairs (x_i, x_j), the weight values W_{ij} measure the social interactions between x_i and x_j. The social network model provides an abstraction which enables several levels of analysis [29], mainly focusing on the relationships among social actors, such as people [30–33], countries and organizations [34, 35].

Different social relations result in different W, and hence different graphs. Depending on the type of social interactions of interest, W could denote (i) border contiguity [36], or (ii) geographical proximity [37], or (iii) diplomatic exchange [12]



(a) Five Balkan countries

(b) Graph of the corresponding 5 capital cities

Figure 2.1: *The geographical network of 5 Balkan countries in the year 2000 (a) and its corresponding graph (b). The graph is complete, with an edge between every pair of capital cities. To avoid overcrowding the figure, only 5 edges are shown for demonstration. In this network, the edge weight between each pair of cities is defined as the reciprocal of the squared distance (in km^{-2}) between them $W_{ij} = \frac{1}{r_{ij}^2}$, where r_{ij} is the geographical distance between capital cities of countries x_i and x_j.*

between countries, etc. Tables 2.1 & 2.2 illustrate three different graphs based on different social relations between the same set of the five Balkan countries X depicted in Fig. 2.1a. Table 2.1 shows a graph of border contiguitiy, which cotains an edge between countries x_i and x_j, $W_{ij} = 1$ if the two share a land border, otherwise $W_{ij} = 0$. The graph in Table 2.2 contains an edge between every pair of countries (x_i, x_j). The edge weight measures geographical proximity between countries by the inverse squared distance between capital cities.

The conventional interpretation of W as a pair-wise relation of edges is purely egocentric, providing us with the knowledge of individual edges but not what the whole network looks like, and thus is analogous to describing a forest by enumerating branches and leaves. Another interpretation of W which offers a fuller picture of the network is that of W as a *kernel*. A kernel is a *symmetric* function which measures how a pair of vertices (i, j) in the network are *similar*, *close* to each other, with respect

Table 2.1: *Matrix representation of the graph between 5 Balkan countries (Fig. 2.1a) whose edges' weights W between pairs of countries (i, j) denote whether x_i and x_j share a common land border (1) or not (0) [36].*

	1-CRO	2-BOS	3-YUG	4-ALB	5-MAC
1-CRO	-	1	1	0	0
2-BOS	1	-	1	0	0
3-YUG	1	1	-	1	1
4-ALB	0	0	1	-	1
5-MAC	0	0	1	1	-

to the target social activity. For each vertex x_i in the network, the kernel W provides us with a measure of the local, egocentric neighborhood surrounding x_i, from which the network arrangement can be deduced. For example, by examining Table 2.1 as a whole, we recognize that (CRO, BOS, YUG) form a close group of border-sharing countries, while (YUG, ALB, MAC) form another group. With the exception of YUG which appears in both group, there's no connection between (CRO, BOS) and (ALB, MAC). Thus, CRO and BOS are "closer" to each other than to ALB, MAC, while YUG serves as the land bridge between the two groups. Even though Table 2.2 offers a more nuanced picture by replacing the yes/no border-sharing property with real values of capital proximity, the overall picture of the two groups still stand out. This grouping knowledge tells us that if the five countries are on an island, YUG will cut right in the middle of the island, sharing borders with other countries, separating CRO, BOS on one side and ALB, MAC on the other side. And that corresponds exactly to the real map (Fig. 2.1a).

Table 2.2: *Matrix representation of the graph between 5 Balkan countries (Fig. 2.1a) whose edges between pairs of countries (x_i, x_j) have weights equal to the inverse square capital distance (km^{-2}) between x_i and x_j: $W_{ij} = \frac{1}{r_{ij}^2}$, where r_{ij} denotes the geographical distance between capital cities of countries i and j [37].*

	1-CRO	2-BOS	3-YUG	4-ALB	5-MAC
1-CRO	-	$1/299^2$	$1/373^2$	$1/542^2$	$1/560^2$
2-BOS	$1/299^2$	-	$1/196^2$	$1/248^2$	$1/274^2$
3-YUG	$1/373^2$	$1/196^2$	-	$1/338^2$	$1/245^2$
4-ALB	$1/542^2$	$1/248^2$	$1/338^2$	-	$1/172^2$
5-MAC	$1/560^2$	$1/274^2$	$1/245^2$	$1/172^2$	-

2.2 Two-mode network datasets

Having defined the social network abstraction as graphs on a single set of entities (which are termed 1-mode network in social fields), we describe a special case in which there are two exclusive subsets of entities and interactions are only permitted between entities of different sets. The graph model of this type of networks resembles the one shown in Fig. 2.2, which is a bi-partite graph. In social science terminology, this is a two-mode network [13], which is very common in practice, such as surveys, questionnaires (interactions between people and questions), or votes, ratings (interactions between voters, raters and subjects, issues) etc. These datasets are usually given as a discrete function $A_{x,y}$, between an observation $x \in X = \{x_1, x_2, \cdots, x_n\}$ and a feature $y \in Y = \{y_1, y_2, \cdots, y_m\}$. With finite n, m, the data is an $n \times m$ matrix A, similar to the one given by Tables 1.1 & 2.3. For distinguishability purpose, we shall call the entities in X observations and the ones in Y features. Even though most of the times, we are interested in the structure of $\{x_i\}$, the roles of X and Y are interchangeable, just as matrix A can be transposed. At times, with reference to a particular dataset, we may switch terminology and refer to X as countries, people, or participants, and to Y as resolutions, questions, issues.

Fig. 2.3 gives us a graphical example of the structure of countries X, given their votes on resolutions Y in the UN General Assembly [9]. On every UN resolution,

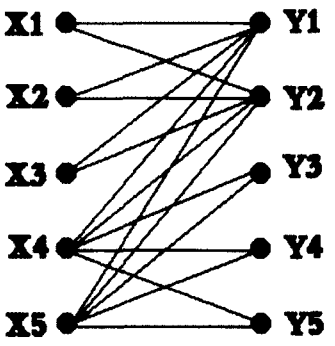


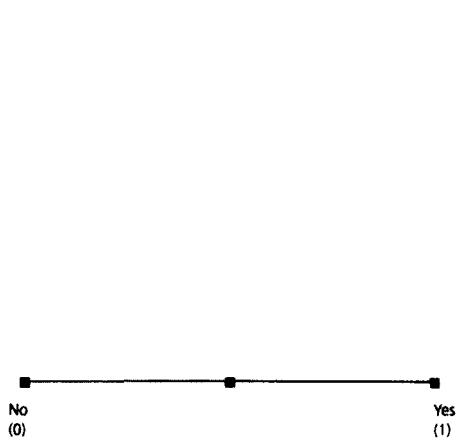
Figure 2.2: *Bipartite graph of observation set* $X = \{X_1, X_2, X_3, X_4, X_5\}$ *and feature set* $Y = \{Y_1, Y_2, Y_3, Y_4, Y_5\}$.

Table 2.3: *Countries' membership status in Inter-Governmental Organizations (IGO) in year 2000 [11]. Cell* (i, j) *takes value 1 if the country* i *was a member of organizatin* j *in 2000. Otherwise, cell value is 0.*

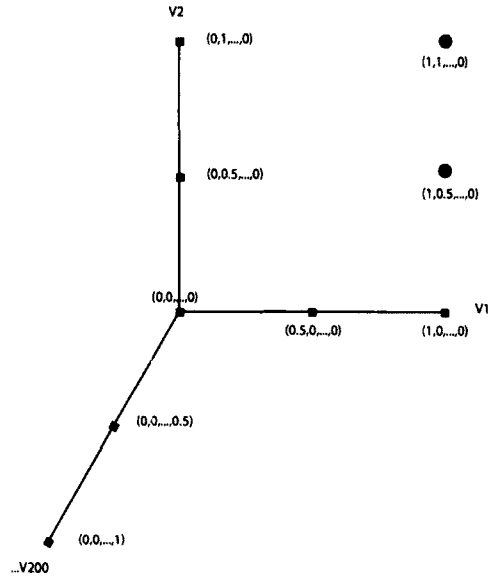
	EU	NATO	NAFTA	UNESCO	WTO	IMF	OECD	⋯
USA	0	1	1	0	1	1	1	
UKG	1	1	0	1	1	1	1	
FRN	1	1	0	1	1	1	1	
SIN	0	0	0	0	1	1	0	
CHN	0	0	0	1	0	1	0	
⋮								

a member country cast a vote of either Yes or No. We can visualize the position of a country on a particular resolution as point on a straight line, as illustrated in Fig. 2.3a. Here Abstain is positioned right in the middle between Yes and No. To this continuum, we assign an "opinion-score" ranging from 0 (No: totally against) to 1 (Yes: totally for). For example, if the resolution at hand here is the R/5/532B resolution in 1950 which condemns South Africa's Apartheid policy as racial discrimination, then we will see South Africa, Australia, Belgium as points toward the left of the spectrum (they voted No), their scores are 0 on the resolution; the Latin American and the Arab nations make up points toward the opposite end (they voted Yes) and thus scoring 1; while the U.S., Russia and most of European countries vacillating in between (they chose to Abstain from voting).

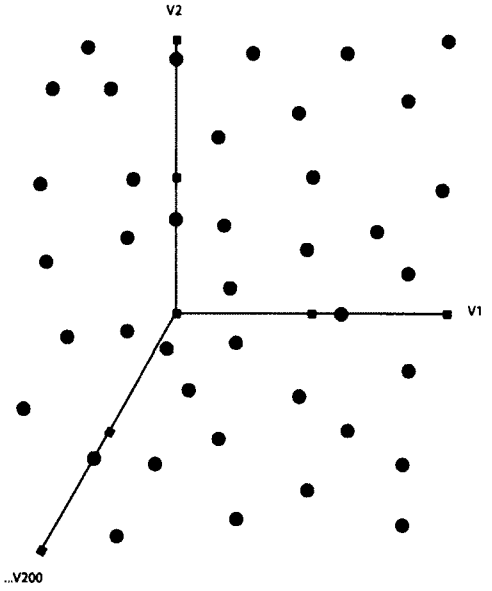
The number of dimensions needed to describe the original data can be very high: we have to look at every resolution in the dataset, which makes it difficult for human inspection. Every year, UN member nations vote on about 100-200 resolutions. If each resolution is visualized as a line like Fig. 2.3a, the whole set of votes on all 200 resolutions constitute a 200-dimensional space, where each dimension is a line similar to Fig. 2.3a. In that multi-dimensional space, each country is a point, whose 200 coordinates describe its positions on all the 200 resolutions. As our human eyes can only perceive up to 3 dimensions, Fig. 2.3b only shows 3 dimensions V_1, V_2, V_200 as representatives, together with a few sample points and their coordinates. An example of a 3-dimensional space is the RGB color space, whose 3 dimensions represent a color value of Red, Green, or Blue, ranging from 0 to 1. At (0,0,0), the resulting color is black; (1,0,0), (0,1,0), (0,0,1), (1,1,1) represent red, green, blue, and white respectively. As we move a color point from (0,0,0) along the R dimension, its color will gradually get redder. Similarly in the voting space, a country's point moving in parallel to a particular dimension V_i from 0 toward the direction of 1 in that dimension signifies that country leaning more and more toward supporting the



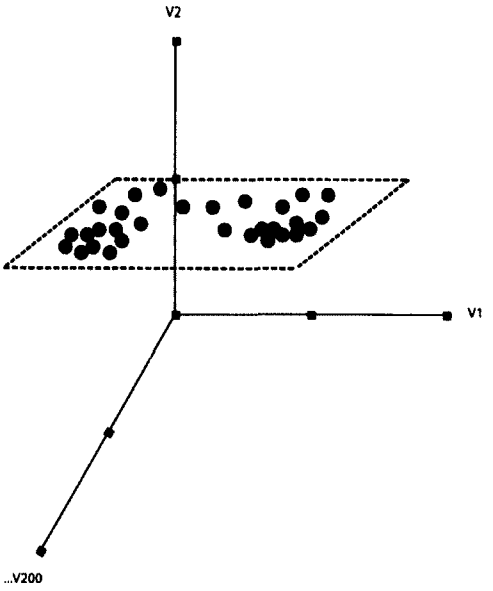
(a) 1-dimensional visualization of UN member votes on a particular UN resolution. The position of a member ranges from No (-10), Abstain (0), to Yes (10).



(b) Multi-dimensional voting space with sample points and coordinates.



(c) Points uniformly distributed in the voting space, if votes are cast randomly.

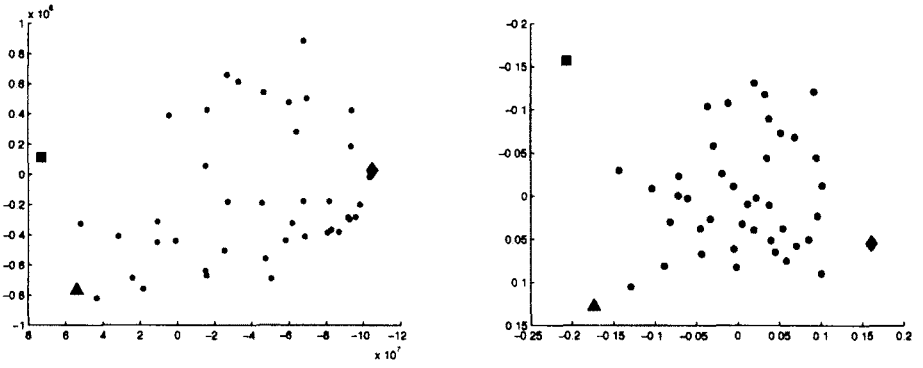


(d) Voting points reflecting low-dimensional political alignment.

Figure 2.3:

resolution represented by V_i.

Real-life data points do not occur randomly but tend to correlate in a structured way according to some underlying real-life process. Revealing the structure that the data points "live in" may allow us to understand the underlying process. If UN members cast their vote in a completely arbitrary manner, we should expect the points to scatter all over the space, with no discernible alignment, as shown in Fig. 2.3c. However, instead of mindlessly tossing coins before casting votes, UN members are expected to be rational players, who cast their votes consistently on related resolutions (e.g. Latin American nations consistently supported resolutions against Apartheid policy.) Additionally, we can resasonably expect the positions of UN members to reflect their political alignment (e.g. UKR and BLR's positions followed the lead of the Soviet Union during the Cold War.) That implies: if we know a nation's (i) position on one resolution, or (ii) the positions of its political neighbors (read "allies"), it is possible to predict its standing in other related resolutions. That further implies that, instead of the original 200-dimensional description, a country's position can be described by a much *reduced* number of dominant issues, topics, or *dimensions* [38], as shown in Fig. 2.3d. *Dimensionality reduction* is the process to recover the low-dimensional structure on which the data lives from the original high-dimensional data space. While applications of dimensionality reduction have been attempted before, we show that (i) the concept of diffusion distance reveals embeddings that are much more informative, and (ii) these embeddings can be conducted across time, revealing the evolution of sociopolitical relationships. As such, they serve as a proxy for the analysis of political datasets.



(a) (b)

Figure 2.4: *2D embeddings of European capital cities, using (a) PCA, (b) MDS given their pairwise geographical distances r_{ij}. Each point in the maps represents a capital city in Europe. Special markers are drawn to denote the capital cities of Iceland (■), Portugal (▲), and Turkey (♦)*

2.3 Dimensionality Reduction In Social Science

Classical dimensionality reduction techniques commonly employed by social scientists include Principal Component Analysis (PCA) [24, 39, 40], Factor Analysis [41–44], and Multidimensional Scaling (MDS) [27, 45–47]. These techniques are usually performed via statistical software packages like SPSS, or SUDAAN [48], Mplus [49].

2.3.1 Principal Component Analysis (PCA)

PCA seeks a set of principal axes $\Psi = \{\psi_k\}$ which represent most of the variance in the data set [50]. These axes are usually given by the eigenvalue decomposition of the covariance matrix:

$$C = A(I_{m\times m} - \frac{1}{m}\mathbb{1}_{m\times m})(I_{m\times m} - \frac{1}{m}\mathbb{1}_{m\times m})^T A^T \tag{2.3.1}$$

where A is the original data matrix, which can be a vote matrix (n countries $\times m$ issues), or questionnaire matrix (n people $\times m$ questions). $\mathbb{1}_{m\times m}$ denotes the $n \times n$ matrix of all 1 entries, and $I_{m\times m}$ the identity matrix of size m. Dimensionality reduc-

tion is achieved by limiting the representation to the first few κ principal components, which account for most of the variability among the data points. Since PCA makes use of data variances, it is vulnerable to outliers. In Fig. 2.4a, due to the extreme position of Iceland to the northwest Portugal to the southwest and Turkey to the southeast, the first component lies in a southwest-northeast direction, resulting in the distortion of the projected map.

2.3.2 Factor Analysis

Beside the conventional projective formulation, there is another probabilistic formulation of PCA [51] which introduces normally distributed latent variable z such that observations $x \in X$ are Gaussian distributed conditioned on the value of z:

$$p(x|z) = \mathcal{N}(v|Fz + \mu, \sigma^2 I) \tag{2.3.2}$$

for some variance σ^2, implying the mean of $x \in X$ is a general linear function of the latent variable z, governed by a $\kappa \times m$ matrix F and an m-dimensional mean vector μ, i.e.

$$x = Fz + \mu + \epsilon \tag{2.3.3}$$

where ϵ denotes additive noise which is m-dimensional Gaussian-distributed noise with covariance $\sigma^2 I$. F can be computed by EM algorithms [52] and its columns can be shown to define the principal subspace spanned by Ψ derived from the standard projective PCA above [51].

Factor Analysis, also called Exploratory Factor Analysis, is a linear Gaussian latent variable model closely related to probabilistic PCA [53]. Instead of isotropic covariance $\sigma^2 I$ as in Eq. 2.3.2, the conditional distribution has a diagonal covariance matrix Σ:

$$p(x|z) = \mathcal{N}(v|Fz + \mu, \Sigma) \tag{2.3.4}$$

in which factor loadings F measure the correlations between observations and the diagonal elements of Σ, measuring independent noise variances, are called uniquenesses. The parameters F, μ, Σ can be determined using an EM algorithm [54]. Eventhough factor analysis has a long tradition in social science, it has been argued theoretically [55, 56] and empirically [57, 58] that Multidimensional Scaling (MDS) is more suitable to voting, rating data than Factor Analysis.

2.3.3 Multidimensional Scaling (MDS)

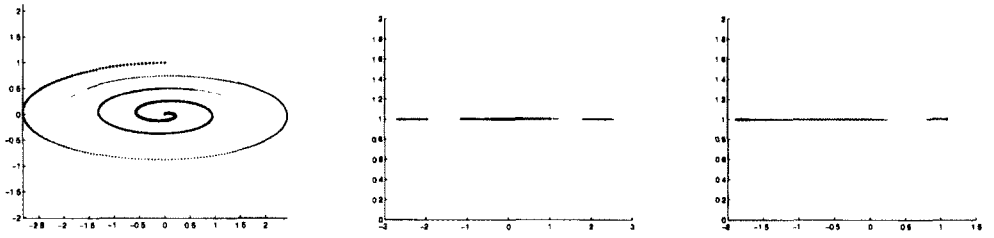
Multidimensional Scaling (MDS) [59, 60] is another useful technique. The Norminal Three-Step Estimation (NOMINATE) model [61] frequently used in voting research literature is a variation of MDS, which iteratively estimates the low-dimensional mappings of the observation and feature sets in alternating order. Given symmetric distance information R_{ij} (a dissimilarity function, the opposite of a kernel, which is a similarity function), MDS finds a low-dimensional embedding $\Psi : x_i \in X \rightarrow \Psi(x_i) = [\psi_k(x_i)]$ such that *all pairwise distances* R_{ij} are preserved. Mathematically, MDS finds Ψ so as to minimize the strain function:

$$\text{Strain} = \sqrt{\sum_{i,j}[R(i,j) - dist_{\Psi}(x_i, x_j)]^2} \tag{2.3.5}$$

or the stress function:

$$\text{Stress} = \sqrt{\frac{\sum_{i,j}[R(i,j) - dist_{\Psi}(x_i, x_j)]^2}{\sum_{i,j} dist^2_{\Psi}(x_i, x_j)}} \tag{2.3.6}$$

where $dist_{\Psi}(x_i, x_j)$ denotes the Euclidean distance between points x_i and x_j in the embedding Ψ.



(a) Data points aligned in a spiral

(b) 1-dimensional embedding by MDS

(c) 1-dimensional embedding by diffusion maps

Figure 2.5: *1-dimensional embeddings of the spiral set of data points in (a), using (b) MDS and (c) diffusion maps. The original data is essentially a line whirled into a spiraling curve. MDS fails to identify this underlying 1-dimensional structure of the data, whereas diffusion maps succeeds.*

Although the embedding by MDS (Fig. 2.4b) successfully recovers the geography of Europe, it could fail to recover structures that are non-linear [62] or low-dimensional structures in "folded" datasets such as the case in Fig. 2.5, since MDS tries to preserve all pairwise linear distances. Here, the original 1000 data points (Fig. 2.5a) is on a single line curved into a spiral. The true underlying structure is the line, i.e. 1-dimensional, but the spiralling curve complicates the distances for MDS. Due to the curve, points which are far apart on the line appear to be closer in the 2-dimensional spiral, in a manner similar to that of a *"folded dimension"* [47], which needs to be *"unfolded"* [63] in order to get at the true alignment of data points. The MDS result in Fig. 2.5b is simply the original spiral projected down onto the vertical axis, completely missing the line structure, because MDS seeks to preserve the proportion of *all pairs of nodes* in the dataset. Diffusion maps, on the other hand, by preserving only the short distances (strong links) while ignoring long distances (weak links), successfully recover the 1-dimensional aligment, as shown in Fig. 2.5c.

Chapter 3

Diffusion Distance & Diffusion Maps

3.1 Motivating Example: The collapse of the Soviet Union

We come back to the UN General Assembly voting database [9] to demonstrate the power of our diffusion approach. Although many scholars may be skeptical about whether votes in the UN General Assembly even matters, left alone reflecting international dynamics, our method is able to review the hidden spatial-temporal structure in the data, particularly, in this example, the political dynamics before and after the disintegration of the Soviet Union.

Suppose there exists an arrangement of countries according to their UN voting patterns. In this arrangement those countries that voted similarly will be close to each other, and those that voted differently willl be far apart. As interactions between these countries change over time, their positions relative to each other keep evolving. If the world is a universe, countries would form stars and constellations revolving around each other in an eternal dance of alliances and conflicts, spurred, not by

physical laws of nature, but by the ever-changing force of political, economic and power interests. The question is whether knowing which countries are "close" can provide useful quantitative information for understanding political allignment and historical events. An illustration of what can be achieved from our diffusion approach is shown in Fig. 3.1 and Fig. 3.4.

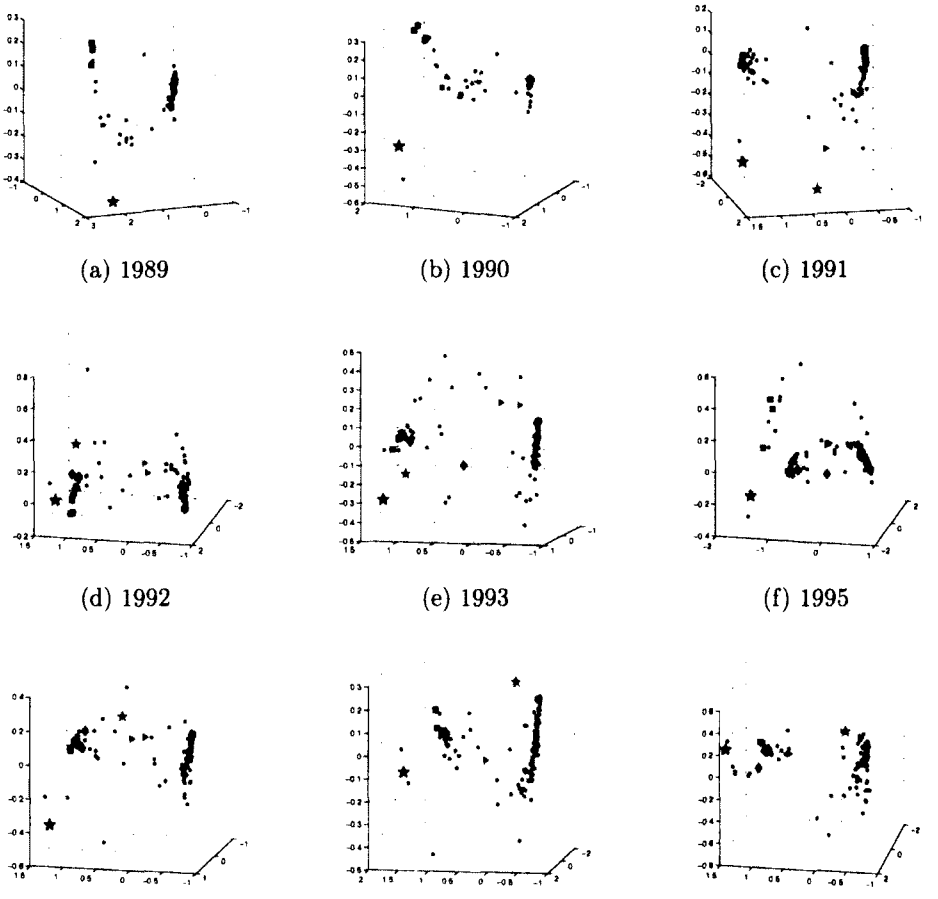
Fig. 3.1 shows the maps of nations according to their UN voting patterns at various time during 1989-2005 [1]. The embedded positions are computed by our diffusion method such that countries are placed closer to each other if they voted similarly, and far apart if they did not. For comparison, we present the embeddings using PCA on the UN voting patterns of the same periods. The apparent failure of PCA in this case is that most of the times, the USA are always positioned in the middle of the map, even close to RUS for some years. Fig. 3.3 compares 3 distance metrics: (a) diffusion distance by our method (which shall be defined in more details later in this article), (b) PCA embedded distance (Euclidean distance between data points embedded by a Principal Component Analysis projection), and (c) Hamming distance (normalized number of resolutions that countries voted different from each other.) Each subfigure plots the ratios of embedding distances in the period 1965-2000:

- $\frac{\text{dist(USA,EU)}}{\text{diam(EU)}}$ as the blue line
- $\frac{\text{dist(RUS,EU)}}{\text{diam(EU)}}$ as the red line
- $\frac{\text{dist(POL,EU)}}{\text{diam(EU)}}$ as the green line

where EU is defined as the states of the European Community.

We can "read" historical events simply by looking at the movement of countries in the diffusion maps in Fig. 3.1. The 1989 diffusion map is polarized with the Western bloc (blue) on the left and the Eastern bloc (red) on the right of Fig. 3.1a. The distance ratio plots in Fig. 3.3a clearly shows the green line (POL-EU) trailing the

[1] See clip 4 on http://www.cs.yale.edu/homes/vision/zucker/embeddings.html.

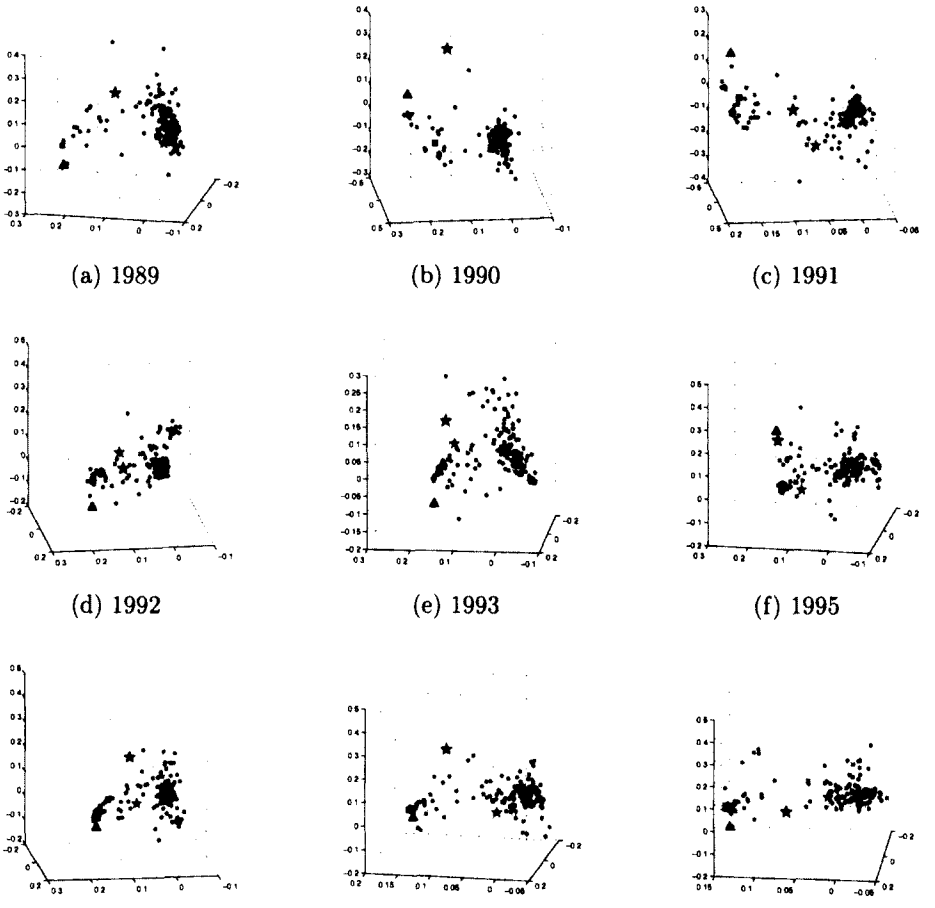


(a) 1989 (b) 1990 (c) 1991

(d) 1992 (e) 1993 (f) 1995

(g) 1997 (h) 2000 (i) 2005

Figure 3.1: *The collapse of the Soviet Union: Diffusion maps of UN voting pattern 1989-2005. Several countries are marked for case study identification:* ★ *(USA),* ■ *(UKG, FRN, BEL, LUX),* ★ *(RUS),* ♦ *(YUG),* ► *(UKR, BLR),* ■ *(POL, HUN),* • *(CHN).*

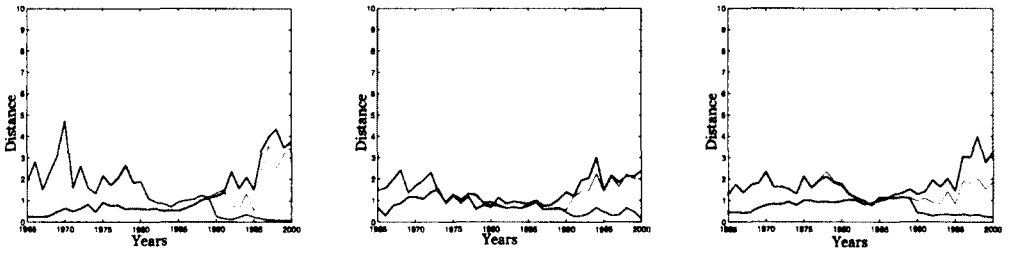


(a) 1989 (b) 1990 (c) 1991

(d) 1992 (e) 1993 (f) 1995

(g) 1997 (h) 2000 (i) 2005

Figure 3.2: *The collapse of the Soviet Union: PCA embeddings of UN voting pattern 1989-2005. Several countries are marked for case study identification:* ★ *(USA),* ■ *(UKG, FRN, BEL, LUX),* ★ *(RUS),* ♦ *(YUG),* ► *(UKR, BLR),* ■ *(POL, HUN),* • *(CHN).*



(a) Diffusion distance (b) Distance of PCA embedding (c) Hamming distance

Figure 3.3: *Ratios of embedding distances between USA-EU (blue), RUS-EU (red), POL-EU (green) in 1965-2000. Here EU is defined as the states of the European Community. Thes plots show how relations between USA, USSR, Poland and the Western European states changed over time, with Poland tailing the USSR until 1989, after which it was completely aligned with the West.*

red line (RUS-EU) prior to 1989, indicating Poland's policy completely dominated by that of the Soviet Union. However, in 1990 (Fig. 3.1a), Poland and Hungary (red squares) switched to the left, followed quickly by Czecholovakia, Bulgaria, and then the three newly independent Baltic republics. Fig. 3.3a clearly reveals a break between the green line and the red line from 1989, showing different trends in Poland and Russia's policies from then on. By 1991 (Fig. 3.1b), Russia (red star), Belarus, and Ukraine (2 red triangles) followed suit, as they moved toward the center. In 1992, after the Soviet bloc fully disintegrated (Fig. 3.1d), its members had all migrated to the left, with Ukraine and Belarus hanging in the middle, leaving China (red circle) on the right, close to the Arabs and the third world. Figs. 3.1d- 3.1f depicts Russia's effort to get close to the West, as Yeltsin vied for Western support for admission to NATO or the EU. The downward trend of the red line during 1992-1995 in Fig. 3.3a indicates Russia's aborted attempt to get close to the EU. After Yeltsin's second election in 1996 and his failure to court the West (Fig. 3.1g), Russia moved to the right of the map. Fig. 3.3a records a sharp ascent of the red line after 1996, implying Russia's abandonment of its westward movement. Further shift eastward occurred after Putin replaced Yeltin in 2000 (Fig. 3.1h), as Russia switched to the right, getting close to

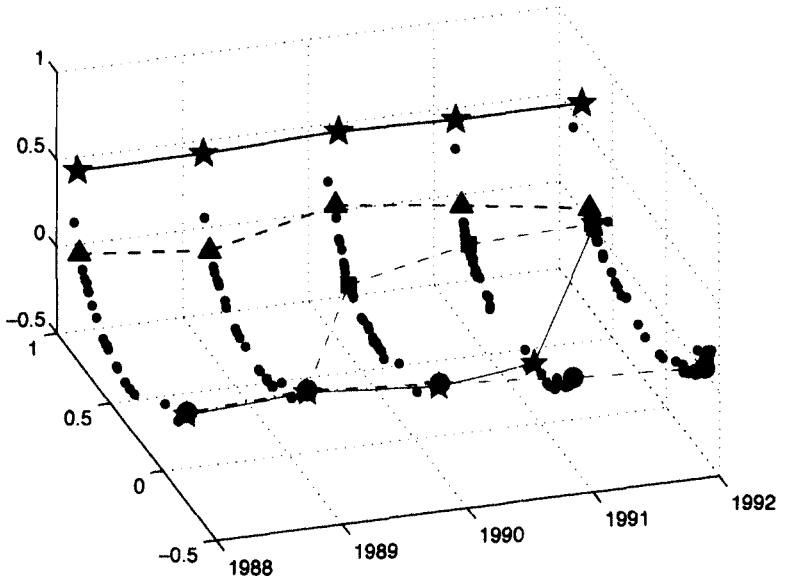


Figure 3.4: *The disintegration of the Soviet Union (1988-1992): The evolution of 2-dimensional diffusion maps of nations according to their voting patterns in the UN Assembly. Each dot denotes the global position of a country in a particular year. Special markers are drawn to denote:* ★ *(USA),* ▲ *(UKG),* ★ *(RUS),* ■ *(POL),* • *(CHN). Several lines are also plotted connecting the "paths" of these countries over time. Note how USA and UKG stayed relatively steady at their positions, while the paths of Communist states started to diverge since 1989. POL was the first to move out of the camp in 1990, followed by RUS, whereas CHN remained in their original position throughout the whole period.*

China again.

The collapse is even more evident in Fig. 3.4, which provides a time-evolution of the event by stringing the 2-dimensional structures of the alignments in Fig. 3.1 along the time dimension. It is apparent from the figure:

- USA and UKG stood close to each other in the 2-dimensional alignment, and their distance remain relatively stable throughout the 5-year period.

- The break-up of the Soviet Union is shown in the diverging lines of RUS, POL and CHN. The Union stayed intact until 1990, when POL moved away, toward the other side of the map. In 1991, RUS inched apart from CHN and the

third-world countries, and then moved completely out by 1992.

3.2 Diffusion Distance

Social phenomena and trade, unlike geography, follow a different distance measure. Goods and social capital *diffuse* from one place to another, perhaps through an intermediate country. Thus nearby countries matter more than distant ones. Since classical techniques preserve all pairwise Euclidean distances between the data points, we argue that not all distances should be preserved uniformly. Instead, *only short distances shoud be maintained, and even attenuated in order to preserve the local structure, while long distances should not be considered for keeping.* The argument is illustrated in Fig. 3.5. In political terms, we see a polarization in which two camps (B, C) closely communicate, but (A, B) barely interact with each other except through intermediary contacts located in the middle tunnel. An embedding which highlights this polarization should tighten the clusters' girth (thus *attenuating short distances*) and stretch the tunnel's length (thus *loosening long distances* and separating the two clusters from each other). Those are the characteristics of *diffusion.*

Think of a substance (e.g. money, population, or political influence) diffusing from a source point out to its neighboring points in amounts proportional to the neighbors' similarity to the source. The substance continues to diffuse to the neighbors of those neighbors, etc. Assuming a fixed amount of substance in the network, we can define $p_t(z|x)$ as the density of substance, originating from source point x, at point z at time t. Thus $p_t(z|x)$ would be high if there are many paths of length $\leq t$ connecting x to z, and low otherwise. If we take point $x = B$ on the right of Fig. 3.5 as the source, after t time steps, most of the substance originated from B should end up at points like $z = C$ on the right cluster, and only a small fraction ends up at points like $z = A$ on the left, because there are significantly more paths from B to C than to

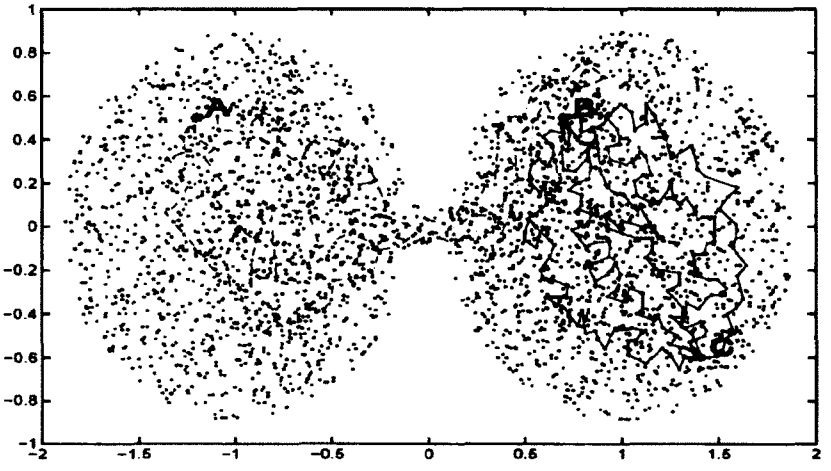


Figure 3.5: *Two tight clusters separated by a narrow path. It is obvious that there are many paths between any pair of nodes from the same cluster (B and C), while there are significantly fewer paths between any pair of nodes from different clusters (A and B).* Reproduced from [64].

A. The intuitive *diffusion distance* [28] between any two points x and y is a weighted difference between the two probability density functions:

$$\begin{aligned} D_t^2(x,y) &= \|p_t(z|x) - p_t(z|y)\|_\omega^2 \\ &= \sum_z (p_t(z|x) - p_t(z|y))^2 \omega(z) \end{aligned} \tag{3.2.1}$$

where $\omega(\bullet)$ is the weight function that normalizes the distance according to the density estimate of each vertex.

International trade can also be viewed as a diffusion process in which money diffuses from country to country. The polarization in Fig. 3.5 can be described in terms of trade during the Cold War. Assuming the trade pattern stays constant, the money will diffuse out to the two sources' trading partners. Thus $p_t(\bullet|USA)$ will be high in the West, and low in the East, while $p_t(\bullet|USSR)$ behaves in the opposite direction. The function $p_t(\bullet|USA)$ provides a notion of "trading sphere" of the USA. Therefore, the diffusion distance between the USA and the USSR can be defined as

the difference between their corresponding spheres $p_t(\bullet|USA)$ and $p_t(\bullet|USSR)$, as described by Eq. 3.2.1.

3.3 Diffusion maps

Diffusion maps is a non-linear dimensionality reduction technique [65–69] which has applications if many different areas [70–80]. Given a symmetric, positive-semidefinite kernel W, the diffusion maps $\Psi_t : x \in V \rightarrow (\psi_{t,0}(x), \psi_{t,1}(x), \ldots, \psi_{t,\kappa}(x))$ for some $\kappa \ll n = |V|$ [28] are defined as:

$$\Psi_t(x) = (\lambda_0^t \upsilon_0(x), \lambda_1^t \upsilon_1(x), \ldots, \lambda_\kappa^t \upsilon_\kappa(x))^T \tag{3.3.1}$$

where $1 = \lambda_0 \geq \lambda_1 \geq \ldots$ and $\{\upsilon_0, \upsilon_1, \ldots\}$ are the corresponding eigenvalues and eigenvectors of

$$\widetilde{W} = D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \tag{3.3.2}$$

and D is the diagonal matrix $D(x,x) = d(x) = \sum_y W(x,y)$. In a trade network, $d(x)$ represents the total value of goods traded by country x. In social network terminology, d is the degree centrality function, which can be used as a density estimate on the vertices [81]. The detailed construction of the diffusion maps is included in Appendix A.

It can be verified that the Euclidean distance in the embedded space spanned by Ψ_t corresponds to the diffusion distance D_t defined in Eq. 3.2.1, with $\omega(x) = \frac{1}{d(x)}$:

$$D_t(x,y) = \|\Psi_t(x) - \Psi_t(y)\| \tag{3.3.3}$$

Dimensionality reduction arises because, at large t, only those eigenfunctions corresponding to large λ survive.

Chapter 4

Diffusion maps of International Relation datasets

We applied diffusion maps to IR datasets. The embedding results reveal interesting low-dimensional structures in the network of countries. The illustrations in this section include snapshots of these 3-dimensional embeddings. Better views of these embeddings are provided on our website http://www.cs.yale.edu/homes/vision/zucker/embeddings.htn

4.1 Geographical distance

Fig. 4.1 [1] provides an experiment with geographical embedding of national capitals, with the kernel $W_{ij} = e^{-\frac{r_{ij}^2}{10^8}}$. The resulting embedding approximates global geographical positions.

[1] See clip 1 on http://www.cs.yale.edu/homes/vision/zucker/embeddings.html.

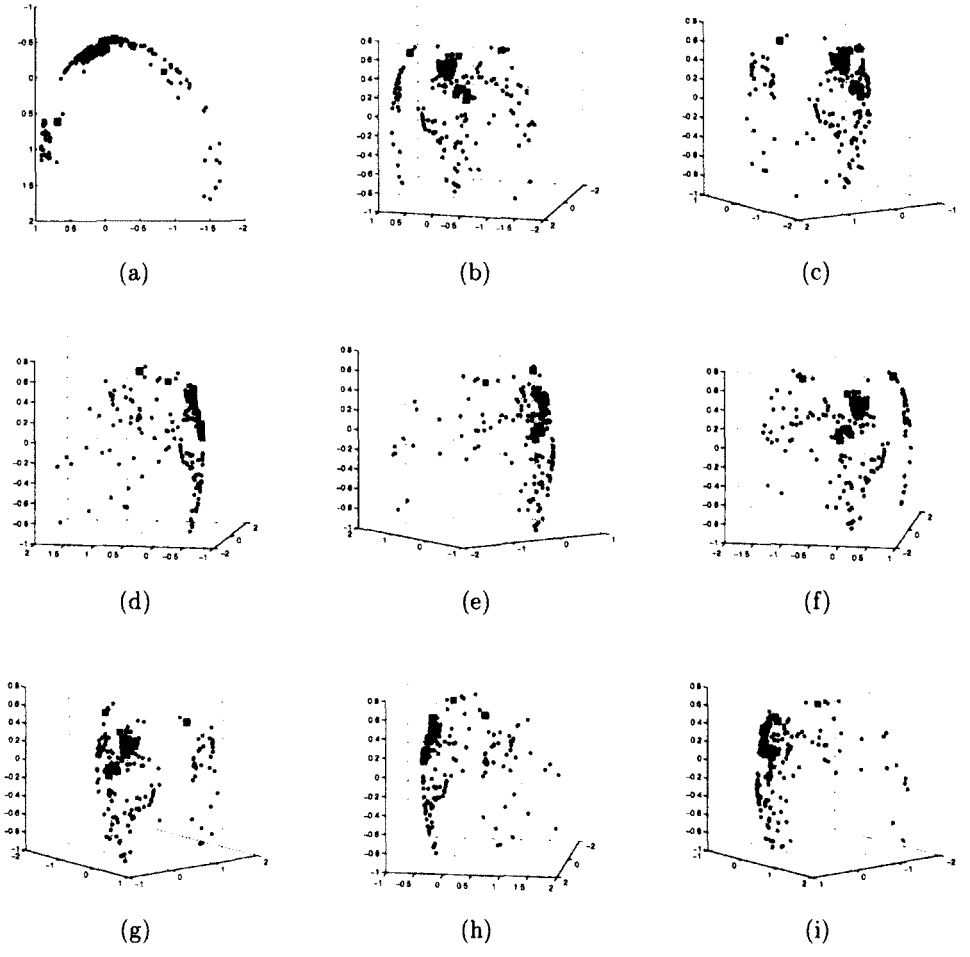


Figure 4.1: *Geographical embedding of national capitals in 3-dimensional space, using the 2^{nd}, 3^{rd}, and 4^{th} vectors of the diffusion map. The edge weight function is defined as $W_{ij} = e^{-\frac{r_{ij}^2}{10^6}}$ where r_{ij} is geographical distance between captitals of nations i and j. Figure (a) provides a top down view, while (b)-(i) show side views of the embedding from different angles, turning from west to east (counterclockwise). Several countries are marked with colored squares for easy identification:* ■ *(USA, UKG, FRN, BEL, ISR),* ■ *(RUS, CHN, POL, HUN, BLR),* ■ *(EGY, SYR, LEB, SAU, KUW).*

4.2 Inter-governmental organization memberships

Since inter-governmental organizations (IGO) play a crucial role in international relations, we ask how various countries are positioned, given their IGO memberships [11]. We consider joining an IGO a deliberate action: when a country i joins a particular IGO k, it has the *intention* of *moving close* to other countries j which are members of k. This is quantified by the ratio of common membership shared by i and j to the total number of IGO's joined by i:

$$\mathrm{INT}_{ij} = \frac{|\,\mathrm{MEM}_i \cap \mathrm{MEM}_j\,|}{|\,\mathrm{MEM}_i\,|} \tag{4.2.1}$$

where MEM_i denotes the set of IGO's joined by country i. The normalization over the total number of IGO's joined by i serves to equalize IGO-intensive countries (such as the US) with IGO-sparse countries (e.g Taiwan). However, due to its asymmetry, INT cannot be used as a kernel. The country-by-country correlation matrix C_{ij} offers a symmetric similarity function, in the same way the candidate-by-candidate correlation matrix was applied to MDS [46] in mapping political candidates from voters' survey scores. Since correlation values range from -1 to 1, shifting and scaling it yields the kernel, whose embedding result is shown in Fig. 4.2 [2]:

$$W_{ij} = \frac{C_{ij} + 1}{2} \tag{4.2.2}$$

[2]See clip 2 on http://www.cs.yale.edu/homes/vision/zucker/embeddings.html.

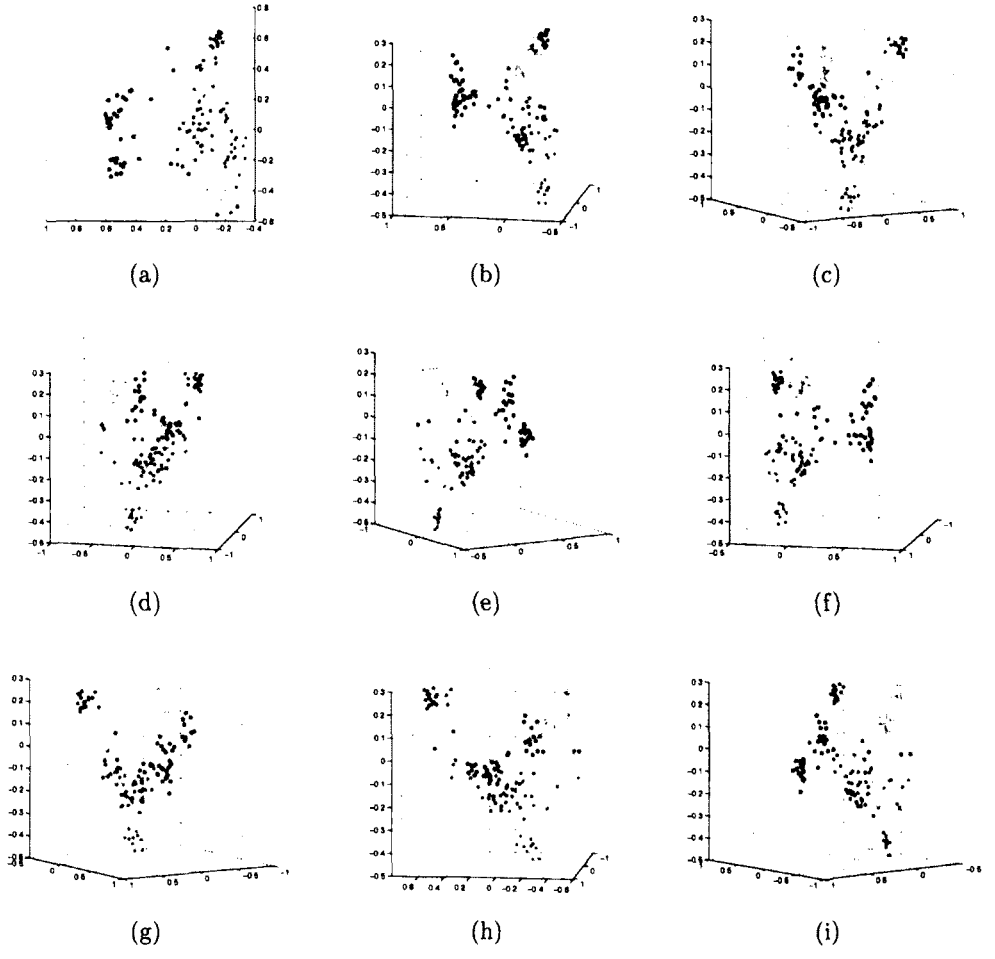


Figure 4.2: *Diffusion map of countries, given their IGO membership in 2000, using the* 2^{nd}, 3^{rd}, *and* 4^{th} *vectors. (a) provides a top-down perspective, while (b)-(i) show side views from different angles, in counterclockwise rotation. The countries are manually colored according to their geographical locations, which shows again that IGO's aligning influence is mostly regional. Legend (with respect to (a)): Caribean (dark blue, upper left); Central & South American (medium blue, lower left); Western European (light blue, upper right); former Soviet states & ISR (yellow, upper right); North African (light red, middle far right); African (light orange, lower right); Middle East (dark orange, middle right); USA & CAN (dark red, middle).*

4.3 UN Assembly voting patterns

The UN Roll Call data [9] for a particular year details for each resolution k, whether a country i voted yes, no, abstained, or was absent. Removing those countries who were absent from at least $1/2$ of the roll calls, we define the matrix:

$$\mathrm{INT}_{ij} = \frac{|\,\mathrm{VOTE}_{i==j}\,|}{|\,\mathrm{VOTE}_i\,|} \tag{4.3.1}$$

where $\mathrm{VOTE}_{i==j}$, denotes the set of issues where country i cast the same vote as j, and VOTE_i is the set of issues on which country i voted. The correlation kernel W follows Eq. 4.2.2 whereas C is the Pearson product-moment correlation matrix [82] of INT. Fig. 4.3 show the embedding result for the year 1965, whose political alignment was heavily influenced by Cold War politics.

We note that diffusion embeddings may reveal the same patterns as other techniques [83], but in the examples below they differ significantly.

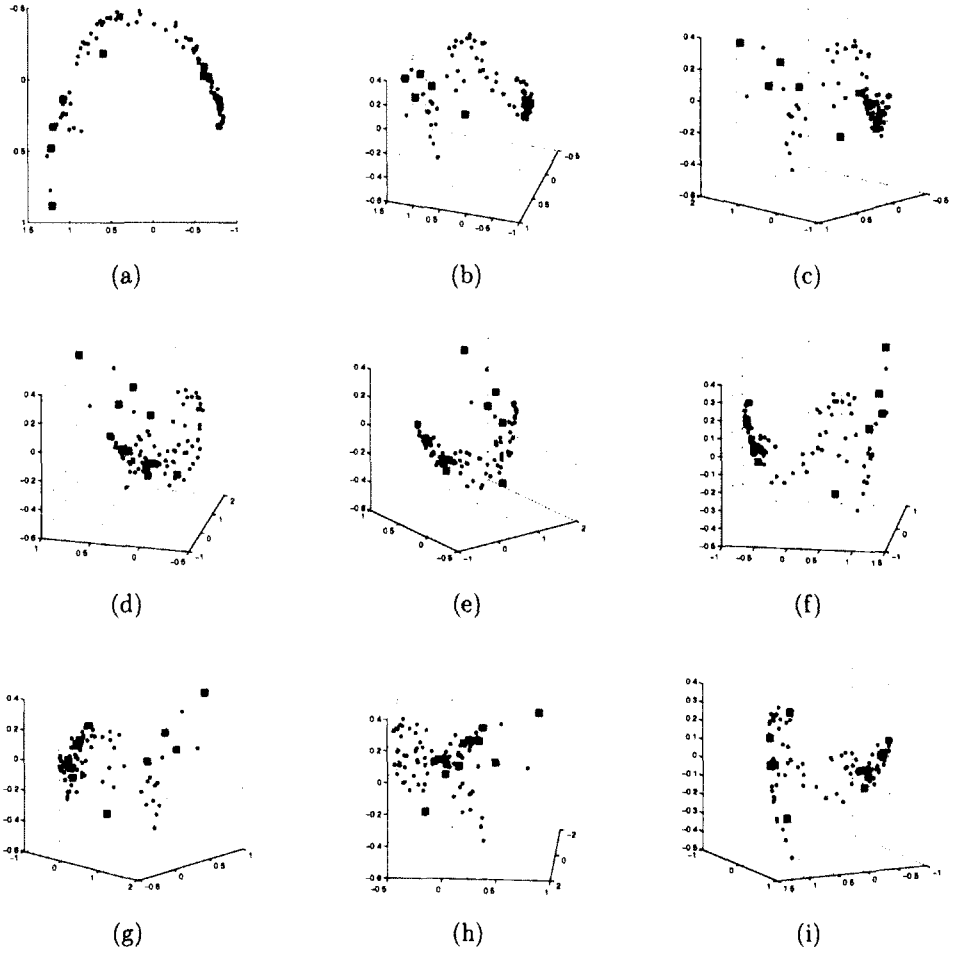


Figure 4.3: *Embedding of nations, based on their UN Voting pattern in 1965. (a) provides a top-down perspective, while (b)-(i) show side views from different angles rotated counterclockwise. Several countries are marked by colored square for easy identification:* ■ *(USA, UKG, FRN, BEL, ISR),* ■ *(RUS, CHN, POL, HUN, BLR),* ■ *(EGY, SYR, LEB, SAU, KUW). This Cold War embedding shows the world spread out on a spectrum with the Western powers on one end, and the Communists on the other end, while the Arabs positioned themselves close to the Eastern Bloc, in an anti-Western stance. Apparently, we can only see one red square in the map, because the whole Communist Bloc concurred with each other on almost every issue. The West, on the other hand, were more spread out.*

4.4 Case study: de Gaulle's France

Fig. 4.4 [3] shows the embedding of the network of UN Assembly members according to their voting patterns at various time during 1957-1975. More detailed maps of 1967 and 1970 are shown in Fig. 4.5 & Fig. 4.6 from different perspectives. This kind of visualization, in which social, political entities move around, attract and repulse each other in a planet-like manner, provides us with a novel historical perspective.

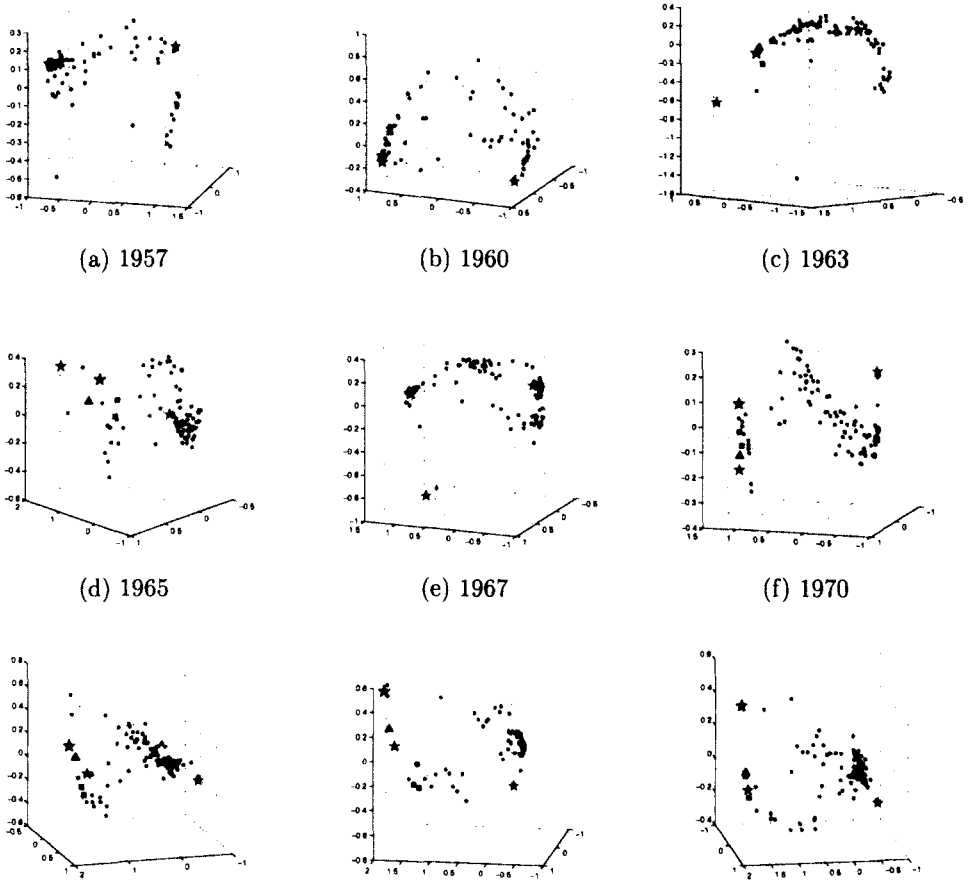
Additionally, Fig. 4.7 plots the ratios of embedding distance in the period 1965-2000:

- $\frac{\text{dist(FRN,EU*)}}{\text{diam(EU*)}}$ as the blue line
- $\frac{\text{dist(UKG,EU*)}}{\text{diam(EU*)}}$ as the red line
- $\frac{\text{dist(FRN,UKG)}}{\text{diam(EU*)}}$ as the green line

where EU* is defined as the states of the European Community, excluding FRN & UK. The distances and diameters are calculated from diffusion distance, distance of PCA embedding, and Hamming distance of the VOTE matrix. The plots of different distance measures show us how diffusion method amplifies the connections between highly connected actors, and also enhances separation between distant parties.

France's self-isolation under de Gaulle's presidency is apparent from the diffusion maps. In 1957 (Fig. 4.4a), France (cyan star, upper left corner) was close to the USA, UK, Belgium, Luxembourg (blue markers). By 1959, France under Charles de Gaulle began to withdraw from NATO military commands and completed that process in 1966. Thus, when we look at the maps as time proceeds, we see France slowly move to the edge of the (blue) Western group in 1960 (Fig. 4.4b), gradually edging further away by 1963 (Fig. 4.4c), planting itself in a distant position from that of the West in

[3] See clip 3 on http://www.cs.yale.edu/homes/vision/zucker/embeddings.html.



(a) 1957 (b) 1960 (c) 1963

(d) 1965 (e) 1967 (f) 1970

(g) 1972 (h) 1973 (i) 1975

Figure 4.4: *De Gaulle's France: Diffusion maps of UN voting pattern 1957-1975. Several countries are marked for case study identification:* ★*(USA),* ▲ *(UKG),* ★ *(FRN),* ■ *(BEL, LUX, GFR),* ★ *(RUS). These maps show France started out close to the Allies in 1957. Then in 1960, France, under de Gaulle's presidency, distanced itsef from the West. The 70s saw France coming back toward the Western fold, once de Gaulle had left.*

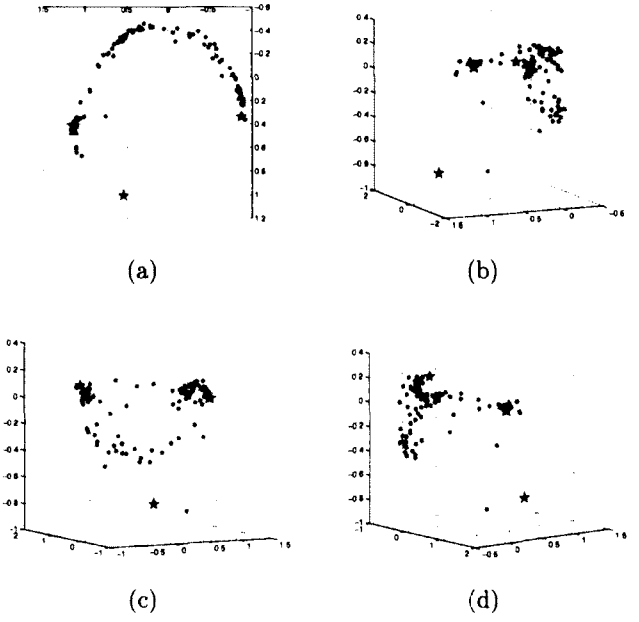
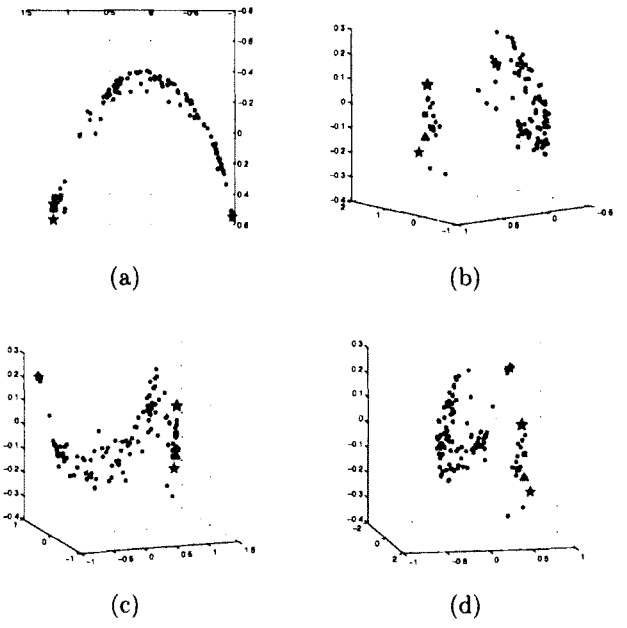
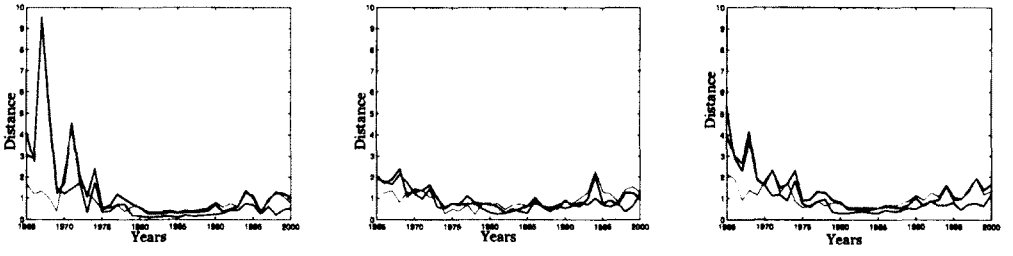


Figure 4.5: *Diffusion map of UN voting pattern in 1967 from different angles. Several countries are marked for case study identification:* ★*(USA),* ▲ *(UKG),* ★ *(FRN),* ■ *(BEL, LUX, GFR),* ★ *(RUS). The map is shown here from top down (a) then from side view with angles rotated counterclockwise (b)-(d).*



(a) (b) (c) (d)

Figure 4.6: *Diffusion map of UN voting pattern in 1970 from different angles. Several countries are marked for case study identification:* ★*(USA),* ▲ *(UKG),* ★ *(FRN),* ■ *(BEL, LUX, GFR),* ★ *(RUS). The map is shown here from top down (a) then from side view with angles rotated counterclockwise (b)-(d).*



(a) Diffusion distance (b) Distance of PCA embedding (c) Hamming distance

Figure 4.7: *Ratios of embedding distances between FRN-EU* (blue), UKG-EU* (red), FRN-UKG (green) in 1965-2000. Here EU* is defined as the states of the European Community, excluding FRN & UK. Thes plots show how relations between France, UK and the rest of the Western European states changed over time, with France standing far apart during the 60s, and coming back to the fold afterward.*

1967 (Fig. 4.4e). The distance ratio plot in Fig. 4.7a shows us the blue line (FRN-EU) started at around 0.8, the green line (FRN-UKG) reaching its peak at 9 in 1967-1968, while the red line (UKG-EU) lying low initially, indicating France's isolated position from that of the Western countries (and UKG) at the time. After de Gaulle left office in 1969, we see the blue line begin to decline steeply, moving in tandem with the red line, implying a reverse course in France' foreign policy, gradually edging closer to that of the rest of West. Indeed, Fig. 4.4f shows France (cyan star, bottom left) moving back toward integration in NATO, its position in 1972-1973 (Fig. 4.4g-4.4h) got closer and closer to that of UKG (blue triangle, top left) (FRN opened up from its self-isolation, allowing UKG to join EC in 1973). By 1975 (Fig. 4.4i) France again stood close to the Western bloc. In the 80s until the end of the Cold War, the distance ratios FRN-EU and UKG-EU (blue & red lines, Fig. 4.7a) ascended slightly, due to the absorption of new members into the EU. The green line (FRNK-UKG), however, remains low throughout the 80s, showing how close FRN and UKG's policies were to each other during that period.

An advantage of the diffusion maps is that they reveal the inherently low dimensional structure among countries, in agreement with prior analysis [19, 84]. It is also apparent from Fig. 4.7 that PCA fails to discover a pattern in the movements of countries in the network, while diffusion distance uncovers the same pattern as the simple Hamming distance. The spectrum given by PCA decays very slowly: it requires 20-30 dimensions to describe all variances in the voting data. Diffusion method, on the other hand, requires only 5-7 dimensions to describe the voting patterns [19]. The diffusion method performs better in amplifying significant events in its distance plot (e.g. the period from 1957-1967 in which France isolated itself). However, the diffusion distance in Fig. 4.7 is computed from only 5 dimensions, whereas the Hamming distance is the aggregated result of votes on all UN resolutions in a particular year.

4.5 Japan and the two Chinas

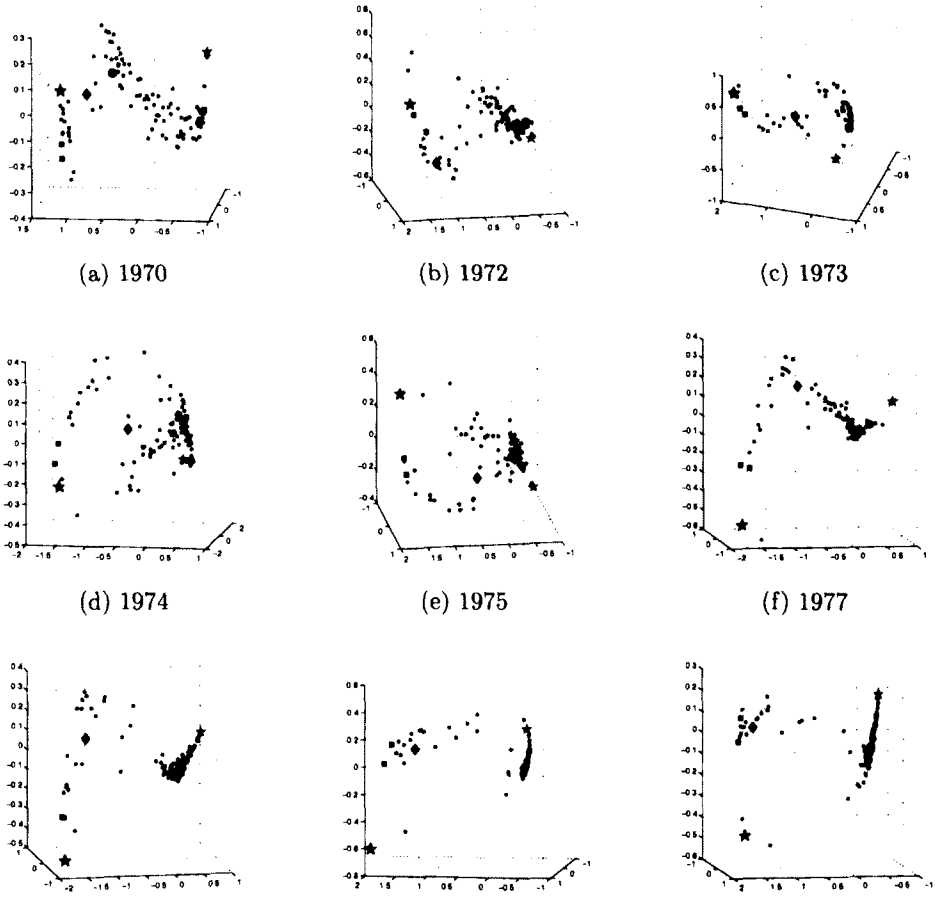
In 1970 (Fig. 4.8a), Taiwan (blue circle) stood close to the Western bloc on the left of the distribution of countries, while Soviet Union and Arab countries planted themselves on the far right, surrounded by third world nations. However, Nixon's visit to China in 1972 prompted Taiwan to move closer to third world countries, courting their votes in order to retain its seat at the UN. Indeed, Fig. 4.8b-4.8c for 1972 & 1973 show Taiwan's votes on the right side of the maps, mingling with the Arab and the third world. Its attempt failed, however, as Taiwan was out of the picture in 1974 (Fig. 4.8d), being replaced by PRC China (red circle).

Prior to 1973 (Figs. 4.8a-4.8b), Japan (blue diamond) positioned itself on the left of the map, close to the Western bloc. However, in response to the 1973 oil crisis (Fig. 4.8c), Japan distanced itself from US & allies and their Middle East policy, moving closer to the oil exporting states. Japan's ambition to Security Council membership also prompted it to seek better relations with other third world nations. We see Japan drifting away from the West, toward the center in 1973 until mid-late 70s (Figs. 4.8d-4.8f). We also see Japan edging back westward in the 80s (Figs. 4.8g-4.8i), moving to the left of the maps, again standing among the Western bloc countries.

4.5.1 Case study: Tito-Stalin split

In the maps of Fig. 4.9 [4], the Soviet Union is represented as a red star, and Yugoslavia as a red triangle. Poland is also shown as a red square, which, in contrast with Yugoslavia, stayed close to Soviet's position throughout the period 1948-1970. By 1948 (Fig. 4.9a), the relation between Yugoslavia and the Soviet Union was still amiable. The break between them occurred in 1949 (Fig. 4.9b), as Yugoslavia suddenly moved away from the Soviet Union and Poland. The split became more evident

[4]See clip 5 on http://www.cs.yale.edu/homes/vision/zucker/embeddings.html.

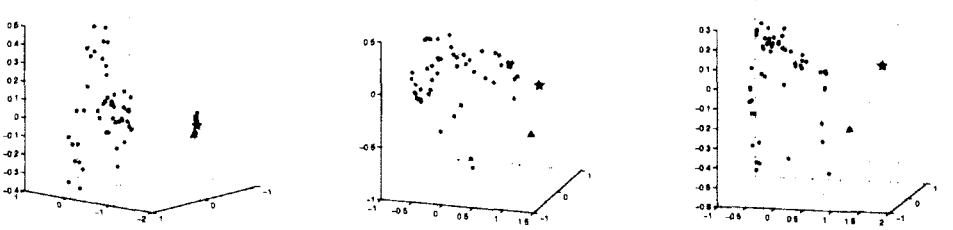


(a) 1970 (b) 1972 (c) 1973

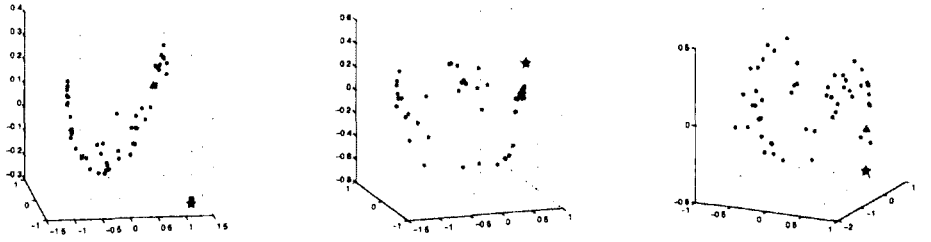
(d) 1974 (e) 1975 (f) 1977

(g) 1980 (h) 1982 (i) 1985

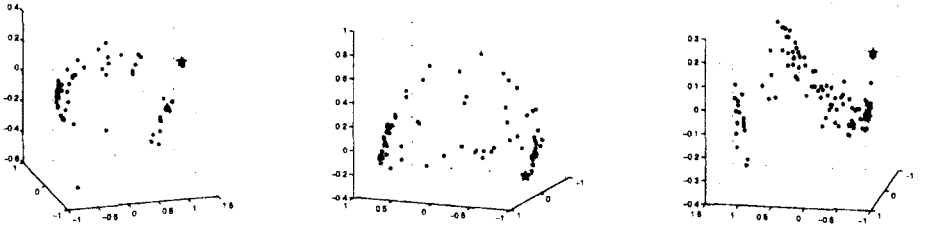
Figure 4.8: *Japan and the two Chinas: Diffusion maps of UN voting pattern 1970-1985. Several countries are marked for case study identification:* ★ *(USA),* ■ *(UKG, FRN),* ♦ *(JPN),* • *(TAW),* ★ *(RUS),* • *(CHN),* ■ *(EGY, SYR). These maps show Taiwan started out on the Western camp, then moved eastward seeking support from the Third World nations in order to prevent its replacement by PRC China in 1974. Japan also started in the Western camp, then moved away from US and Allies, toward oil exporting countries, and other Third World nations in its bid for Security Council membership in the late 70s, only to rejoin the West again in the 80s.*



(a) 1948 (b) 1949 (c) 1950

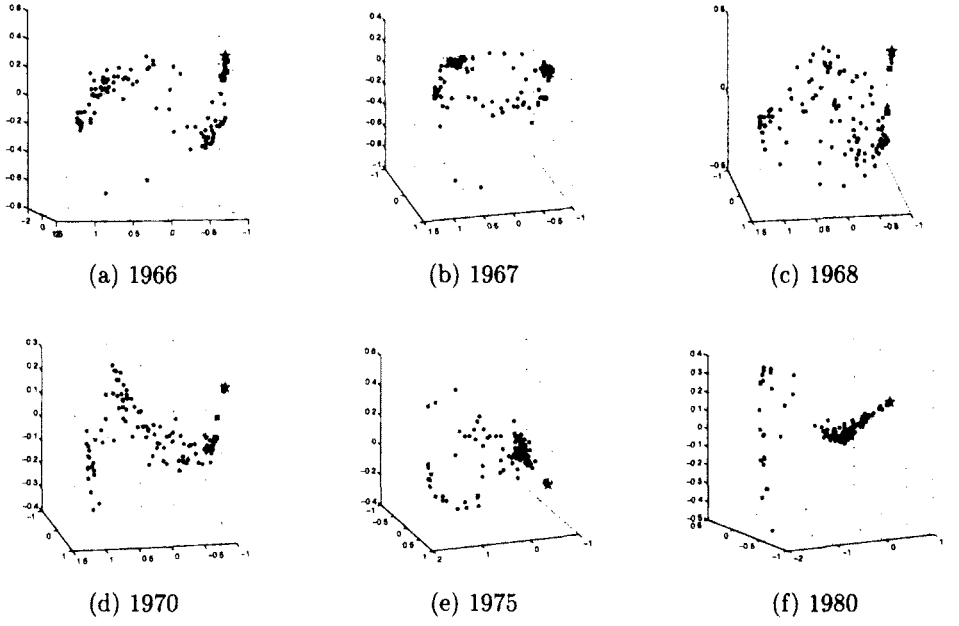


(d) 1952 (e) 1953 (f) 1955



(g) 1957 (h) 1960 (i) 1970

Figure 4.9: *Tito-Stalin split: Diffusion maps of UN voting pattern 1948-1970. Several countries are marked for case study identification:* ★ *(RUS),* ■ *(POL),* ▲ *(YUG). These maps show the ups and downs in the camaraderie between Yugoslavia and the Soviet Union. The close cooperation broke down in 1949, and continue to deteriorate from there. The relationship was briefly (and unsuccessfully) rekindled after Stalin's death in 1953 as Soviet 'caught up' with the rest of the world, and tried to court Yugoslava back.*



(a) 1966 (b) 1967 (c) 1968

(d) 1970 (e) 1975 (f) 1980

Figure 4.10: *The case of Romania: Diffusion maps of UN voting pattern 1966-1980. Several countries are marked for case study identification:* ★ *(RUS),* ■ *(ROM),* ■ *(EGY, SYR). These maps demonstrate Romania's pursuit of independent policies, keeping a more neutral stance, away from that of the Soviet Union, without actually leaving the Communist bloc.*

as Yugoslavia drifted even further away in 1950 (Fig. 4.9c). By 1952 (Fig. 4.9d), the Soviet bloc was mostly isolated from the rest of the world. By 1953 (Fig. 4.9e), after Stalin's death, Soviet attempted to court Yugoslavia back to the fold of its influence. Therefore, Soviet moved up to join the main curve, and at the same time, getting closer to Yugoslavia. However, since then, Yugoslavia and Soviet Union drifted apart again, as observed in Figs. 4.9f-4.9i.

4.5.2 Case study: Soviet-Romania relation

From the mid-60s onward, Romania pursued a foreign policy somewhat independent of the Soviet Union, without really leaving the Soviet bloc. It stayed close to the Arabs and other fairly radical Third World countries. Fig. 4.10a begins with Romania (red square) staying close to the Soviet Union (red star) and the Arab countries (Egypt

and Syria as green squares). Then in 1967-1968, we see Romania moved away from the Soviet bloc (Figs. 4.10b-4.10c). In the 70s (Figs. 4.10d-4.10f), when the distance between the Soviet bloc and the Arab world increases, Romania always stayed a step away from the Soviet bloc, and closer to the Arab countries [5].

4.5.3 Case study: US-Israel relation

Fig. 4.11 [6] shows the diffusion maps of nations by their UN voting patterns at various time during 1966-1995. Fig. 4.12 plots the ratios of embedding distance in the period 1965-2000:

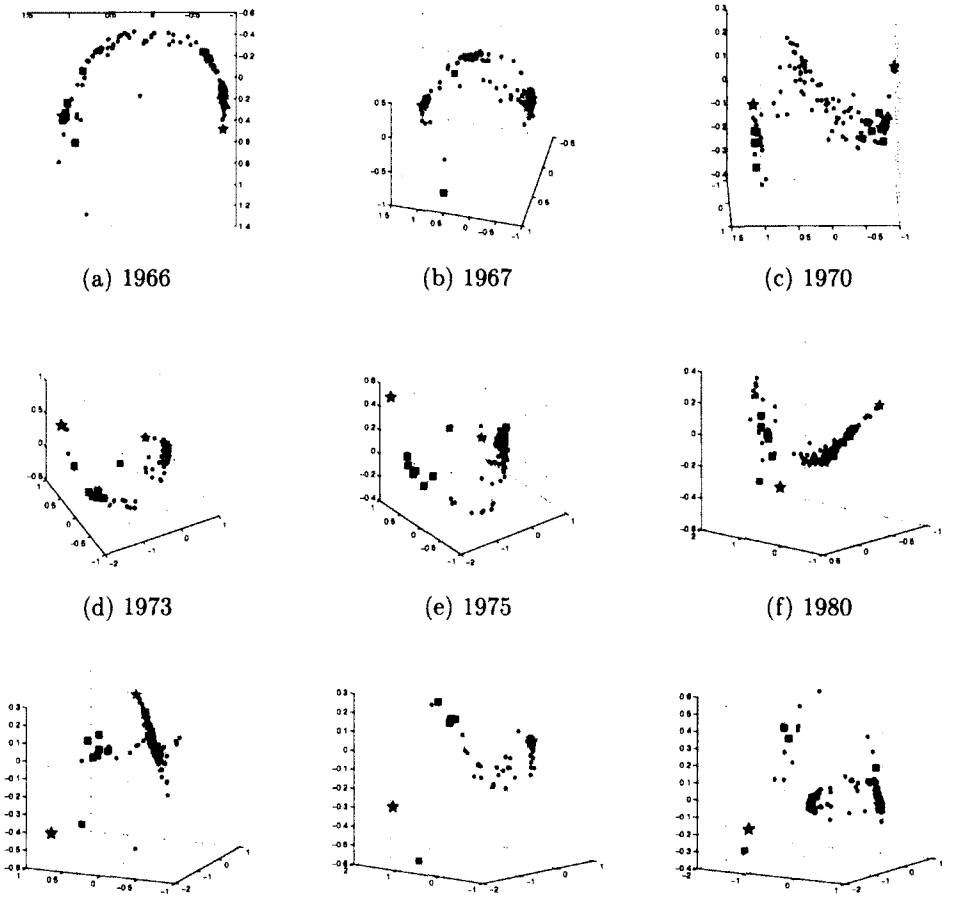
- $\frac{\text{dist(USA,EU*)}}{\text{diam(EU*)}}$ as the blue line
- $\frac{\text{dist(ISR,EU*)}}{\text{diam(EU*)}}$ as the red line
- $\frac{\text{dist(USA,ISR)}}{\text{diam(EU*)}}$ as the green line

where EU* is defined as the states of the European Community, excluding FRN & UK. The distances and diameters are calculated from diffusion distance, distance of PCA embedding, and Jaccard distance of the VOTE matrix in the same manner as the plots in Fig. 3.3. Similarly, Fig. 4.13 compares the ratios of embedding distance in the period 1965-2000:

- $\frac{\text{dist(USA,ARAB*)}}{\text{diam(ARAB*)}}$ as the blue line
- $\frac{\text{dist(ISR,ARAB*)}}{\text{diam(ARAB*)}}$ as the red line
- $\frac{\text{dist(USA,ISR)}}{\text{diam(ARAB*)}}$ as the green line

where ARAB* is defined as the countries in the Arab League in 1945: IRQ, EGY, SYR, LEB, JOR, SAU, YEM.

[5] See clip 6 on http://www.cs.yale.edu/homes/vision/zucker/embeddings.html.
[6] See clip 7 on http://www.cs.yale.edu/homes/vision/zucker/embeddings.html.

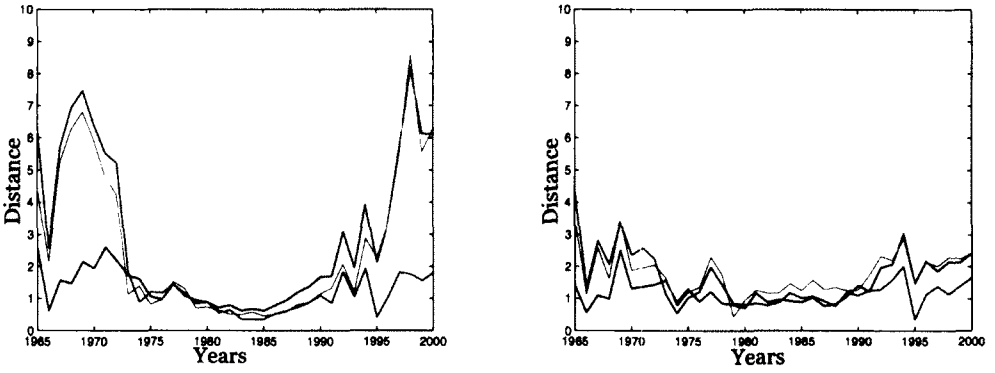


(a) 1966 (b) 1967 (c) 1970

(d) 1973 (e) 1975 (f) 1980

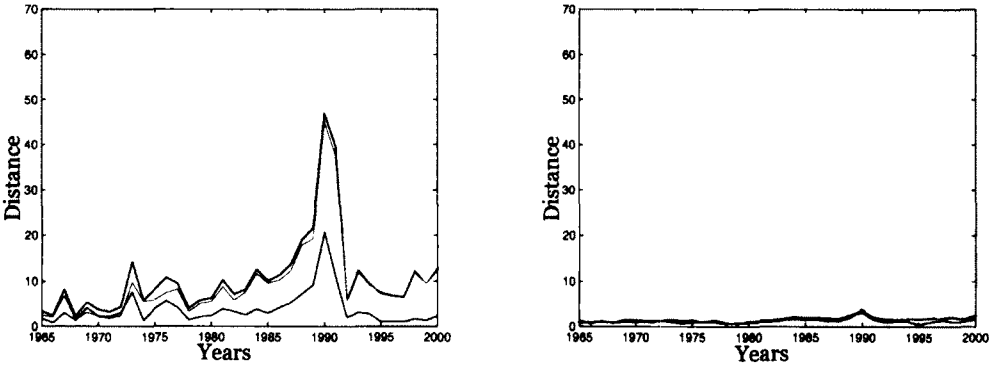
(g) 1985 (h) 1990 (i) 1995

Figure 4.11: *USA-Israel relation: Diffusion maps of UN voting pattern 1966-1995. Several countries are marked for case study identification:* ★ *(USA),* ■ *(EU nations),* ■ *(ISR),* ★ *(RUS),* • *(CHN),* ★ *(EGY),* ▲ *(SYR),* ■ *(Arab countries). These maps show how the USA-ISR relations evolve, getting closer and closer together over time.*



(a) Diffusion distance (b) Distance of PCA embedding

Figure 4.12: *Ratios of embedding distances between USA-EU* (blue), ISR-EU* (red), USA-ISR (green) in 1965-2000. Here EU* is defined as the states of the European Community, excluding FRN & UK. Thes plots show how relations between USA, ISR and the Western European states evolve over time.*



(a) Diffusion distance (b) Distance of PCA embedding

Figure 4.13: *Ratios of embedding distances between USA-ARAB* (blue), ISR-ARAB* (red), USA-ISR (green) in 1965-2000. Here EU* is defined as the countries in the Arab League in 1945: IRQ, EGY, SYR, LEB, JOR, SAU, YEM. Thes plots show how relations between USA, ISR and the Arab League evolve over time.*

The voting pattern of 1966 (Fig. 4.11a) begins with the world polarized by the West (blue) and the Soviet (red), while other countries spanned the spectrum in between. The Six Day War of 1967 (Fig. 4.11b) saw the Soviet-Arabs (red-green) alliance closing in. France under deGaulle imposed an arms embargo on Israel (cyan square), and adopted a more pro-Arab policy. France is shown staying away from the rest of the West in its stance. Hostile exchanges between Israel and the Arabs persisted throughout 1967-1970. At the same time, the US (blue star) and Israel began moving closer together, sharing military information with each other (Fig. 4.11c).

The year 1973 witnessed the Yom Kippur War, after which Israel emerged victorious but more isolated from the neutral countries (Fig. 4.11d). The result of the war led to the Arab oil embargo and subsequently the oil crisis 1973-1974, forcing other countries, including UK and France, to oppose US' support for Israel, in order to court OPEC's favor. Israel was most isolated from the US and Allies during 1970-1975 (Fig. 4.12a). On the other hand, Figs. 4.11d & 4.11e show the US being pushed into a corner, which Israel, isolated from others, was also slowly moving to. The situation slightly improved in 1980 with the signing of Camp David Accords and the peace treaty between Israel and Egypt. Fig. 4.11f also tells the fracture in the Soviet-Arabs relations after Soviet's invasion of Afghanistan in 1979.

The year 1985 (Fig. 4.11g) saw Israel being pushed further away from the main group, after its attack on Iraq's nuclear facility built by France in 1981, and its involvement in the First Lebanon War, which drew condemnation from the UN Assembly. However, with Reagan's personal support for Israel, US-Israel (green line, Fig. 4.12a) relation picked up from there, solidified by numerous strategic military cooperation agreements. The trend continued into 1990 (Fig. 4.11h) with Israel being granted major non-NATO ally status in 1989, while its position with the Arabs and others worsened following the First Intifada and Israel's raids into southern Lebanon in 1988.

The collapse of the Soviet Union in 1991 led to major changes in the world order, breaking up old alliances, and thus slightly improved the global map (Fig. 4.11i). As a result of Clinton's mediation, several peace treaties was signed between Israel and Arab countries. However, the trend started in the 80s by Reagan had already been established: the US firmly planted itself in a corner in total support of Israel, far away from the Third World goup, and even from the Western allies, as evident in Figs. 4.11i & 4.12a.

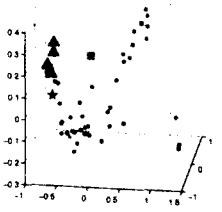
4.5.4 Case study: the Middle East conflict

Figs. 4.14-4.16 show the diffusion embeddings according to the UN voting patterns based on: (i) *all resolutions* (LHS figures) and (ii) *only Middle East-related resolutions* (RHS figures). The member nations are colored according to their votes in Resolution GA11317 (Nov 2012) granting non-member observer state status to Palestine: For, Against, & **Abstain**. We believe that the general embeddings in the left-column figures (a, c, e) show a trend in which the votes of USA and ISR getting ever closer to each other, and away from others. Additionally, the Middle East embeddings in the right-column figures (b, d, f) make an even stronger case, in which different stages of USA-ISR relations can be identified.

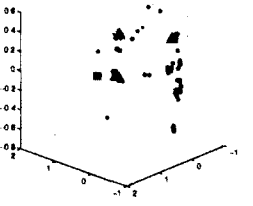
- Late '40s: ISR was mostly isolated (Figs. 4.14a-4.14b). Its positions was closer to the USA & the West than to the rest.

- '50s - Early '60s: The general embeddings in Figs. 4.14c,4.14e show ISR approaching the USA and the Western bloc. However, the Middle East embeddings in Figs. 4.14d,4.14f tell a slightly different story: FRN & UKG moved over to ISR's side during the Suez crisis 1956 while the USA still stood with the other nations in the Middle East issues.

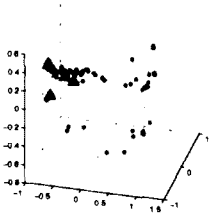
- Late '60s: Fig. 4.15a again shows ISR being isolated after the Yom Kippur



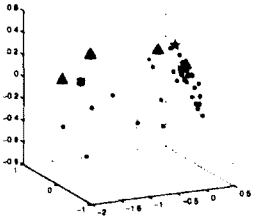
(a) All resolutions 1946-1955



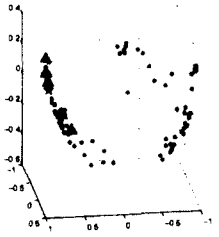
(b) Only Middle East resolutions 1946-1955



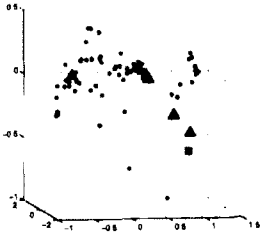
(c) All resolutions 1951-1960



(d) Only Middle East resolutions 1951-1960

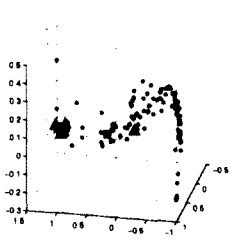


(e) All resolutions 1956-1965

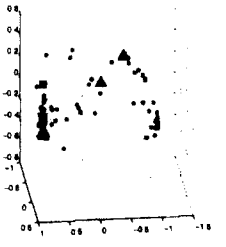


(f) Only Middle East resolutions 1956-1965

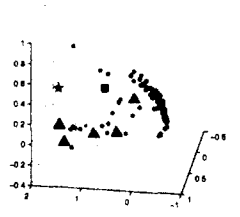
Figure 4.14: *Diffusion embeddings of the UN voting patterns in the late 40s untill the early 60s: LHS (a, c, e) using* **all resolutions** *in the specified periods; RHS (b, d, f) using* **only Middle East-related resolutions** *in the specified periods. UN member nations are colored by their votes in Resolution GA11317 (Nov 2012):* ***For****, Against, &* ***Abstain****. Several countries are marked by special symbols:* ★ *(USA),* ■ *(ISR),* ▲ *(UKG),* ▲ *(FRN, SPN, SWD, DEN).*



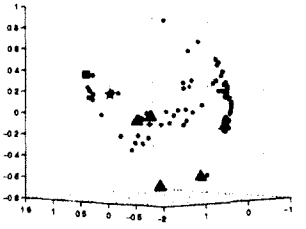
(a) All resolutions 1966-1970



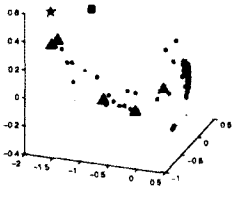
(b) Only Middle East resolutions 1966-1970



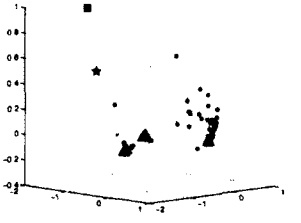
(c) All resolutions 1971-1975



(d) Only Middle East resolutions 1971-1975

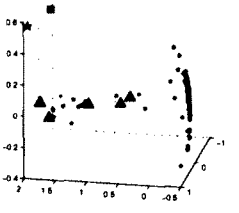


(e) All resolutions 1976-1980

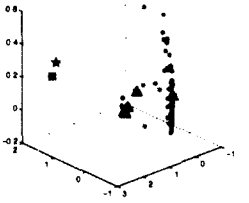


(f) Only Middle East resolutions 1976-1980

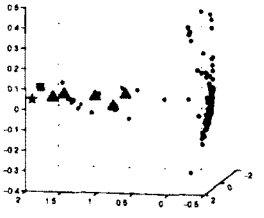
Figure 4.15: *Diffusion embeddings of the UN voting patterns in the late 60s through the 70s: LHS (a, c, e) using* **all resolutions** *in the specified periods; RHS (b, d, f) using* **only Middle East-related resolutions** *in the specified periods. UN member nations are colored by their votes in Resolution GA11317 (Nov 2012): For, Against, & Abstain. Several countries are marked by special symbols: ★ (USA), ■ (ISR), ▲ (UKG), ▲ (FRN, SPN, SWD, DEN).*



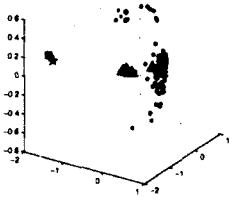
(a) All resolutions 1981-1985



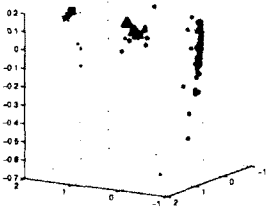
(b) Only Middle East resolutions 1981-1985



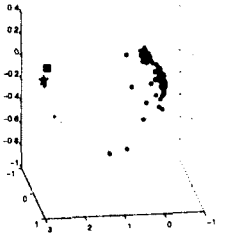
(c) All resolutions 1986-1995



(d) Only Middle East resolutions 1986-1995



(e) All resolutions 2001-2010



(f) Only Middle East resolutions 2001-2010

Figure 4.16: *Diffusion embeddings of the UN voting patterns in the 80s and beyond: LHS (a, c, e) using* **all resolutions** *in the specified periods; RHS (b, d, f) using* **only Middle East-related resolutions** *in the specified periods. UN member nations are colored by their votes in Resolution GA11317 (Nov 2012):* **For**, *Against*, *&* ***Abstain.*** *Several countries are marked by special symbols:* ★ *(USA),* ■ *(ISR),* ▲ *(UKG),* ▲ *(FRN, SPN, SWD, DEN).*

War in 1967. Nevertheless, Fig. 4.15b displays a change of attitude of the USA toward ISR under Johnson's presidency. FRN, however, withdrew its support after the 1967 war.

- Early '70s: In general, ISR appeared to be politicaly isolated after the Yom Kippur War and the oil crisis of 1973 (Fig. 4.15c). On the other hand, we may argue, by the ME embedding in Fig. 4.15d, that the USA under Nixon's presidency wholeheartedly supported ISR while other nations keep a safe distance away.

- Late '70s and early '90s: In general, Fig. 4.15e shows ISR closing on the Western bloc, including the USA, the UKG, and FRN. However, Fig. 4.15f reveals that the USA went all-out in support of ISR in the Middle East issues, away from the positions of other countries. This trend continues toward the early '90s, through the collapse of the USSR (Figs. 4.16a-4.16d). Both the general and ME embeddings aggree on the trend that USA-ISR relations were getting ever closer and perhaps, more and more isolated. However, there is a timing difference in these embeddings. The ME embeddings show that the USA slowly approached ISR in the ME issues, finally throwing its weight behind ISR in the early '70s, and has been standing by ISR's side in the ME conflict ever since. The general embeddings contend that the "honeymoon" in USA-ISR relations started in the late '80s. From this time lag, we may infer that, because USA chose to support ISR since the '70s, cementing a strong alliance, ISR has been following the USA's lead in every other international issues.

- 2001-2010: The color distributions in both Fig. 4.16e and Fig. 4.16f seem to display a clear cut on the Palestinian issue. The countries which abstained from voting in Resolution GA11317 mingled closely with the supporting countries. Of the 9 countries which voted against Resolution GA11317, Canada and Panama

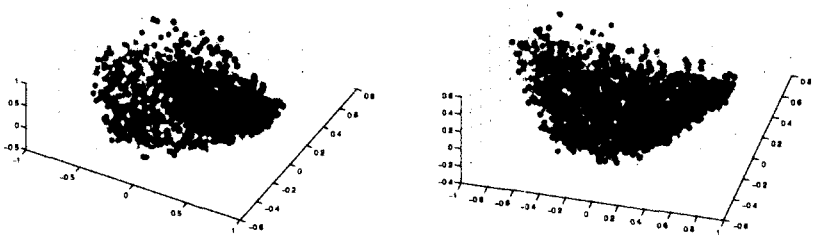
appear to be the reluctant ones, their positions being closer to the other camp.

Chapter 5

Diffusion maps in supervised applications

5.1 Effect of irrelevant features

Section 4.5.4 provides an example in which we have a labeling function f (in this case the votes on Resolution GA11317) and we would like to organize the countries in a way that reflects their opinions on GA11317. In other words, we want countries that are mapped closed to each other in Ψ_t to have similar values in f. The left-column maps in Figs. 4.14-4.16 merely plot the GA11317 votes on the diffusion embeddings computed from every resolution, on different issues, topics. The UN General Assembly vote on a wide range of issues, topics, each of them having many resolutions. In practice, countries may disagree on one topic and agree on another, which means resolutions of the same topic are likely to be strongly correlated (or strongly anti-correlated) while resolutions from different topics are less likely to be so. Since Resolution GA11317 is directly related to the Middle East conflict, the votes on resolutions of other issues are hardly relevant. As presented in Section 4.5.4, the right-column maps in Figs. 4.14-4.16 plot the GA11317 votes on the diffusion em-



(a) MMPI Scale 2 (D/Depression) (b) MMPI Scale RC4 (Antisocial Behavior)

Figure 5.1: *Diffusion maps of the MMPI dataset [3, 4] using all questions with a Gaussian kernel, using the first 3 eigenfunctions. The people are colored based on their scores on (a) Scale 2 (D/Depression) and (b) Scale RC4 (Antisocial Behavior)*

beddings computed from only resolutions on Middle East issues, which reveal more nuanced, and at times different representations of the relations between countries in the Middle East conflict, more relevant to the targeted Resolution GA11317.

We argue that: *in supervised context, diffusion method may fail to find structure which agrees with a particular function of interest f, from datasets such as surveys, questionnaires, ratings, voting records.* These types of datasets, either by design or by nature, usually consist of multiple groups of features, most of which are irrelevant to f, both semantically and statistically. However, because features of the same groups are likely to be strongly correlated (or strongly anti-correlated), the irrelevant features tend to form consistent patterns in the observation space X, which may interfere with the discovery of any other useful structure in X that is relevant to f.

Fig. 5.1 further demonstrates the effect of irrelevant features in the Minnesota Multiphasic Personality Inventory (MMPI) dataset. The MMPI is a standardized, highly used psychometric test of adult personality and psychopathology [3], that has been in its revised version, the MMPI-II for 13 years [4, 85]. Over the past 60+ years, this item set has been widely accessed and well researched with a wide variety of different clinical and nonclincal samples [41, 42, 86–89]. The sample MMPI-2 dataset in our

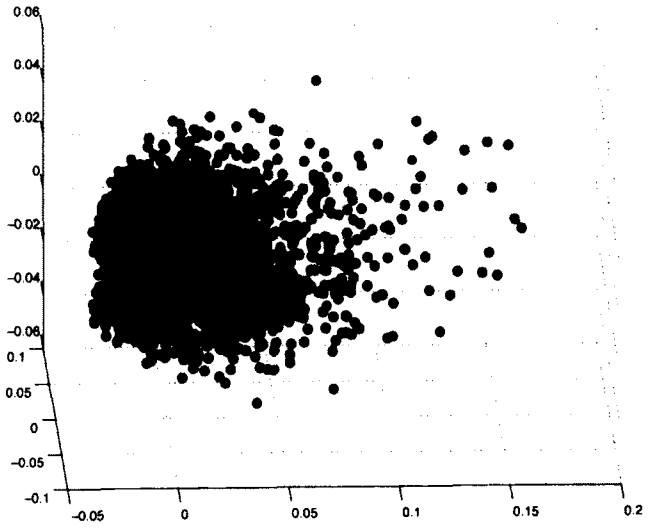


Figure 5.2: *Diffusion maps, using all questions from the NESARC dataset, with the first 3 eigenfunctions from a Gaussian kernel. The people are colored by their responses to question S2BQ1A21 - More than once drive a vehicle while drinking (W1).*

experiment contains a survey response $N \times M$ matrix A of $N = 2428$ persons answering $M = 567$ questions. We applied a Gaussian kernel $k(x_i, x_j) = e^{(-frac||A_i - A_j||^2 \beta^2)}$ where A_i, A_j are response vectors on all 567 questions of participants x_i, x_j respectively. Fig. 5.1 plots the diffusion maps for two of its restructured clinical scales, the widely used Depression and Antisocial Behavior scales. The smooth transition from cool color (low value) to hot color (high value) in Fig. 5.1a shows that the diffusion maps using all questions can organize people in this particular data sample according to their MMPI Scale 2 (D/Depression) scores. However, it performs rather badly in organizing people according to their Restructured Clinical Scale RC4 (Antisocial Behavior) scores (Fig. 5.1b).

Fig. 5.2 presents another demonstration of the same phenomenon using all questions currently available from the National Epidemiologic Survey of Alcohol and Related Conditions (NESARC) dataset [5, 6]. It is the largest comorbidity study, covering questions on alcohol consumption, tobacco and drug uses, psychological disorders and other problems using a national population sample. Because of both its wide

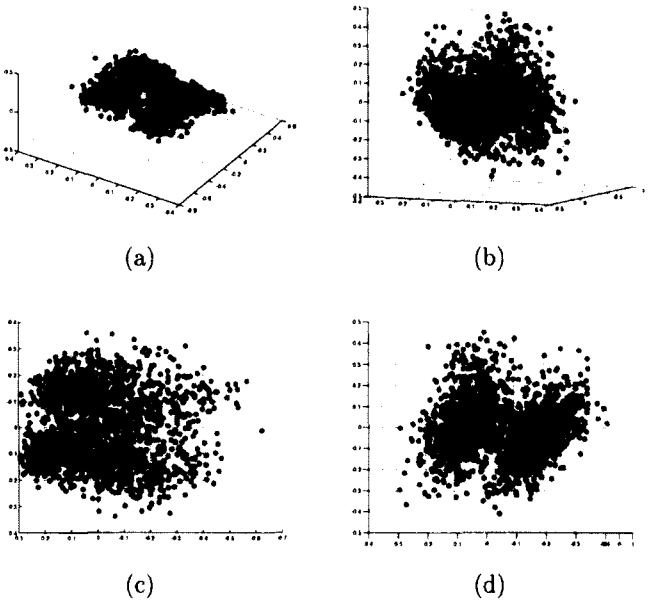
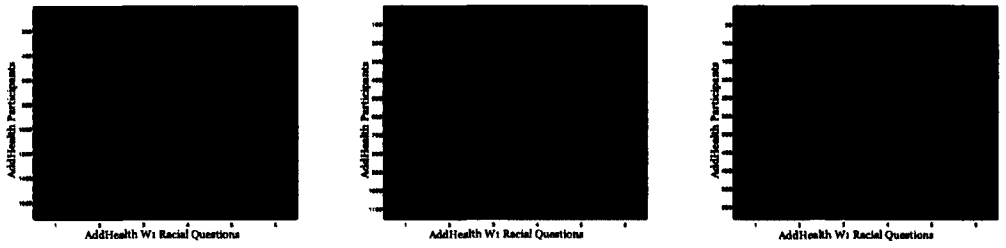


Figure 5.3: *Diffusion embedding of AddHealth Wave-1 questionnaire participants, using the first 3 eigenfunctions. A Gaussian kernel was used. Participants are colored based on the alcohol questions of BINGE DRINKING FREQUENCY at W-2, red corresponds to cases where respondents have moved from non-binging status (in this case -from non-drinking at baseline) to binging (having had at least one bingeing incident in the past year (5 or more alcoholic drinks on an occasion), at follow-up, a year later; whereas blue to non-binging cases. The subfigures are of the same 3-dimensional map from different viewing angles.*

item pool and the representativeness of the sample, this dataset has attracted lots of attention from the social science and public health research community [43, 90–95]. Due to the large size of the dataset, 5000 persons (out of 43000) were randomly picked from the whole population of the dataset. Their corresponding points, colored by their responses to question S2BQ1A21 (More than once drive a vehicle while drinking), appear to be completely unorganized with respect to the labeling function S2BQ1A21.

As another example, we applied diffusion maps to the national sample Adolescent Health Longitudinal Study (ADDHealth) Wave 1 questionnaire [7, 8], which surveyed students at age 13-14 (Wave-1) in order to identify alcohol abuse behavior (e.g. bing-

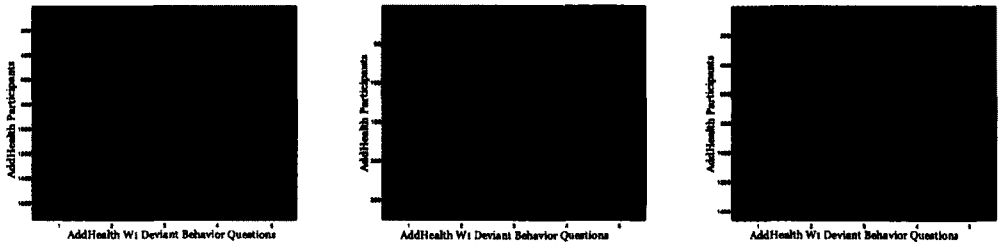


(a) (b) (c)

Figure 5.4: *AddHealth W1 questionnaire responses on racial questions. (a) Responses of all 1735 participants on 6 racial questions; We then applied diffusion methods to the responses in (a) and used the first eigenfunction to divide the participants into the negative end (b) and positive end (c) of the first eigenfunction. Note the opposite patterns of responses between (b) and (c).*

ing) a year later (Wave-2). Ample research have been carried out on this dataset to study and predict adolescent substance use, mental health and other issues [44, 96–98]. The dataset composes of 1735 persons $\times$ 374 questions, most of which are binary. We first zero-centered every question, replacing all the blank values with 0. The zero-centered matrix D was re-normalized such that every question has answers ranging from 0 to 1. We then computed the Gaussian kernel, using $\beta = 100$. Here the function f is the binary function which separates 190 bingers (10.95%) from the remaining non-bingers. Fig. 5.3 shows the diffusion map of questionnaire participants using the first 3 eigenfunctions. The participants are shown as blue and red dots on the map. Red dots correspond to bingers while blue dots are non-bingers. The only recognizable structure, using the first 3 eigenfunctions, is 4 big clusters (Fig. 5.3a), which are divided according to races (Caucasian vs. non-Caucasian) and gender (male vs. female). Other eigenfunctions do not present any discernible structure. The bingers appear to be randomly distributed.

The questionnaire dataset can be shown to consist of smaller groups of questions which are correlated (or anti-correlated) to each other in various degrees. Fig. 5.4 focuses on the group of racial questions:



(a) (b) (c)

Figure 5.5: *AddHealth W1 questionnaire responses on questions about deviant behaviors. (a) Responses of all 1735 participants on 5 deviant behavior questions; We then applied diffusion methods to the responses in (a) and used the first eigenfunction to divide the participants into the negative end (b) and positive end (c) of the first eigenfunction. Note the opposite patterns of responses between (b) and (c).*

- 6.1 WHAT IS YOUR RACE (WHITE)
- 6.2 WHAT IS YOUR RACE? (BLACK)
- S1Q6A RACE-WHITE-W1
- S1Q6B RACE-AFRICAN AMERICAN-W1
- A6_1 RACE, WHITE-PQ
- A6_2 RACE, BLACK/AFRICAN AM.-PQ

while Fig. 5.5 focuses on the group of questions about deviant behaviors:

- S29Q12 SELL DRUGS-W1
- S24Q1 EVER HAVE SEX-W1
- S5Q9 EXPELLED FROM SCHOOL-W1
- S26Q5 NON-ROMANCE SEX W/ ANYONE-W1
- S5Q7 RECEIVED OUT-OF-SCHL SUSPENSION-W1

The absolute values of correlation among the first group is at least 0.6, and the correlation among the second group is 0.26 on average.

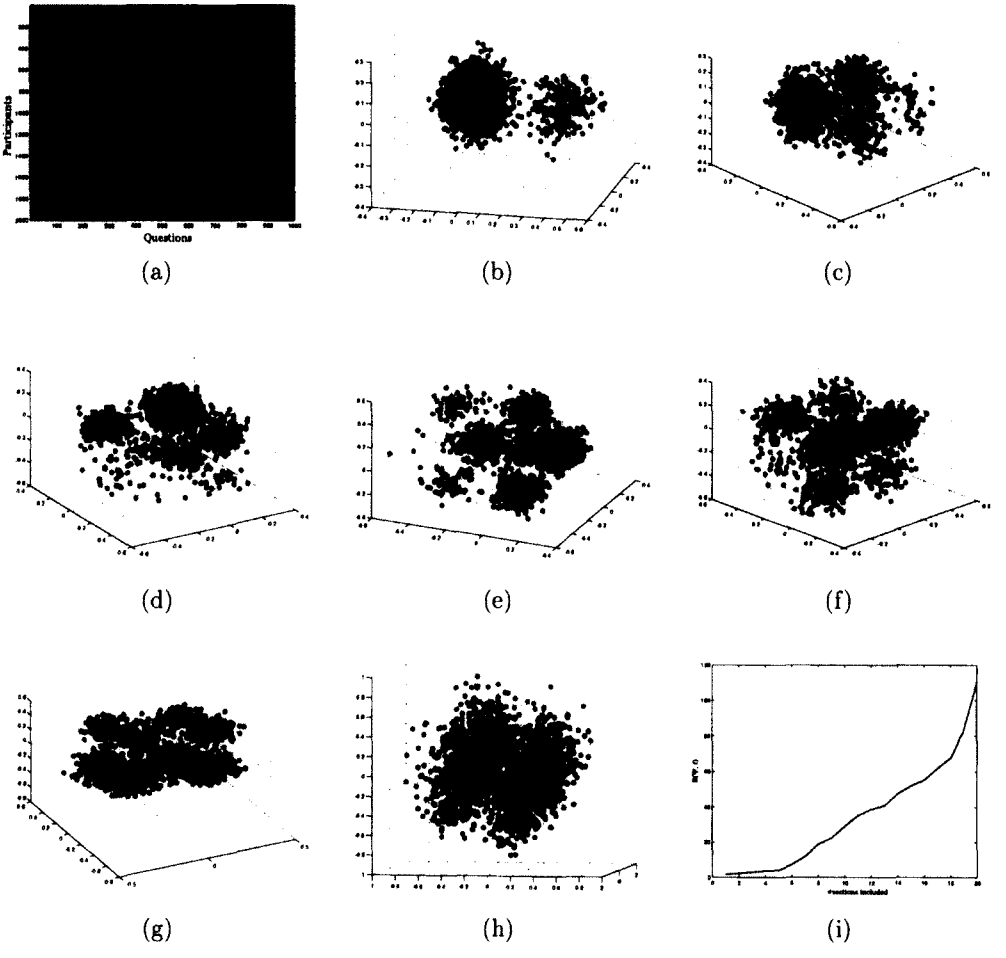
5.2 Synthetic Experiments

We tested this effect of irrelevant features in synthetic dataset of multiple sections, each represents a topical group of binary questions. To replicate the behavior of the real data, we investigated the responses in Fig. 5.5a, built a Gaussian kernel from it. The first eigenfunction divides the participants into two groups which have opposite patterns of response values, as shown in Figs. 5.4b & 5.4c. This property holds true for the group of deviant behavior questions in Figs. 5.5b & 5.5c, and for other group of correlated questions. Thus, given $N = |X|$ as the number of participants and $\delta \in (0, 0.5)$ as Bernoulli mean parameter, for each section S_t of the synthetic data, we first randomly divide the population of n participants into two mutually exclusive groups X_t^a and X_t^b (of random proportions): $X_t^a \cup X_t^b = X, X_t^a \cap X_t^b = \varnothing$. Let χ_t be the identifying function on the participants according to the population division in section S_t such that $\chi_t(i) = 1\ \forall x_i \in X_t^a$ and $\chi_t(i) = 0\ \forall x_i \in X_t^b$. The number of questions m_t in section S_t is a function of δ. The answer of participant x_i to question q_j is modeled as:

$$\begin{cases} A_{ij} \sim \mathrm{B}(1, \delta) & \text{if } q_j \in S_t \ \&\ x_i \in X_t^a \\ A_{ij} \sim \mathrm{B}(1, 1-\delta) & \text{if } q_j \in S_t \ \&\ x_i \in X_t^b \end{cases} \tag{5.2.1}$$

Fig. 5.6a shows the synthetic questionnaire datasets of 2000 participants, 20 sections, each contains 50 binary questions. The Bernoulli mean parameter δ is set at 0.3. We used the first section S_1's population division χ_1 as the control dimension (i.e. assuming that χ_1 divides bingers from non-bingers.)

Therefore, let $f = \chi_1$. Figs. 5.6b-5.6h plot f on different diffusion embeddings. These embeddings were computed from Gaussian kernels based on increasing number of question sections. As the number of questionnaire sections climbs up, the red points (bingers) gradually disperse among the blue points (non-bingers). We define an error function $\Xi(\Psi, f)$ which measures the smoothness of f on the embedding Ψ:

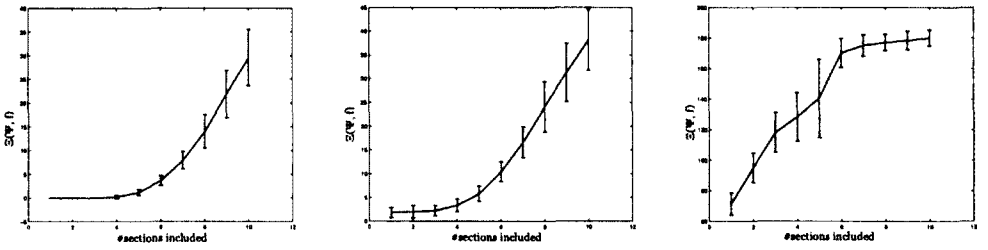


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 5.6: *Experimental results on (a) synthetic questionnaire of 2000 participants, $\delta = 0.3$, 20 sections of 50 questions each. Let $f = \chi_1$, the identifying function of question section S_1. Following are the plots of f on different diffusion maps, computed from Gaussian kernels based on sections (b) S_1 only; (c) S_1, S_2; (d) S_1, S_2, S_3; (e) $S_1, \cdots, S_4$; (f) $S_1, \cdots, S_5$; (g) $S_1, \cdots, S_{10}$; (h) $S_1, \cdots, S_{20}$ (every sections of questions in the dataset). (i) plots the error function Ξ with respect to f of the diffusion embeddings as more sections are used.*



(a) $\delta = 0.1, m_t = 20\ \forall S_t$ (b) $\delta = 0.3, m_t = 50\ \forall S_t$ (c) $\delta = 0.45, m_t = 250\ \forall S_t$

Figure 5.7: *Experimental results on synthetic questionnaire of 2000 participants,* $f = \chi_1$*, the identifying function of question section* S_1*. The plots show the error function* Ξ *as the number of questionnaire sections increases, under different settings of* δ *and* m_t*.*

$$\Xi(\Psi, f) = \sum_{i=1}^{N} (f(x_i) - \bar{f}_\Psi(x_i))^2 \tag{5.2.2}$$

where $\bar{f}_\Psi(x_i)$ is the average of f on the ν-nearest neighborhood of x_i, weighted by inverse diffusion distances. By default, we set $\nu = 10$. Fig. 5.6i plots the values of Ξ as the number of questionnaire sections used to compute the diffusion embeddings increases. The error curve steadily climbs up, implying that as more irrelevant features were included in the computation, the information given by S_1 was diluted, so that more and more bingers (red point) end up in neighborhoods consisting mostly of non-bingers (blue points). The error function Ξ plots in Fig. 5.7 under different settings of δ, in repeated experiments, shows the same characteristic.

5.3 Correlation hierarchical clustering of features

We have shown empirically that, under a supervised context, diffusion method is vulnerable to irrelevant features. We also showed that correlated patterns exist in the feature space of survey and voting record datasets. Our first attempt in selecting relevant features is to tackle the structure of the feature space, which will hopefully provide further information on the clustering of observations. In specific, we seek

to organize the resolutions according to how countries voted on them, or the survey questions according to how people answered them, with the goal of uncovering themes that summarize them. Given the lack of a prior on themes among resolutions and questions – how many there are or, even, whether any exist – we adapt a hierarchical clustering algorithm. For each cluster in the hierarchy, we seek a set of "summary questions" that best approximate large groups of questions underlying the embeddings. This reduces the dimension of the data set; and if the summary questions are combinations of small numbers of questions, they are more interpretable. Different from factor analysis, which is a top-down approach leading to factors that are linear combinations of all questions, ours is a bottom-up approach, grouping and summarizing questions as we go, according to the correlations within the "local groups".

5.3.1 Algorithm

Any pair of resolutions are related if they are highly correlated either positively or negatively. For example, during the Cold War period, a UN resolution condemning Israel in Middle East issues will most likely be rejected by the West and supported by the Arabs; however, another UN resolution in support of Israel would lead to the exact opposite voting pattern. Therefore, we study the absolute value of data correlation as a topical similarity function. More formally, this leads to a relatively standard objective function that only depends on dot products. It can be modified using the kernel trick to incorporate non-linearities, in particular those that arise with our diffusion kernel.

We treat each resolution as a vector of responses $\boldsymbol{q}_i$ normalized so $\sum_j \boldsymbol{q}_i(j) = 0$ and $\|\boldsymbol{q}_i\| = 1$. We denote $Q = \{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n\}$, the set of votes to all resolutions. On the way to designing an objective function, we first seek to find a set of "summary questions" $S = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k\}$ and a clustering $C = \{c_1, \ldots, c_k\}$ of questions with

summary questions with the following properties:

$$\bigcup_{i=1}^{k} c_i = Q \tag{5.3.1}$$

$$c_i \cap c_j = \varnothing, i \neq j \tag{5.3.2}$$

$$\|\boldsymbol{s}_i\| = 1 \tag{5.3.3}$$

Equations (5.3.1) and (5.3.2) guarantee that each question is assigned to a single cluster. We now maximize the similarity between each question and the summary question to which it is assigned. The objective function is: $\varrho(C,S) = \sum_{i=1}^{k} \sum_{\boldsymbol{q}_j \in c_i} |\langle \boldsymbol{q}_j | \boldsymbol{s}_i \rangle|^2$

In bioinformatics this is called the diametric clustering objective function [99], and it has an equivalent metric clustering minimization problem. Using the fact that $|\langle \boldsymbol{q}_j | \boldsymbol{s}_i \rangle|^2 \leq 1$

$$\begin{aligned} \arg\max_{C,S} \varrho(C,S) &= \arg\min_{C,S} \{n - \varrho(C,S)\} \\ &= \arg\min_{C,S} \sum_{i=1}^{k} \sum_{\boldsymbol{q}_j \in c_i} d(q_j, s_i)^2 \end{aligned}$$

where $d(\boldsymbol{v},\boldsymbol{w}) = \sqrt{1 - |\langle \boldsymbol{v}|\boldsymbol{w} \rangle|^2}$. $d(\cdot,\cdot)$ is a pseudometric, which is to say (i) $d(\boldsymbol{v},\boldsymbol{v}) = 0$; (ii) $d(\boldsymbol{v},\boldsymbol{w}) = d(\boldsymbol{w},\boldsymbol{v})$; (iii) $d(\boldsymbol{u},\boldsymbol{v}) + d(\boldsymbol{v},\boldsymbol{w}) \geq d(\boldsymbol{u},\boldsymbol{w})$.

The maximization version of this problem suggests one heuristic, while the minimization problem suggests another. The first is a modification of Lloyd's algorithm.

procedure MODIFIEDLLOYD($\{q_1, \ldots, q_n\}$)
  cluster = initialclustering()
  while $\varrho_{old} \neq \varrho_{new}$ **do**
    $\varrho_{old} = \varrho_{new}$
    for $i = 1$ to k **do**
      $V = concat(q \in c_i)$ ▷ $V = [q_{c1}|\cdots|q_{cm}]$
      $v_i = SVD(V)$ ▷ v_i is largest left sing. vect.

end for

end while

for $j = 1$ to n **do**

put q_j in the cluster that maximizes $|\langle v_i | q_j \rangle|$

end for

recompute ϱ_{new}

end procedure

This algorithm increases the objective function ϱ at each stage. In fact, each for-loop increases ϱ.

Proof. The second for-loop is straightforward, as each question is assigned to the cluster that maximizes the objective. Therefore if any questions change cluster, the objective function will increase.

Let V be defined as above. Then $V = U * D * W^T$ where U and W are unitary and D is a diagonal matrix of singular vectors. Then

$$\sum_{q_j \in c_i} |\langle \boldsymbol{q_j} | \boldsymbol{s} \rangle|^2 = \|\boldsymbol{s}^T V\|^2$$

$$= \sum_i D_{ii}^2 \langle \boldsymbol{u_i} | \boldsymbol{s} \rangle$$

where $\boldsymbol{u_i}$ are the columns of U. This is maximized by setting $\boldsymbol{s}$ to be equal to the largest singular vector $\boldsymbol{u}_1$

Therefore each stage of the algorithm increases ϱ. Since there are a finite number of clusterings, and hence values for ϱ and each stage of the algorithm increases ϱ, it converges, though possibly not to the global optimum. □

Although Lloyd's algorithm guarantees a local maximum in the objective function, for our application we seek a related – but in a local sense, slightly different – condition: we guarantee that the absolute correlation distance cannot exceed a threshold.

We start with n individiual singleton clusters of observations X and a data matrix A of m countries (rows) and n resolutions (columns) (e.g. Table 1.1). We also have a correlation threshold $\theta \in (0, 1)$ and a cooldown ratio $\alpha \in (0, 1)$. We repeatedly iterate through the following steps, merging clusters until only one remains:

```
procedure GREEDYCLUSTER(A, θ)
    unallocated = A
    for c in unallocated do
        remove c from unallocated
        for q in unallocated do
            if abs(corr(c, q) < θ) then
                remove q from unallocated
                assign q to cluster c
            end if
        end for
    end for
    reassign questions to most correlated cluster center
    return clusters
end procedure
procedure GREEDYTREE(A, θ, α)
    tree = ∅
    while numclusters > 1 do
        clusters = GreedyCluster(A, θ)
        add clusters to a new layer in tree
        set A to projection onto the largest singular vector of each cluster
        θ = θα
    end while
    return tree
end procedure
```

Performance is very similiar to the Lloyd algorithm, which could in effect be inserted into the first procedure. The result of GreedyTree is a hierarchy of clusters,

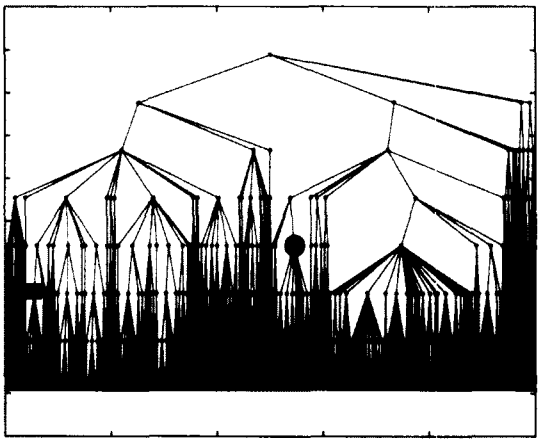


Figure 5.8: *Clustering result of UN resolutions during the period 1998-2002. Two individual clusters are marked with* • *and* ■ *symbols for demonstration.*

whose bottom layer contains singletons and top layer contains a single cluster which includes all features. Each of the layers in between are formed by merging clusters in the layer immediately under it. One possible way toward selecting features relevant to the labeling function f is to include f, as if it is also a feature, into the input dataset before calling GreedyTree. We can examine the resulting hierarchy to see which cluster of features f ends up in.

5.3.2 Experimental Results

UN Resolutions We applied the clustering algorithm on the set of UN resolutions during the period 1998-2002 [9], with $\theta = 0.95$ and $\alpha = 0.8$. Fig. 5.8 shows the clustering hierarchy with two clusters • and ■. Below are the tag descriptions of the resolutions in cluster ■ which contain Middle East-related resolutions:

- *Palestine, Assistance*
- *Palestine, Displaced Persons*
- *Palestine, Scholarships, Grants*
- *UNRWA*

- *Palestine, Refugee Properties*
- *Jerusalem, University*
- *Israel, Geneva Convention*
- *Israel, Settlements*
- *Human Rights, Occupied Territories*
- *Syrian Golan*

and cluster • which is composed of resolutions mostly from Human Rights topic:

- *Human Rights, Iraq*
- *Human Rights, Iran*
- *Human Rights, Iran*
- *Human Rights, Iraq*
- *Human Rights, Democratic Congo*
- *Human Rights, Sudan*
- *Crimes Against Women*
- *Human Rights, Iran*
- *Human Rights, Iraq*
- *Human Rights, Sudan*
- *Human Rights, Democratic Congo*
- *Armaments, Transparency*
- *Human Rights, Iran*
- *Human Rights, Democratic Congo*

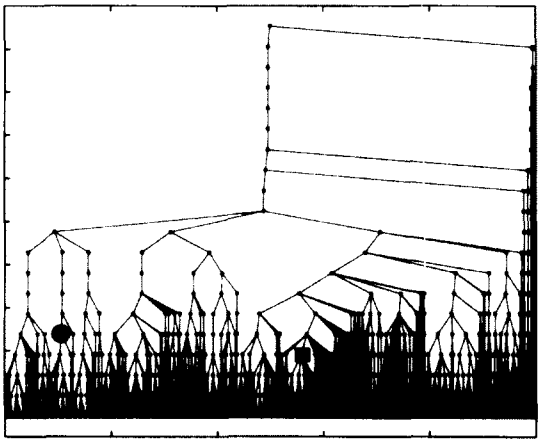


Figure 5.9: *Clustering result of IGO according to memberships in the year 2000. Two individual clusters are marked with • and ■ symbols for demonstration.*

- *Human Rights, Iraq*
- *Human Rights, Sudan*
- *Armaments, Transparency*
- *Executions, Arbitrary*
- *Human Rights, Sudan*
- *Human Rights, Iraq*
- *Human Rights, Democratic Congo*

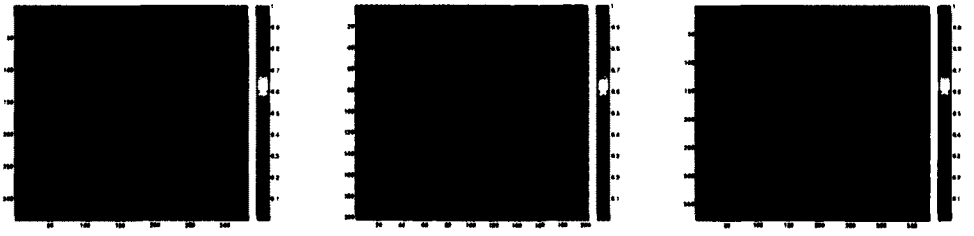
Inter-governmental Organizations (IGO) The clustering hierarchy of IGO's, according to memberships in the year 2000, is displayed in Fig. 5.9. Cluster ■ contain organizations in the European region:

- *CERN: Eur. Org. for Nuclear Research*
- *COE: Council of Europe*
- *EBRD: Eur. Bank for Reconstruction & Dev.*
- *ECPTA: Eur. Conf. of Postal & Telecom. Admin.*

- *EMBC: Eur. Molecular Biology Conf.*
- *EMBL: Eur. Molecular Biology Lab*
- *EMPPO: Eur. & Mediterranean Plant Protection Org.*
- *EPFSC: Eur. Postal Financial Services Commission*
- *EPO: Eur. Patent Office*
- *EUFMD: Eur. Commission for the Control of Food-and-mouth Disease*
- *EUROCONTROL: Eur. Org. for the Safety of Air Navigation*
- *EUROMET: Eur. Collaboration on Measurement Standards*
- *HCPIL: Hague Conf. on Private International Law*
- *NATO: North Atlantic Treaty Org.*
- *OSCE: Org. for Security & Cooperation in Europe*
- *OTIF: Central Office for International Railway Transport*

Cluster • contain organizations in the Caribbean region:

- *ACS: Assoc. of Caribbean States*
- *ASBLAC: Assoc. of Supervisory Banks of Latin America and Caribbean*
- *CARICOM: Caribbean Community*
- *CDB: Caribbean Development Bank*
- *CFATF: Caribbean Financial Action Task Force*
- *CXC: Caribbean Examinations Council*
- *ECCB: Eastern Caribbean Central Bank*
- *IACI: Inter-American Children's Institute*



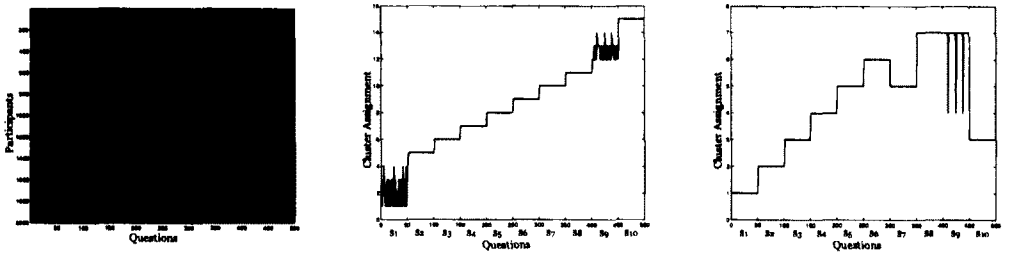
(a) IGO membership (2000) (b) UN votes (2000) (c) AddHealth W1 questionnaire

Figure 5.10: *Absolute value of correlations among the features of different datasets: (a) IGO membership of 333 organizations in year 2000; (b) UN voting records of 202 resolutions in year 2000; (c) AddHealth W1 questionnaire of 374 questions.*

- *IACSS: Inter-American Inst. of Agr. Sci.*
- *OAS: Pan-American Union/Org. of American States*
- *OECS: Organization of Eastern Caribbean States*
- *OPANAL: Agency for Prohibition of Nuclear Arms in Latin America*
- *PAHO: Pan-American Health Organization*

5.3.3 Drawbacks

The correlation hierarchical clustering algorithm provides us with a structured view of the feature space. Taking a horizontal slice at any level reveals clusters of features which are thematically related to each other. The summary variables are useful in averaging out small noises in related features, and also help us achieve dimensionality reduction. However, a few disadvantages must be taken into consideration. First of all, not all datasets have highly-structured feature space like the IGO membership and UN voting records. Even though there are (barely) recognizable structures in the question space of the AddHealth W1 questionnaire dataset (Fig. 5.10c), the correlation values are significantly lower than those presented in Figs. 5.10a-5.10b.



(a) (b) (c)

Figure 5.11: *Experiment on synthetic binary questionnaire with* $N = 2000$, *10 sections of* $m_t = 50$ *questions each,* $\delta = 0.3$*: (a) The data matrix; (b) Cluster assignments of 500 questions at level 6 of the hierarchy; (c) Cluster assignments of 500 questions at level 7 of the hierarchy.*

Additionally, we have no prior statistics to justify the threshold value at any horizontal layer. For example, it may make sense to have a cluster of Human Rights resolutions at layer 5, while resolutions about Arms Control and Border Conflicts separated at level 4. The best clustering is most likely to compose of clusters from different layers. Fig. 5.11 demonstrates such a case using synthetic binary questionnaire of $N = 2000$ persons, 10 sections of $m_t = 50$ questions each, Bernoulli mean parameter $\delta = 0.3$. We ran GreedyTree procedure ($\alpha = 0.5$, initial $\theta = 0.9$) on the dataset shown in Fig. 5.11a. Figs. 5.11b & 5.11c show the cluster assignments of all 500 questions at level 6 and 7, respectively. The clustering hierarchy has 15 clusters at level 6 and 7 clusters at level 7. Fig. 5.11b tells us that most of the sections have been correctly grouped at level 6, with the exceptions of S_1 and S_9, each being divided into 3 subclusters. Our best option is to retain the clustering of other sections and merge the subclusters of S_1 and S_9 separately. However, that is impossible without prior knowledge of proper threshold values, as shown in Fig. 5.11c: at level 7, S_1 is correctly clustered, while $\{S_5, S_7\}$, $\{S_8, S_9\}$, and $\{S_3, S_{10}\}$ are merged together.

Most importantly, using this algorithm can only associate f with feature clusters of only one theme while n practice, the labeling function f may not be constrained within a single theme. With the UN voting records, our most common task is to

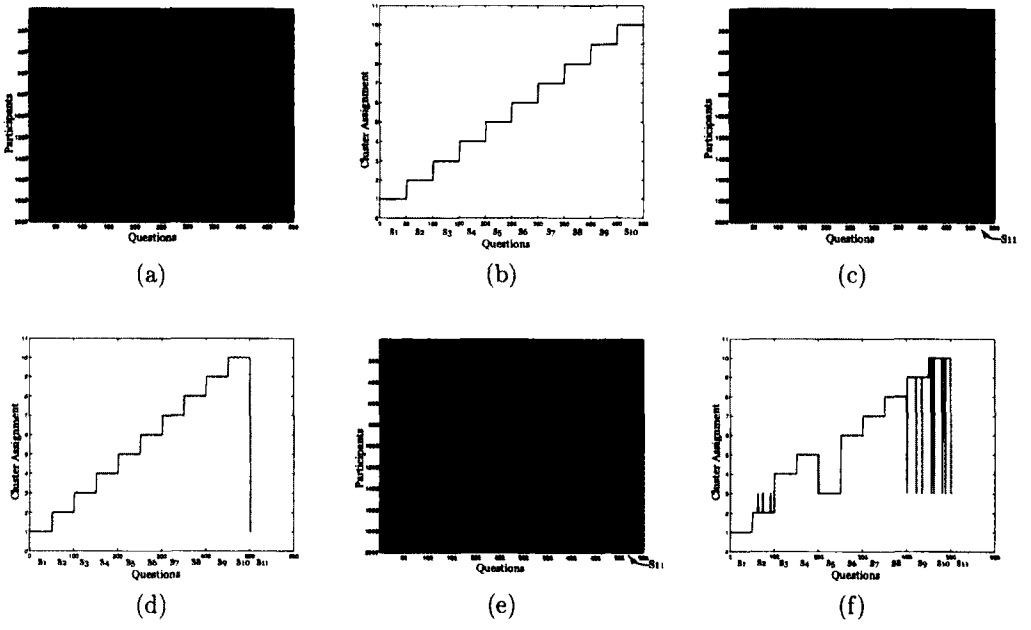


Figure 5.12: *Experiment on synthetic binary questionnaire with $N = 2000$, 10 sections of $m_t = 50$ questions each, $\delta = 0.25$: (a) The original data matrix; (b) Cluster assignments of 500 questions at level 6 of the hierarchy showing perfect clustering; (c) The extended data matrix with $f_1 = \chi_1$ represented as the "11^{th} section" of questions 501 to 550; (d) Cluster assignments of 550 questions in (c) showing perfect clustering and $f_1 = \chi_1$ assigned to the same cluster with questions of S_1 (the red-colored segment); (e) The extended data matrix with $f_2(x_i) = 1$ if $(\sum_{j \in J} A_{ij}) > 10$, $f_2(x_i) = 0$ otherwise. J is a random set of 20 questions which include 3 questions from S_2, 2 from S_3, 2 from S_4, 4 from S_5, 1 from S_7, 1 from S_8, 3 from S_9 and 4 from S_{10}. f_2 is represented as the "11^{th} section" of questions 501 to 550; (f) Cluster assignments of 550 questions in (d) showing minor clustering noise and f_2 assigned to the same cluster with questions of S_5 (the red-colored segment).*

infer the votes in one future resolution based on existing resolutions. The unknown resolution likely belongs to only one of a few strongly-defined topics of interest in the UN General Assembly. However, in other datasets such as the MMPI, AddHealth W1 questionnaires, the function f may not exclusively belong to any semantic group of questions. Fig. 5.12 demonstrates such a case using synthetic binary questionnaire of $N = 2000$ persons, 10 sections of $m_t = 50$ questions each, Bernoulli mean parameter $\delta = 0.25$. Running GreedyTree procedures ($\alpha = 0.5$, initial $\theta = 0.9$) on the dataset shown in Fig. 5.12a yields a perfect 10-cluster clustering at level 6 of the hierarchy (Fig. 5.12b). Fig. 5.12c shows the same questionnaire appended by $f_1 = \chi_1$ which is represented as the "11^{th} section" of questions 501 to 550. Running GreedyTree on this dataset also yields perfect 10-cluster result at level 6 with the "$S_1 1$" (the red segment) correctly assigned to the same cluster as S_1. We then chose a different labeling function $f_2(x_i) = 1$ if $(\sum_{j \in J} A_{ij}) > 10$, $f_2(x_i) = 0$ otherwise, whereas J is a random set of 20 questions which include 3 questions from S_2, 2 from S_3, 2 from S_4, 4 from S_5, 1 from S_7, 1 from S_8, 3 from S_9 and 4 from S_{10}. Fig. 5.12e shows the questionnaire appended by f_2 which is represented as "S_{11}" of questions 501 to 550. Running GreedyTree on this dataset produces an almost perfect clustering with minor noise. However, f_2 is assigned to the same cluster as S_5, implying that the labeling function f_2 belongs to the same topic as only S_5, which is incorrect considering the way f_2 is constructed.

5.4 Second-order correlation of features

Our main interest is in selecting features which *"organize the data points in the same way as f does."* If f is binary, the desired features should also be able to divide the population into two camps, similar to the division of f. If f is continuous, the selected features should spread the population accordingly from f's high end to its

low end. In other words, the resulting structure induced by the chosen features have to *correlate* with the structure implied by f. There fore, if we have a kernel function K_f induced by f, and a kernel function K_y induced by a particular feature y, we can compute the correlation between K_f and K_y as

$$Corr(K_f, K_y) = \frac{< K_f, K_y >_F}{||K_f||_F ||K_y||_F} \tag{5.4.1}$$

where $||K||_F^2 =< K, K >_F$. A simple choice for the kernels of f and y is their autocorrelations:

$$K_f = (f - \mu_f)(f - \mu_f)^T K_y = (y - \mu_y)(y - \mu_y)^T \tag{5.4.2}$$

which make K_f and K_y first-order correlation measures, and $Corr(K_f, K_y)$ therefore a second-order correlation:

$$Corr(K_f, K_y) = E_{u,u',v,v'}[(u - \mu_f)(u' - \mu_f)(v - \mu_y)(v' - \mu_y)] \tag{5.4.3}$$

whereas $u, u' \in range(f)$ and $v, v' \in range(y)$. This choice is similar to the usage of the cross-covariance operator in [100, 101] which measures the dependence between reproducing kernel Hilbert spaces (RKHSs). In our finite, single-feature case, the cross-covariance operator [101] becomes a covariance (correlation) operator:

$$C_{fy} = E_{uv}[(u - \mu_f) * (v - \mu_y)] = E_{uv}[(u - \mu_f)^T (v - \mu_y)] \tag{5.4.4}$$

Then, we then have:

$$C_{fy}^2 =< C_{fy}, C_{fy} >= E_{u,u',v,v'}[(u - \mu_f)(u' - \mu_f)(v - \mu_y)(v' - \mu_y)] = Corr(K_f, K_y) \tag{5.4.5}$$

5.4.1 Filtering Algorithm

procedure CORRELATIONSCORE(A, f, y)

$\tilde{f} = f - \mu_f$

$r = A(y)$ ▷ r is responses to y

$\tilde{r} = r - \mu_r$

$K_f = \tilde{f} * \tilde{f}^T$

$K_y = \tilde{r} * \tilde{r}^T$

$s = tr(K_f * K_y)/\sqrt{tr(K_f * K_f) * tr(K_y * K_y)}$

return s

end procedure

procedure SELECTFEATURES(A, f, k)

for $y \in Y$ **do**

$S(j)$ = CorrelationScore(A, f, y)

end for

set $J = \{j_1, \cdots, j_k\}$ the indices of the top-k highest in S

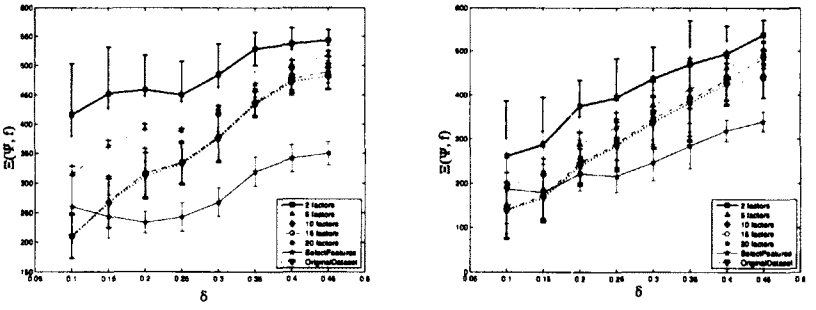
return $\{y_{j1}, \cdots, y_{jk}\}$

end procedure

We use the CorrelationScore procedure to compute $Corr(K_f, K_y)$ for each feature y in the dataset. The procedure SelectFeatures then picks the k top-correlated features. For convenience, we usually use $k = 10$ or $k = 20$. However, a more careful approach is to gradually increase the value of k, starting from 1, and observe the smoothness of f on the induced diffusion embeddings, according to the error function $\Xi(\Psi, f)$ defined in Eq. 5.2.2.

5.4.2 Synthetic Experiments

Fig. 5.13 shows two experiments with random synthetic questionnaire data to compare factor analysis and SelectFeature-enhanced diffusion maps. We repeatedly run experiments on randomized synthetic questionnaire of 2000 participants, Bernoulli mean parameter δ varies from 0.1 to 0.45. In (a), there are exactly 10 sections in



(a) (b)

Figure 5.13: *Experiments on synthetic questionnaire of 2000 participants, δ varies from 0.1 to 0.45. In (a) the questionnaire has exactly 10 sections of 20 binary questions each. In (b) the questionnaire has a random number of sections (from 2 to 10), each has its own random number of questions (from 20-30). In each repetition of the experiment, the function f is created by randomly selecting 20 questions, $f(x_i) = 1$ if the sum of the answers by x_i to these questions exceeds 10, and $f(x_i) = 0$ otherwise. The first 5 lines plot the error functions Ξ of the diffusion maps induced by different number of factors computed by factor analysis. The cyan-colored line plots the error function of the diffusion maps induced by 40 features filtered by the SelectFeatures procedure. To complete the comparison, the last blue dotted line plots the error function of the diffusion maps induced by All questions in the questionnaire.*

the questionnaire, each section consists of exactly 20 questions. Experiment (b) introduces more uncertainty into the process by making the number of sections in each questionnaire a random number in $[2, 10]$; each section has its own random number of questions (in $[20, 30]$). In each repetition, we randomly chose a set J of 20 questions. The labeling function f is defined as:

$$f(x_i) = \begin{cases} 1 & \text{if } \sum_{j \in J} A_{ij} > 10 \\ 0 & \text{otherwise} \end{cases} \tag{5.4.6}$$

We used factor analysis to uncover 2, 5, 10, 15 and 20 factors, projected the dataset onto them to use as low-dimensional coordinates. We used the SelectFeatures procedure to filter 40 questions most correlated with f. We then compute the diffusion maps with the Gaussian kernel on these selected questions. For comparison, we also computed the diffusion maps with the Gaussian kernel on *all quesions*. Several observations can be drawn from Fig. 5.13:

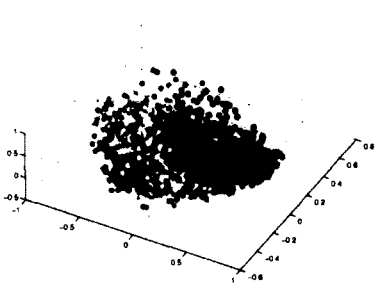
- Increasing values of δ disturb the discovery of factors and degrades the performance of factor analysis-induced maps.
- As the number of factors parameter increases, the performance also rises. The difference in performance is due to the loss of information by analyzing fewer sections than what present in the data.
- Since 10 is the true number of questionnaire sections in experiment (a) (or the maximum number of sections in experiment (b)), the diffusion maps induced by factor analysis of 10 factors performs best with respect to the error function Ξ. Factor analysis using 15 or 20 factors can do not much better than using 10 factors. Generally, the performance of diffusion maps using all questions is on par with that of factor analysis using 10, 15 or 20 factors.
- The diffusion maps computed from features filtered by SelectFeatures has the

best performance, with respect to Ξ. At low values of δ, it behaves slightly poorly, compared to factor analysis, only if the latter is supplied with precise apriori information. However, as *delta* increases, it picks up immediately and performs significantly better than factor analysis.

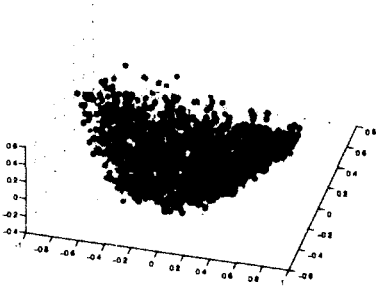
5.4.3 MMPI dataset

We applied SelectFeatures procedure to the MMPI dataset [4]. Fig. 5.14 shows the performance of SelectFeatures against two labeling functions: Depression score and Antisocial behavior score. With the depression score function, using the questions filtered by SelectFeatures appears to be no worse than using all questions (Figs. 5.14a & 5.14e). Moreover, SelectFeatures yielded a perfectly smooth antisocial behavior score function in Fig. 5.14f, in comparison with Fig. 5.14b. In descending order of their corresponding correlation scores, the 30 questions most correlated with depression scale are:

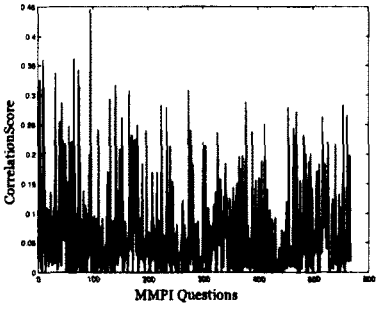
- 95. I am happy most of the time.
- 65. Most of the time I feel blue.
- 9. My daily life is full of things that keep me interested.
- 73. I am certainly lacking in self-confidence.
- 31. I find it hard to keep my mind on a task or job.
- 3. I wake up fresh and rested most mornings.
- 140. Most nights I go to sleep without thoughts or ideas bothering me.
- 75. I usually feel that life is worthwhile.
- 165. My memory seems to be all right.



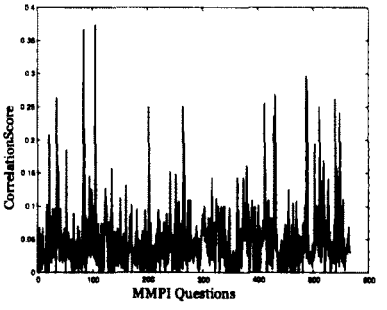
(a) MMPI Scale 2 (D/Depression), all questions used



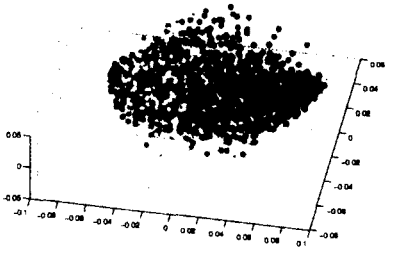
(b) MMPI Scale RC4 (Antisocial Behavior), all questions used



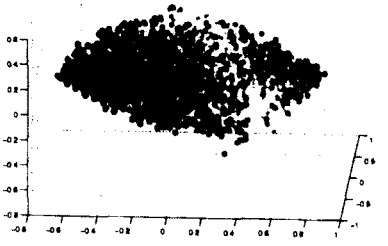
(c) MMPI Scale 2 (D/Depression), correlation scores



(d) MMPI Scale RC4 (Antisocial Behavior) correlation scores



(e) MMPI Scale 2 (D/Depression), top-30 correlated questions used



(f) MMPI Scale RC4 (Antisocial Behavior), top-30 correlated questions used

Figure 5.14: *Diffusion maps of the MMPI dataset [3, 4] with Gaussian kernels, using the first 3 eigenfunctions. The left column uses Scale 2 (D/Depression) score and the right column uses Scale RC4 (antisocial behavior) score as labeling functions: (a) & (b) diffusion maps computed from all questions (replicated from Fig. 5.1); (c) & (d) plots of correlation scores of the question responses against the two labeling functions; (e) & (f) diffusion maps calculated from the top-30 most correlated questions.*

- 273. Lile is a strain for me much of the time.
- 10. I am about as able to work as I ever was.
- 130. I certainly feel useless at times.
- 377. I am not happy with myself the way I am.
- 43. My judgment is better than it ever was.
- 223. I believe I am no more nervous than most others.
- 554. When my life gets difficult, it makes me want to just give up.
- 454. The future seems hopeless to me.
- 233. I have difficulty in starting to do things.
- 469. I sometimes feel that I am about to go to pieces.
- 141. During the past few years I have been well most ofthe time.
- 561. I usually have enough energy to do my work.
- 516. My life is empty and meaningless.
- 152. I do not tire quickly.
- 39. My sleep is fitful and disturbed.
- 411. At times I think I am no good at at all.
- 180. There is something wrong with my mind.
- 56. I wish I could be as happy as others seem to be.
- 464. I feel tired a good deal of the time.
- 109. I seem to be about as capable and smart as most others around me.

- 196. I frequently find myself worrying about something.

while the 30 questions most correlated with RC4 antisocial behavior scale (in descending order of correlation score) are:

- 105. In school I was sometimes sent to the principal for bad behavior.
- 84. I was suspended from school one or more times for bad behavior.
- 487. I have enjoyed using marijuana.
- 489. I have a drug or alcohol problem.
- 431. In school my marks in classroom behavior were quite regularly bad.
- 35. Sometimes when I was young I stole things.
- 540. I have gotten angry and broken furniture or dishes when I was drin...
- 412. When I was young I often did not go to school even when I shouid h...
- 264. I have used alcohol excessively.
- 511. Once a week or more I get high or drunk.
- 202. My parents often objected to the kind of people I went around with.
- 548. Ive been so angry at times that Ive hurt someone in a physic...
- 429. Except by doctors orders I never take drugs or sleeping pills.
- 266. I have never been in trouble with the law.
- 21. At times I have very much wanted to leave home.
- 502. I have some habits that are really harmful.
- 52. I have not lived the right kind of life.

- 518. I have made lots of had mistakes in my life.
- 85. At times I have a strong urge to do something harmful or shocking.
- 379. I got many beatings when I was a child.
- 37. At times I feel like smashing things.
- 542. I have become so angry with someone that I have felt as if I would...
- 134. At times I feel like picking a fist fight with someone.
- 240. At times it has been impossible lor me to keep from stealing or sh...
- 250. At times l have been so entertained by the cleverness of some crim...
- 94. Much of the time I feel as If I have done something wrong or evil.
- 373. I have done some bad things in the past that I never tell anybody ...
- 362. I can remember "playing sick" to get out of something.
- 316. I have strange and peculiar thoughts.
- 527. After a bad day, I usually need a few drinks to relax.

Fig. 5.15 presents a comparison between factor analysis (FA) and diffusion maps with feature selection on the MMPI dataset, using RC4 score for labeling function. The blue line plots the error function $\Xi(FA, RC4)$ of the embedding using the FA factors as coordinates, whereas the number of factors increases from 5 to 50 along the horizontal axis. Its shape reveals the true number of factors is about 25. The green line plots the error function $\Xi(\Psi, RC4)$, where Ψ is the diffusion coordinates computed from all questions. The red line plots $\Xi(\Psi, f)$, in which Ψ is the diffusion coordinates calculated from top-30 questions filtered by SelectFeatures. The plots indicate that diffusion maps enhanced by SelectFeatures significantly reduces the error, compared to that of diffusion maps without SelectFeatures and factor analysis.

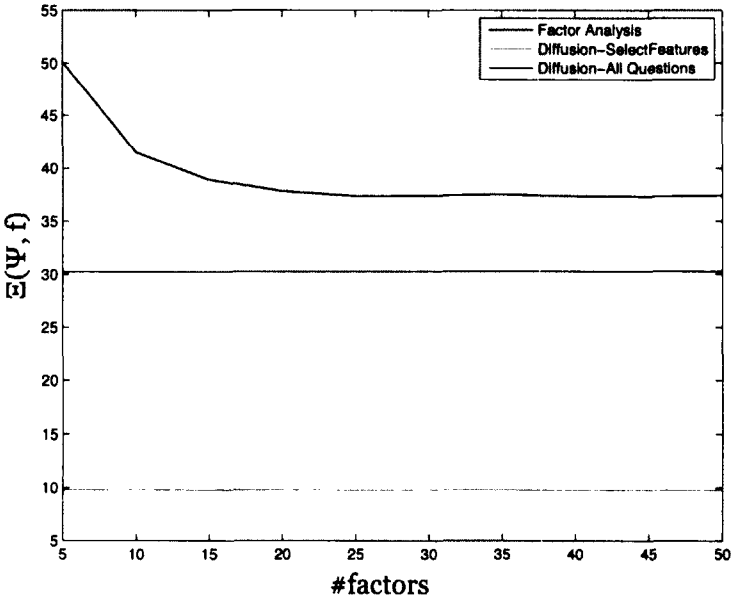
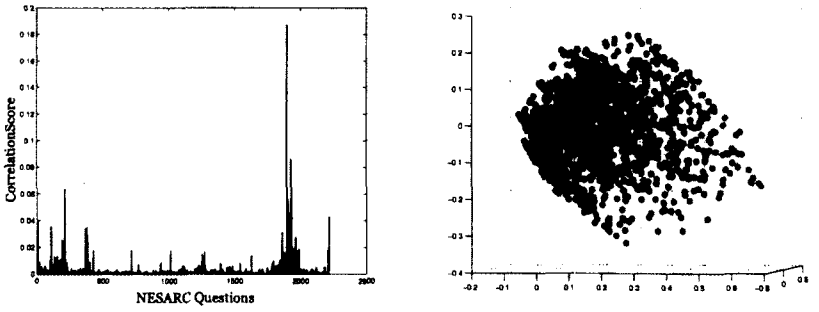


Figure 5.15: *Comparison between factor analysis (FA) and diffusion maps with feature selection on the MMPI dataset, using RC4 score for labeling function. The blue line plots the error function $\Xi(FA, RC4)$ of the embedding using the FA factors as coordinates. The number of factors increases from 5 to 50 along the horizontal axis. The green line plots the error function $\Xi(\Psi, RC4)$, where Ψ is the diffusion coordinates computed from all questions. The red line plots $\Xi(\Psi, f)$, in which Ψ is the diffusion coordinates calculated from top-30 questions filtered by SelectFeatures.*



(a) Scores of correlation with drunk driving behavior

(b) Diffusion maps on questions with correlation scores above 0.02

Figure 5.16: *(a) Plot of correlation scores between the driving-while-drinking question S2BQ1A21 and other non-alcohol questions. The red dotted line shows the applied threshold value 0.02 to obtain the top correlated questions; (b) Diffusion maps of 5000 NESARC participants, whose points are colored by their responses to drunk driving questions S2BQ1A21. The diffusion coordinates were computed from a Gaussian kernel of questions whose correlation score to S2BQ1A21 is above 0.02.*

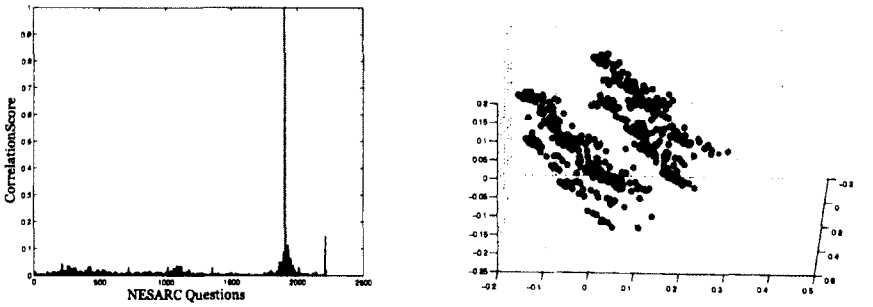
5.4.4 NESARC questionnaire

We applied our algorithm to the NESARC questionnaire data to see how SelectFeatures procedure improve the embedding in Fig. 5.2. Fig. 5.16a shows the resulting plot of correlation scores and the threshold 0.02 being applied for filtering purpose. The questions whose correlation scores are above 0.02 are mostly about (i) smoking and illicit or non-medical prescription drug use, and (ii) antisocial behaviors:

- GENDER
- HISTDX2. HISTRIONIC PERSONALITY DISORDER (LIFETIME DIAGNOSIS)
- PARADX2. PARANOID PERSONALITY DISORDER (LIFETIME DIAGNOSIS)
- S10Q1B63. EVER TROUBLE YOU OR CAUSE PROBLEMS AT WORK/SCHOOL OR WITH FAMILY/OTHER PEOPLE

- S10Q1B64. EVER TROUBLE YOU OR CAUSE PROBLEMS AT WORK/SCHOOL OR WITH FAMILY/OTHER PEOPLE
- S11AQ1A1. OFTEN CUT CLASS, NOT GO TO CLASS OR GO TO SCHOOL AND LEAVE WITHOUT PERMISSION(W1)
- S11AQ1A13. EVER SCAM OR CON SOMEONE FOR MONEY, TO AVOID RESPONSIBILITY OR JUST FOR FUN
- S11AQ1A14. EVER DO THINGS THAT COULD EASILY HAVE HURT YOU OR SOMEONE ELSE, LIKE SPEEDING OR DRIVING AFTER HAVING TOO MUCH TO DRINK(W1)
- S11AQ1A15. EVER GET MORE THAN 3 TICKETS FOR RECKLESS/CARELESS DRIVING, SPEEDING, OR CAUSING AN ACCIDENT(W1)
- S11AQ1A21. EVER FORGE SOMEONE ELSE'S SIGNATURE, LIKE ON A LEGAL DOCUMENT OR CHECK
- S11AQ1A23. EVER ROB OR MUG SOMEONE OR SNATCH A PURSE
- S11AQ1A24. EVER MAKE MONEY ILLEGALLY, LIKE SELLING STOLEN PROPERTY OR SELLING DRUGS(W1)
- S11AQ1B21. HAPPEN BEFORE AGE 15 (FORGE SOMEONE ELSE'S SIGNATURE)
- S11AQ1B24. HAPPEN BEFORE AGE 15 (MAKE MONEY ILLEGALLY, LIKE SELLING STOLEN PROPERTY OR SELLING DRUGS)
- S11AQ1C1. HAPPEN SINCE AGE 13 (CUT CLASS, W1)
- S11AQ1C13 HAPPEN SINCE AGE 15 (SCAM OR CON SOMEONE FOR MONEY)

- S11AQ1C14. HAPPEN SINCE AGE 15 (RECKLESS, W1)
- S11AQ1C15. HAPPEN SINCE AGE 15 (MORE THAN 3 TICKETS, W1)
- S11AQ1C18. HAPPEN SINCE AGE 15 (START FIRE ON PURPOSE TO DESTROY SOMEONE ELSE'S PROPERTY)
- S11AQ1C21. HAPPEN SINCE AGE 15 (FORGE SOMEONE ELSE'S SIGNATURE)
- S11AQ1C23. HAPPEN SINCE AGE 15 (ROB OR MUG SOMEONE OR SNATCH A PURSE)
- S11AQ1C24. HAPPEN SINCE AGE 15 (ILLEGAL INCOME, W1)
- S11AQ5B. DID ALL EXPERIENCES BEFORE AGE 15 HAPPEN WHILE OR AFTER USING A MEDICINE OR DRUG
- S3AQ1A. EVER SMOKED 100+ CIGARETTES (W1)
- S3AQ1B. EVER SMOKED 50+ CIGARS (W1)
- S3AQ1C. EVER SMOKED A PIPE 50+ TIMES
- S3AQ1D. EVER USED SNUFF 20+ TIMES (W1)
- S3AQ1E. EVER USED CHEWING TOBACCO 20+ TIMES (W1)
- S3AQ93. SMOKED PIPE WHEN HAD SOME OF THESE EXPERIENCES WITH TOBACCO IN LAST 12 MONTHS
- S3BQ1A4. EVER USED AMPHETAMINES (W1)
- S3BQ1A5. EVER USED CANNABIS (W1)
- S3BQ1A6. EVER USED COCAINE OR CRACK (W1)



(a) Scores of correlation with Vandalizing behavior

(b) Diffusion maps on questions with correlation scores above 0.05

Figure 5.17: *(a) Plot of correlation scores between the vandalizing behavior question S2BQ1A17 and other non-alcohol questions. The red dotted line shows the applied threshold value 0.05 to obtain the top correlated questions; (b) Diffusion maps of 5000 NESARC participants, whose points are colored by their responses to driving-while-drinking questions S2BQ1A21. The diffusion coordinates were computed from a Gaussian kernel of questions whose correlation score to S2BQ1A17 is above 0.02.*

- S3BQ1A7. EVER USED HALLUCINOGENS (W1)
- S3CD5Q12E7. HAPPEN PRIOR TO LAST 12 MONTHS WITH CANNABIS
- S3CQ12A7. MORE THAN ONCE DRIVE A VEHICLE UNDER THE INFLUENCE OF MEDICINE/DRUG (W1)
- S3CQ12D7. PPY MORE THAN ONCE DRIVE A VEHICLE UNDER THE INFLUENCE OF MEDICINE/DRUG (W1)
- SCHIZDX2. SCHIZOID PERSONALITY DISORDER (LIFETIME DIAGNOSIS)

Fig. 5.16b repeats the experiment in Fig. 5.2, this time using only questions filtered by SelectFeatures. The drunk-drivers are more well organized in this embedding, with a transition from blue color (low values) to red (high values), compared to Fig. 5.2.

As another experiment, we used the vandalizing behavior (part of antisocial behavior section in the questionnaire) S11AQ1A17 (EVER DESTROY/ BREAK/ VANDALIZE SOMEONE ELSE'S PROPERTY) as the labeling function. Fig. 5.17a shows

the resulting plot of correlation scores and the threshold 0.05 being applied for filtering purpose. The filtered questions are all about antisocial behaviors:

- ANTISOCDX2. ANTISOCIAL PERSONALITY DISORDER (WITH CONDUCT DISORDER)
- S11AQ1A11. EVER TIME HAVE WHEN YOU LIED A LOT, OTHER THAN TO AVOID BEING HURT
- S11AQ1A13. EVER SCAM OR CON SOMEONE FOR MONEY, TO AVOID RESPONSIBILITY OR JUST FOR FUN
- S11AQ1A14. EVER DO THINGS THAT COULD EASILY HAVE HURT YOU OR SOMEONE ELSE, LIKE SPEEDING OR DRIVING AFTER HAVING TOO MUCH TO DRINK(W1)
- S11AQ1A18. EVER START FIRE ON PURPOSE TO DESTROY SOMEONE ELSE'S PROPERTY OR JUST TO SEE IT BURN
- S11AQ1A20. EVER STEAL SOMETHING FROM SOMEONE/SOMEPLACE WHEN NO ONE WAS AROUND
- S11AQ1A22. EVER SHOPLIFT
- S11AQ1A24. EVER MAKE MONEY ILLEGALLY, LIKE SELLING STOLEN PROPERTY OR SELLING DRUGS(W1)
- S11AQ1A25. EVER DO SOMETHING YOU COULD HAVE BEEN ARRESTED FOR, REGARDLESS OF WHETHER YOU WERE CAUGHT OR NOT
- S11AQ1A27. EVER GET INTO A LOT OF FIGHTS THAT YOU STARTED

- S11AQ1A30. EVER HIT SOMEONE SO HARD THAT YOU INJURED THEM OR THEY HAD TO SEE A DOCTOR
- S11AQ1A31. EVER HARASS, THREATEN OR BLACKMAIL SOMEONE
- S11AQ1B17. HAPPEN BEFORE AGE 15 (VANDALIZE, W1)
- S11AQ1B18. HAPPEN BEFORE AGE 15 (ARSON, W1)
- S11AQ1B20. HAPPEN BEFORE AGE 15 (STEAL, W1)
- S11AQ1B22. HAPPEN BEFORE AGE 15 (SHOPLIFT, W1)
- S11AQ1B25. HAPPEN BEFORE AGE 15 (COULD BE ARRESTED, W1)
- S11AQ1C13. HAPPEN SINCE AGE 15 (SCAM OR CON SOMEONE FOR MONEY)
- S11AQ1C14. HAPPEN SINCE AGE 15 (RECKLESS, W1)
- S11AQ1C17. HAPPEN SINCE AGE 15 (VANDALIZE, W1)
- S11AQ1C20. HAPPEN SINCE AGE 15 (STEAL, W1)
- S11AQ1C22. HAPPEN SINCE AGE 15 (SHOPLIFT, W1)
- S11AQ1C24. HAPPEN SINCE AGE 15 (ILLEGAL INCOME, W1)
- S11AQ1C25. HAPPEN SINCE AGE 15 (COULD BE ARRESTED, W1)
- S11AQ1C30. HAPPEN SINCE AGE 15 (INJURE OTHERS, W1)
- S11AQ1C31. HAPPEN SINCE AGE 15 (HARASS, THREATEN, W1)

Fig. 5.17b, maps the 5000 questionnaire participants using only the filtered questions listed above. Compared to Fig. 5.2, the drunk-drivers are more well organized in this embedding. We could see two groups of points, one including mostly blue

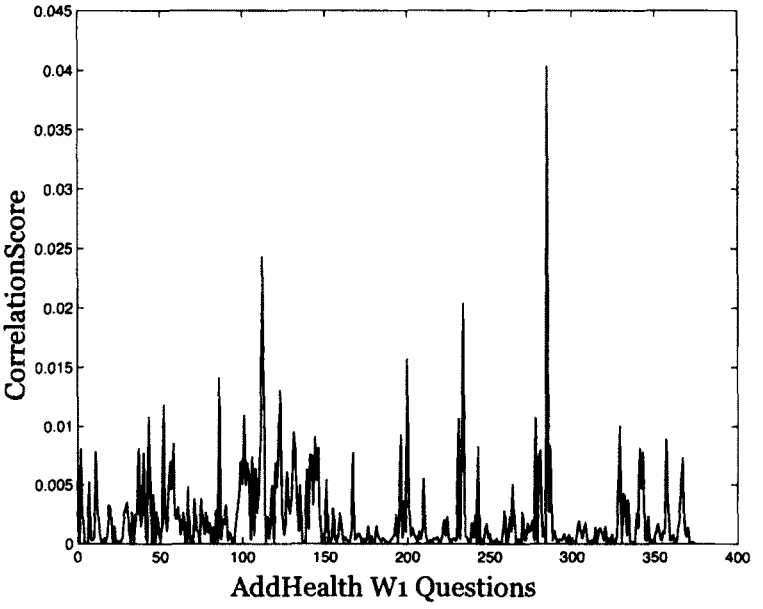


Figure 5.18: *Correlation scores between the W-2 binging kernel (computed from W-2 question about binging frequency) and the kernels computed from non-alcohol questions in the AddHealth dataset. The red dotted line shows the applied threshold value 0.01 to obtain the 14 questions, which are used to render the diffusion maps in Fig. 5.19.*

points, and the other containing mostly red points. This shows a strong relationship between drunk driving (S2BQ1A21) and antisocial behaviors.

5.4.5 AddHealth questionnaire

We applied SelectFeatures procedure to the AddHealth W1 questionnaire. The correlation scores in Fig. 5.18 are extremely low, indicating that most of the questions do not carry a lot of information about the bingers. So, we focused on the most informative responses. Taking all the questions whose correlation scores are higher than 0.01, we obtained the following 13 questions

- S28Q1 EVER SMOKED A CIGARETTE-W1
- S29Q14 TAKE PART IN A GROUP FIGHT-W1

- C46 THINK HAS KISSED/NECKED-PQ
- S22Q1 DID YOU HAVE A ROMANCE-W1
- S18Q16 RELY ON GUT FEELINGS-W1
- S33Q4 PAST YR-FRIENDS ATTEMPT SUICIDE-W1
- S29Q15 LOUD/ROWDY IN A PUBLIC PLACE-W1
- S3Q51 TYPICAL HOURS OF SLEEP-W1
- S29Q3 LIE TO PARENTS ABOUT WHEREABOUT-W1
- S3Q40 FREQ-RIDE A MOTORCYCLE-W1
- S26Q5 NON-ROMANCE SEX W/ ANYONE-W1
- C32 BAD TEMPER-PQ
- S1Q6B RACE-AFRICAN AMERICAN-W1

Fig. 5.19 shows the diffusion embedding of all participants, with a Gaussian kernel on the 14 questions, using the first 3 eigenfunctions, at several different rotations.

From Fig. 5.19a, we see that the participants belong to 2 large clusters located along the second eigenfunction.

From Fig. 5.20, most of African American are located on the upper portion of the second eigenfunction. Since there are only 16 bingers (out of 191) who are African American, we turned our focus to the non-African American group. We filter out all participants whose second eigenfunction value is higher than 0.03, and repeat the kernel correlation computation, based on the remaining group of participants. The kernel correlation scores are shown in Fig. 5.21.

Thresholding the correlation scores in Fig. 5.21 at 0.015, we obtained the following 10 questions:

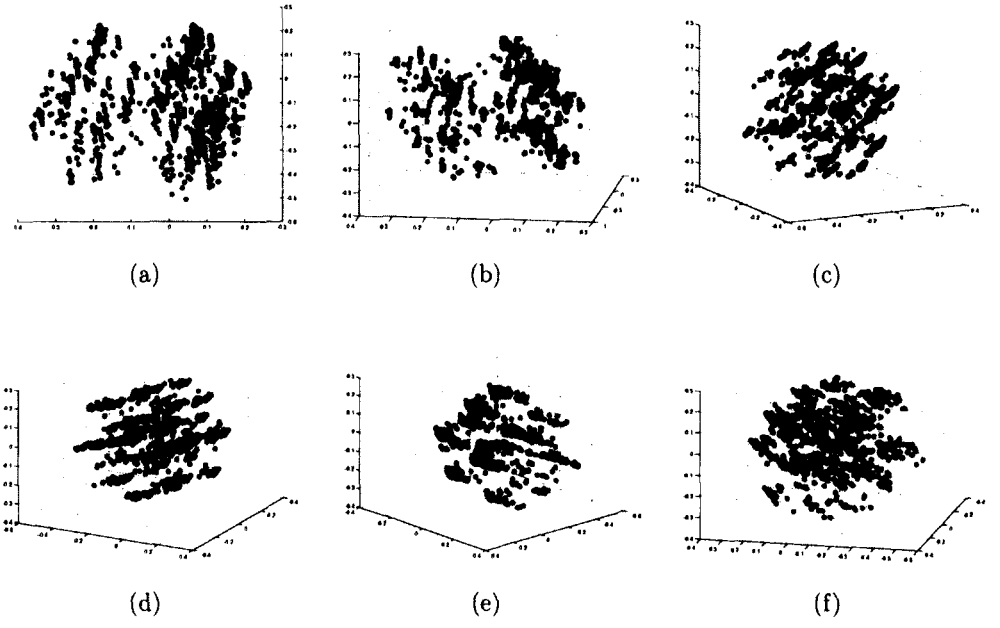


Figure 5.19: *Diffusion maps of AddHealth W1 questionnaire participants using 13 questions whose correlation scores to the binging kernel (Fig. 5.18) is more than 0.01, and a Gaussian kernel. Bingers are colored in red. The subfigures are of the same 3-dimensional map from different viewing angles.*

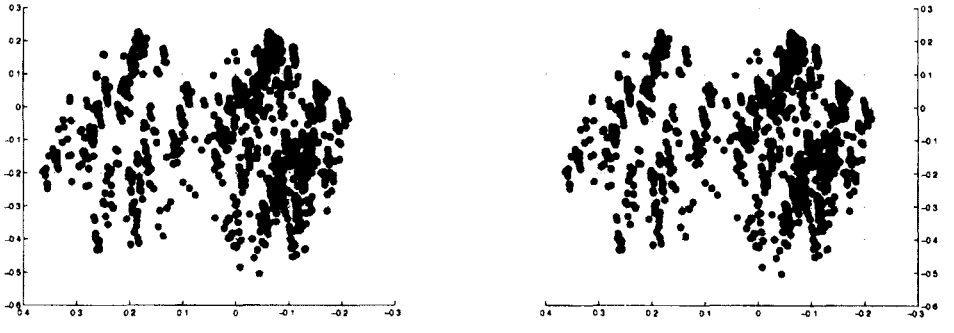


Figure 5.20: *Thresholding on the second eigenfunction of the embedding shown in Fig. 5.19a: (a) Participants whose second eigenfunction has value greater than 0.05 are colored in red; (b) Participants whose answered Yes to "S1Q6B RACE-AFRICAN AMERICAN-W1" are colored in red.*

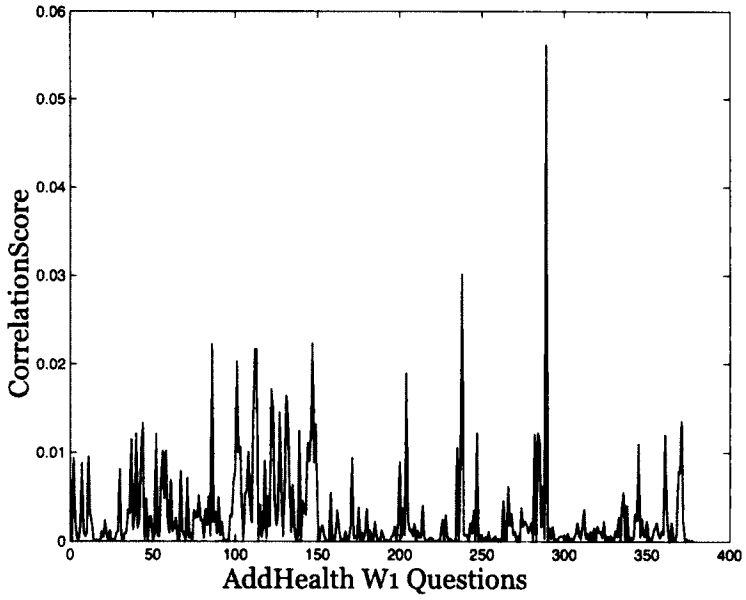


Figure 5.21: *Correlation scores between the W-2 binging kernel and the kernels computed from W1 non-alcohol questions in the AddHealth dataset, based on non-African American respondents. The red dotted line shows the applied threshold value 0.015 to obtain the 14 questions, which are used to render the diffusion maps in Fig. 5.22.*

- S28Q1 EVER SMOKED A CIGARETTE-W1
- C46 THINK HAS KISSED/NECKED-PQ
- S18Q16 RELY ON GUT FEELINGS-W1
- S29Q14 TAKE PART IN A GROUP FIGHT-W1
- S29Q15 LOUD/ROWDY IN A PUBLIC PLACE-W1
- S29Q3 LIE TO PARENTS ABOUT WHEREABOUT-W1
- S22Q1 DID YOU HAVE A ROMANCE-W1
- S35Q6 WANT TO LEAVE HOME-W1
- S33Q1 PAST YR-THINK ABOUT SUICIDE-W1
- S26Q5 NON-ROMANCE SEX W/ ANYONE-W1

More than half of these questions are about behavioral undercontrol and rule breaking, especially when considered in the context that these are 11-12 year olds [102]. More interestingly, this filtered list of social and behavioral characteristics strongly agrees with the recent findings by the IMAGEN project [103] of factors differentiating bingers from non-bingers (their Extended data Fig 7), in order of strength of relationship: smoking, romantic events, deviance events/deviance valence (suicidal ideation), and family events. These same domains of content differentiate those who became bingers in the prospective analysis from those who did not. Brain activation patterns in reward response and emotional reactivity and several candidate genes relating to alcohol abuse also predict these drinking differences, albeit with much lower strength [103]. This strong agreement is a signal that our findings are on the right track, reaching the same results via only responses in a limited dataset.

Working only with the W1 questions in the filtered list above, Fig. 5.22 shows the diffusion embedding of non-African American participants, with a Gaussian kernel of these 10 questions, using the first 3 eigenfunctions.

According to Fig. 5.22, the first 3 eigenfunctions divide the non-African American participants into 12 clusters. The first 2 eigenfunctions organize these clusters in parallel lines (Fig. 5.22a). Fig. 5.24 shows that the smokers and non-smokers are separated along the parallel direction, while Fig. 5.25 shows that the direction perpendicular to the parallel direction resembles the sum of answers to questions "S22Q1 DID YOU HAVE A ROMANCE-W1" and "C46 THINK HAS KISSED/NECKED-PQ". Moreover, the third eigenfunction organizes the clusters into layers, which correspond to the difference between answers to "S22Q1 DID YOU HAVE A ROMANCE-W1" and "C46 THINK HAS KISSED/NECKED-PQ" (Fig. 5.26). Thresholding the third eigenfunction, we found 5 layers:

- $V3 < -0.15$: kids said no romance while parents thought their kids have kissed/necked.
- $-0.15 \leq V3 < -0.07$: kids said no romance and parents had no idea.
- $-0.07 \leq V3 < 0$: both answered yes.
- $0 \leq V3 < 0.1$: mixed set of both answered yes or both answered no.
- $0.1 \leq V3$: kids had romance and parents thought no kissing/necking took place.

Since the bingers are the main target, we observe that the bingers are located with high density in certain portions of the parallel clusters. Fig. 5.27 indicates that in each cluster, the 4th eigenfunction spreads along the parallel direction, while the 6th eigenfunction spreads along the vertical direction. Therefore, as we are able to locate each individual cluster using the first 3 eigenfunctions, we can locate the area of high binger density in that cluster with the 4th and 6th eigenfunctions. Fig. 5.28

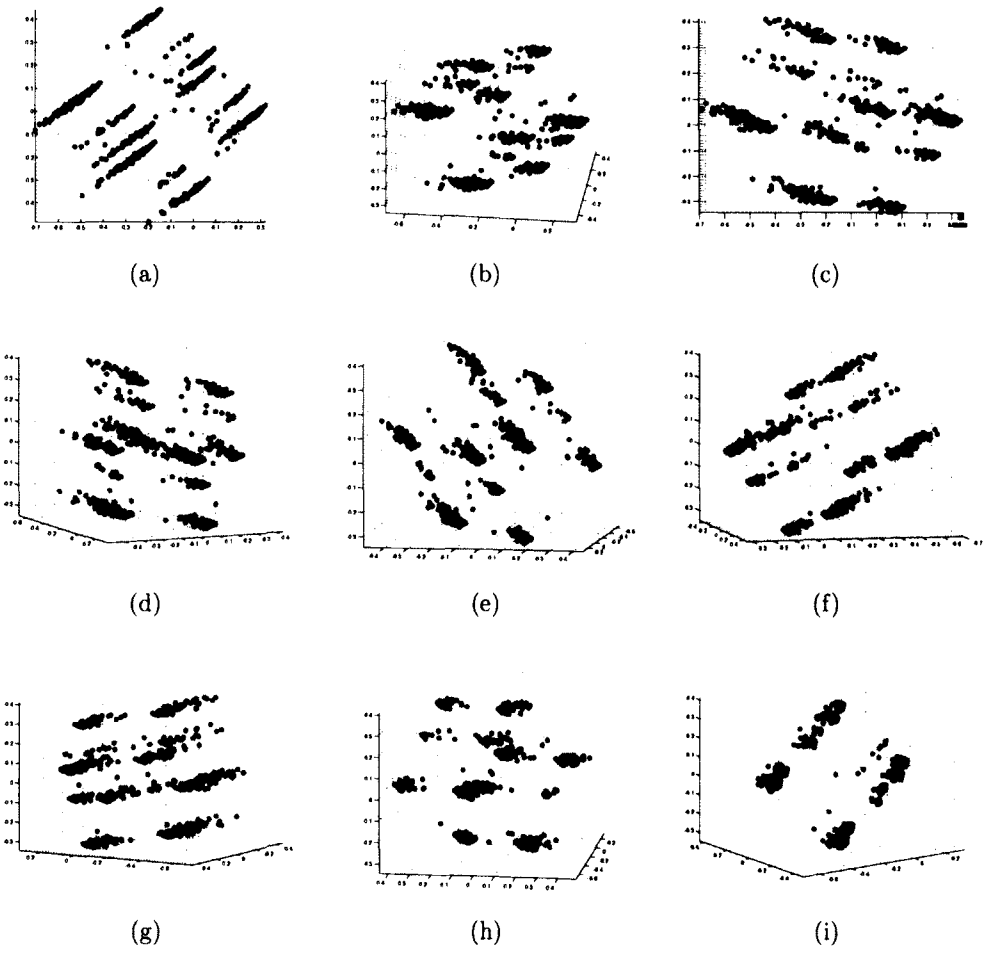


Figure 5.22: *Diffusion maps of AddHealth W1 questionnaire's non-African American participants, using 10 questions whose correlation scores to the binging kernel is more than 0.015, and a Gaussian kernel. Bingers are colored in red. The subfigures are of the same 3-dimensional map, from different viewing angles.*

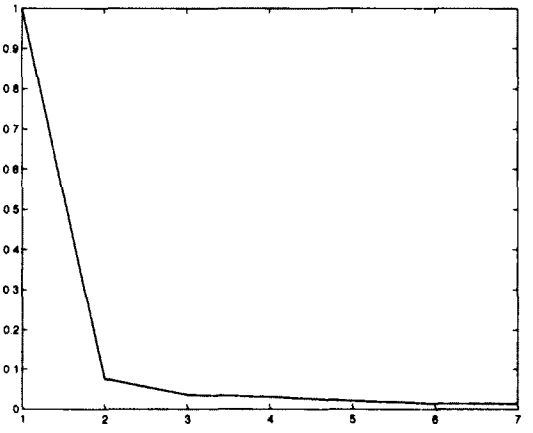
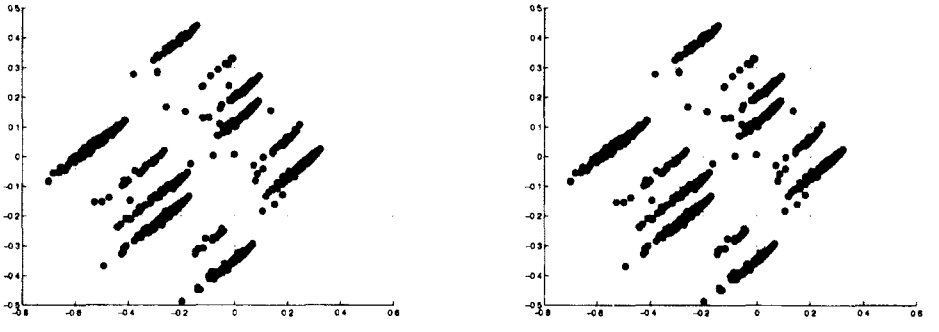
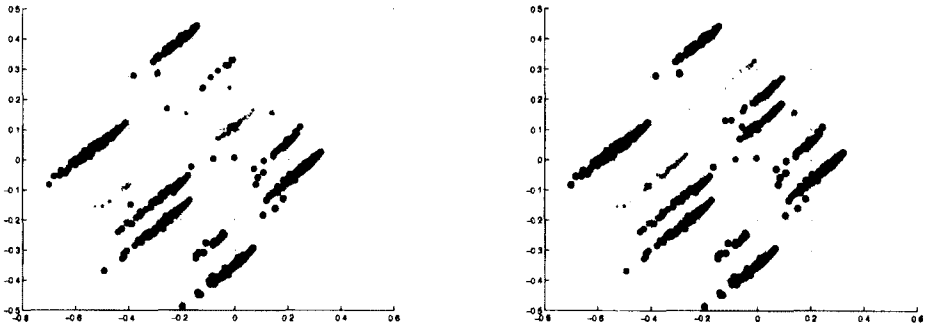


Figure 5.23: *The spectrum of the Gaussian kernel, built on the responses of non-African American to 10 questions whose correlation scores to the binging kernel is more than 0.015. The graph is limited to the first 7 eigenvalues.*



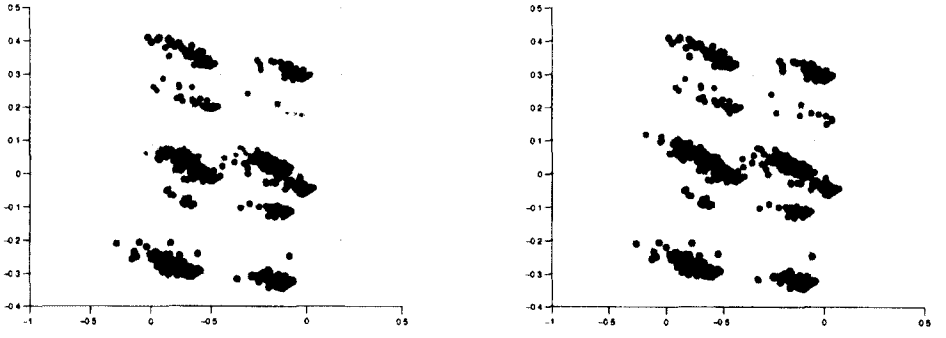
(a) (b)

Figure 5.24: *Thresholding on the first two eigenfunctions of the embedding shown in Fig. 5.22a: (a) Participants whose projections of the first two eigenfunctions onto the diagonal (2, 1) direction (SW-NE) value lower than 1.6 are colored in red; (b) Participants who answered Yes to "S28Q1 EVER SMOKED A CIGARETTE-W1" are colored in red.*



(a) (b)

Figure 5.25: *Thresholding on the first two eigenfunctions of the embedding shown in Fig. 5.22a: (a) Participants are colored based on the projection values of the first two eigenfunctions onto the diagonal (-1, 2) direction (NW-SE); (b) Participants are colored based on the sum of answers to "S22Q1 DID YOU HAVE A ROMANCE-W1" and "C46 THINK HAS KISSED/NECKED-PQ".*



(a) (b)

Figure 5.26: *Thresholding on the third eigenfunction of the embedding shown in Fig. 5.22c (This is the vertical direction of the 3-dimensional map in Fig. 5.22): (a) Participants are colored based on third eigenfunction; (b) Participants are colored based on the difference of answers to "S22Q1 DID YOU HAVE A ROMANCE-W1" and "C46 THINK HAS KISSED/NECKED-PQ".*

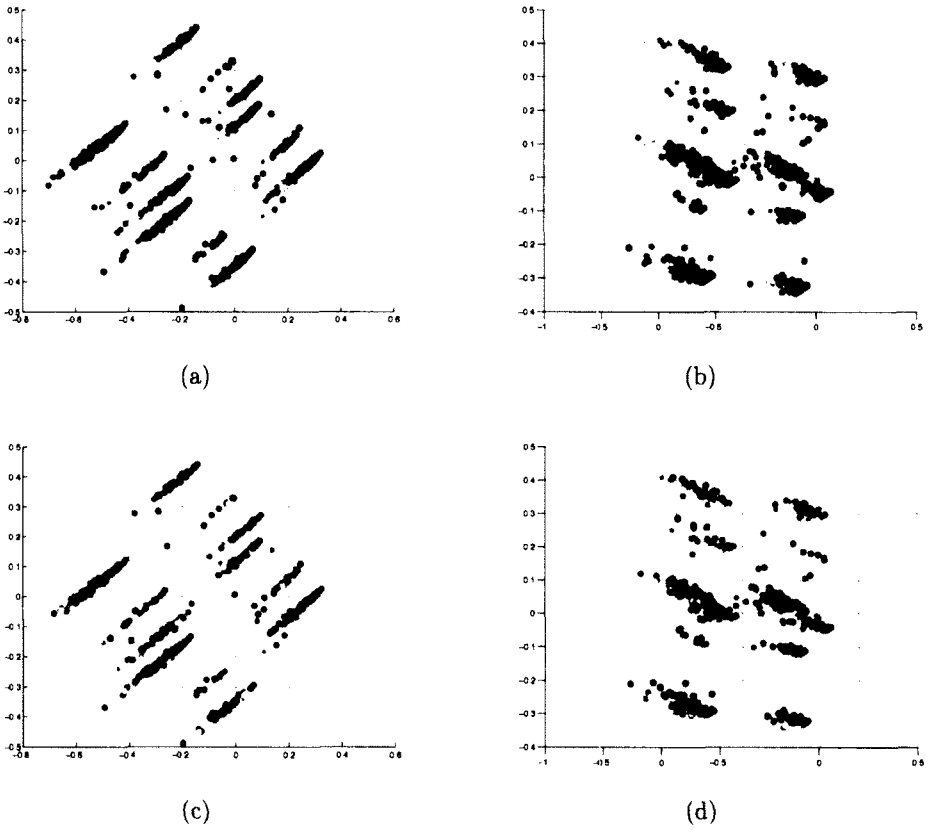


Figure 5.27: *Thresholding of local clusters using 4th and 6th eigenfunctions: (a) & (c) are taken from the same angle as Fig. 5.22a while (b) & (d) are captured from the same perspective as Fig. 5.22c; (a) & (b) Participants are colored based on 4th eigenfunction; (c) & (d) Participants are colored based on 6th eigenfunction. These eigenfunctions are explained in Fig. 5.29.*

provides a demonstration of using the first 3 eigenfunctions to locate one particular cluster, then thresholding the 4th eigenfunction to locate the bingers.

Investigating the 4th and 6th eigenfunctions, Fig. 5.29a indicates that these two functions separate people with suicidal thoughts from those who don't have such thoughts. The addition of the 5th eigenfunction in Fig. 5.29b widens the separation between the two groups. Fig. 5.29c shows certain relation between questions "S33Q1 PAST YR-THINK ABOUT SUICIDE-W1" and "S18Q16 RELY ON GUT FEELINGS-W1": Among the participants who have the same answers to "S18Q16

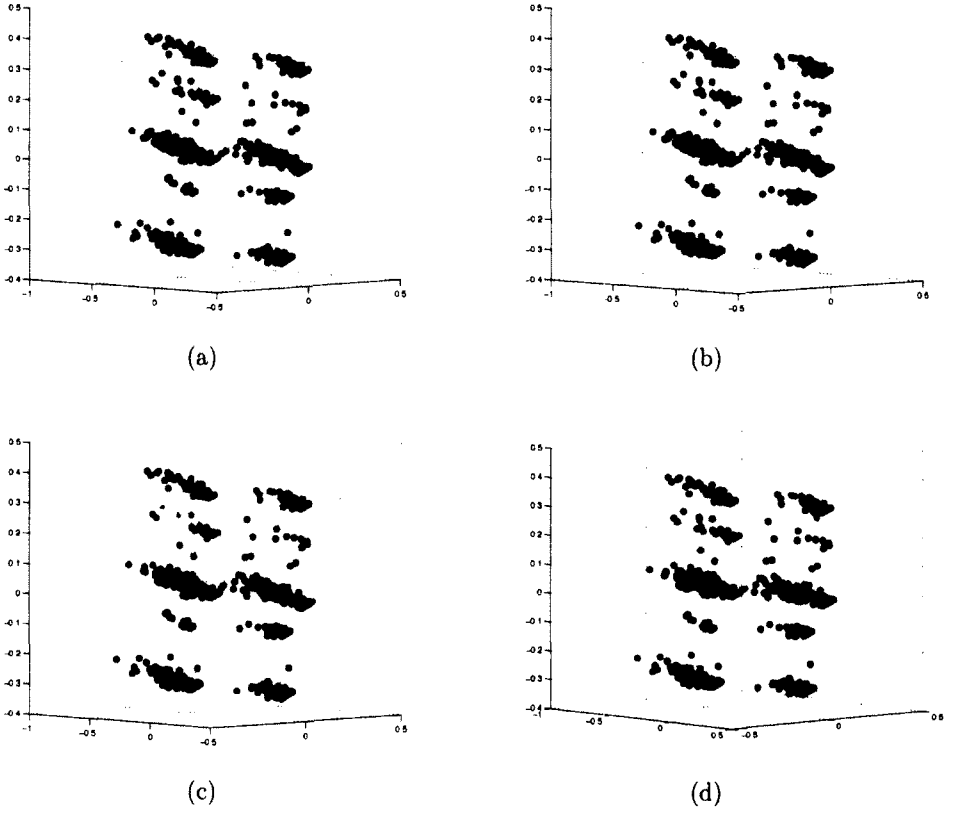


Figure 5.28: *Locating the bingers in a local cluster, diffusion map based on first 3 eigenfunctions: (a) The overall diffusion map with bingers being colored red; (b) The targeted cluster of participants (bingers and non-bingers included) is colored red, located by thresholding the first 2 eigenfunctions; (c) Participants in the targeted cluster is colored by the 4th eigenfunction; (d) Thresholding the 4th eigenfunction on the targeted cluster, using threshold value of 0.1.*

RELY ON GUT FEELINGS-W1", the 5th eigenfunction value separates those who have suicidal thoughts from others. Fig. 5.30 shows the embedding using 1st, 2nd and 5th eigenfunctions, which reveals some interesting relationship between the 5th eigenfunctions, suicidal thoughts, and gut feelings.

Fig. 5.31 summarizes the experiments in Waves 2 and 3 of the same AddHealth dataset, on the same participants. From Fig. 5.31a, the 10 W2 non-alcohol questions most correlated with the W2 binge drinking frequency are:

- S27Q44 SINCE MOLI,TRIED MARIJUANA-W2.
- S27Q48 3 FRIENDS-SMOKE POT ¿1 A MONTH-W2.
- S27Q1 SMOKED A CIGARETTE-W2.
- S23Q1 EVER TOUCH ANOTHERS GENITALS-W2.
- S23Q2 EVER HAVE SEX-W2.
- S27Q10 HOW MANY FRIENDS SMOKE-W2.
- S3Q46 NIGHT FROM HOME W/O PERMISS-W2.
- S33Q3 MOM-USE OF BIRTHCONTROL-W2.
- S33Q2 MOM-YOU HAVING SEX WITH STEADY-W2.
- S33Q1 MOM-FEEL ABOUT SEX NOW-W2.

Fig. 5.31b embeds the participants using these 10 W2 questions, and plots the W2 binging frequency labeling function on the map. Likewise, Fig. 5.31c yields another group of 10 W3 questions most correlated with the W3 binge drinking question (i.e., when respondents were age 19-20):

- S28Q103 3 BEST FRNDS H/MANY DRINK MON-W3.

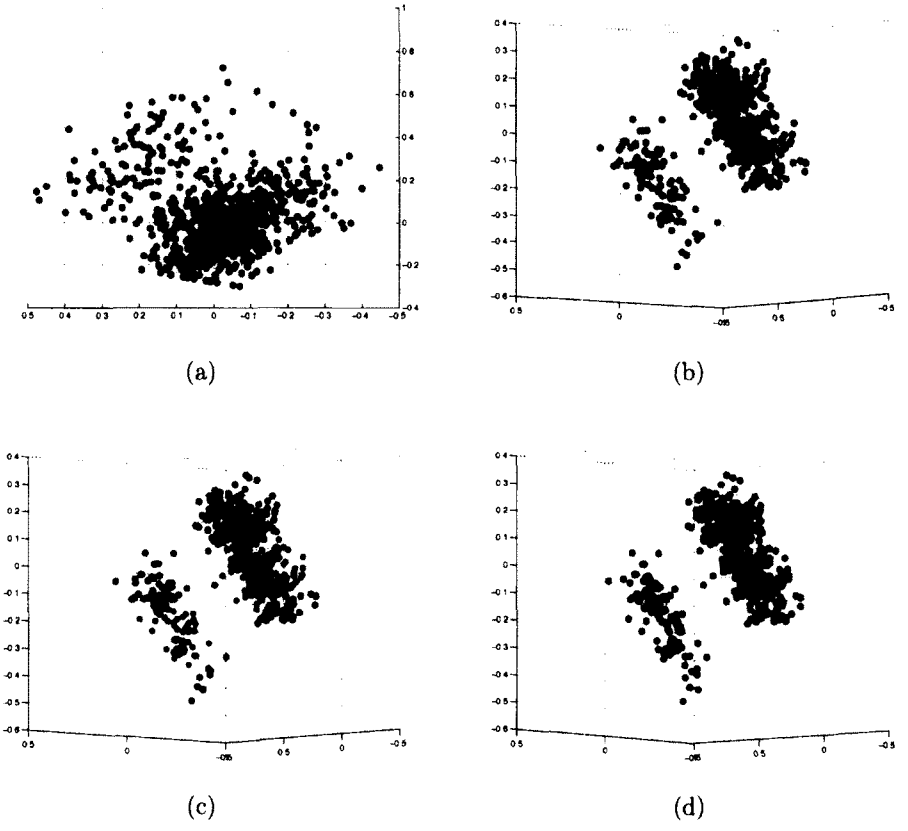
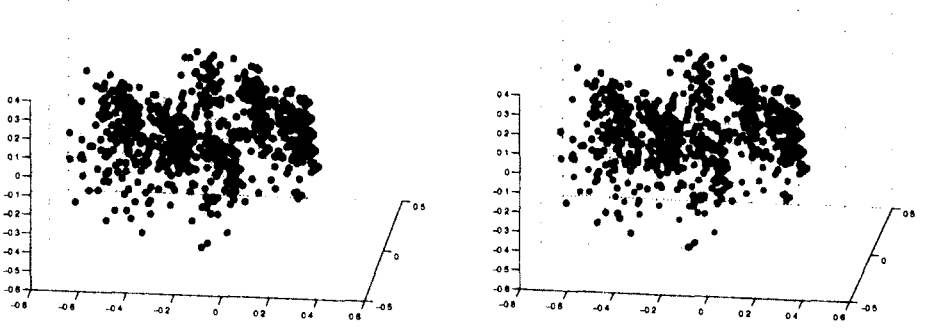


Figure 5.29: *Diffusion embeddings of non-African American participants, using the 10 selected W1 non-alcohol questions and the 4th, 6th and 5th eigenfunctions: (a) Embedding by 4th and 6th eigenfunctions, participants are colored based on answers to "S33Q1 PAST YR-THINK ABOUT SUICIDE-W1"; (b) Embedding by 4th, 6th and 5th eigenfunctions, participants are colored based on answers to "S33Q1 PAST YR-THINK ABOUT SUICIDE-W1", the addition of the 5th eigenfunction enhances the distinction between the two groups; (c) Embedding by 4th, 6th and 5th eigenfunctions, participants are colored based on answers to "S18Q16 RELY ON GUT FEELINGS-W1"; (d) Embedding by 4th, 6th and 5th eigenfunctions, bingers are colored in red.*

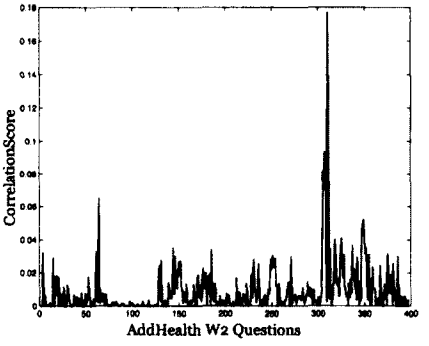


(a) (b)

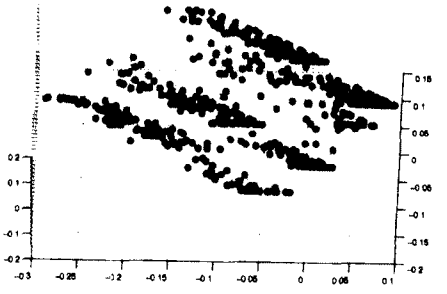
Figure 5.30: *Diffusion embeddings of non-African American participants by the 1st, 2nd and 5th eigenfunctions: (a) Participants are colored based on answers to "S33Q1 PAST YR-THINK ABOUT SUICIDE-W1"; (b) Participants are colored based on answers to "S18Q16 RELY ON GUT FEELINGS-W1".*

- S28Q37 SINCE JUN95 DRINK 2/3 TIMES-W3.
- S28Q108 SINCE JUN95 USED MARIJUANA-W3.
- S28Q104 3 BEST FRNDS H/MANY BINGE-W3.
- S28Q117 SINCE JUN95 USED OTH DRUG-W3.
- S12Q16 PAST 7 DAYS NEVER SWEAR-W3.
- S28Q105B SINC JUN95 TAKN TRANQUILIZER-W3.
- S26Q1 12 MO,OFT DAMAGE PROP/NOT YOUR-W3.
- S28Q1 EVER TRIED CIG SMOKING-W3.
- S28Q28 TRUE-LOOK FOR EXCITEMENT-W3.

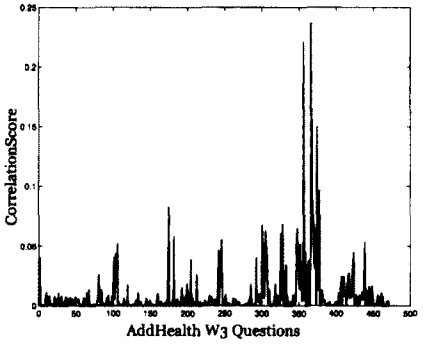
Fig. 5.31d uses these W3 questions to embed the participants, colors them by their binge drinking during W3. AddHealth Wave-3 questionnaire was conducted 5 years after Wave-2, meaning the participants were about 19-20 years old when they took W3 survey. The time gap and the age difference explains a stark contrast between



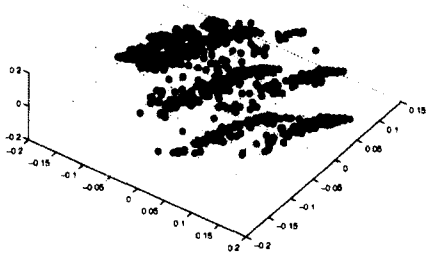
(a) Correlation scores of W-2 non-alcohol questions to W-2 binging behavior.



(b) Diffusion embedding of participants, using Gaussian kernel of 10 most correlated W-2 questions in (a).



(c) Correlation scores of W-3 non-alcohol questions to W-3 binging behavior.



(d) Diffusion embedding of participants, using Gaussian kernel of 10 most correlated W-3 questions in (c).

Figure 5.31: *Diffusion embeddings of non-African American participants in AddHealth W-2 (a, b) and W-3 (c, d) questionnaires. (Note that W-3 refers to responses when respondents were ages 19-20, i.e., 5 years after W-2)*

the W1, W2 list of relevant questions and the W3 list. The interested questions in W1, W2 show that early teens' binging behavior is related to different factors such as cigarette smoking, sex/romance, deviant behaviors, violence. However, in W3, the questions gear mostly toward smoking and drugs, with a rule-breaking, sensation seeking orientation, and involving substantially more drug use, of a variety of different types [104].

Chapter 6

Conclusions

We applied diffusion geometry to sociopolitical and public health datasets, revealing hidden patterns in the relationships among nations. Our experimental results yield interesting historical narratives. The fact that our findings, drawn only from the datasets, largely concur with those of political scientists and behavioral scientists, and indicate great potential of diffusion geometry in the analysis of security-related datasets, and in parsing other behavioral science, and public health data.

We presented a variety of experiments, both with synthetic and real data, to demonstrate the noise effect of irrelevant features in correlated datasets (such as surveys, votings, ratings). We showed that, under a supervised context, applying diffusion maps in an exploratory approach to the original dataset leads to unsatisfactory performance. We proposed a feature selection algorithm using second-order correlation, which demonstrably enhances the performance of diffusion methods in supervised applications on several public health datasets.

Appendix A

Diffusion maps & Diffusion distance

A.1 Random Walk

We begin by taking a microscopic view at the diffusion process in social network: a randomwalk [105]. Let us follow the path of a person "migrating" from country to country inside the network of capital cities, whose edge weights are defined according to the demographic gravitational model [106] $W_{ij} = \alpha \frac{N_i N_j}{r_{ij}^2}$, where N_i denotes the population of country i, r_{ij} provides the distance between capital cities of countries i and j, and α is a positive constant. We assume that for every time step t, if the person is at country i, the probability that he moves to country j in the next step is proportional to the demographic attraction W_{ij}. Setting the proportionality parameter to 1 for convenience, we then construct the transition matrix M whose entries M_{ij} denotes the probability the person moves from country i to country j.

$$M = D^{-1}W \tag{A.1.1}$$

where D is a diagonal matrix $D_{ii} = d_i = \sum_j W_{ij}$, called the degree matrix. M is by definition a stochastic matrix whose every row sums to one. Therefore, M can be interpreted as defining a random walk on G. If we examine a 1-by-n vector p_t of probability the person appears in each country at time t, we have:

$$p_{t+1}^T = p_t^T M = p_t^T D^{-1} W \tag{A.1.2}$$

Unless G is a regular graph (whose vertices have the same degree), M is an asymmetric matrix, which is difficult to deal with. Instead, we may opt to examine another symmetric matrix related to M:

$$\widetilde{M} = D^{1/2} M D^{-1/2} = D^{-1/2} W D^{-1/2} \tag{A.1.3}$$

The entries of $\widetilde{M}$ are thus $\widetilde{M}_{ij} = \frac{W_{ij}}{\sqrt{d_i}\sqrt{d_j}}$. Because $\widetilde{M}$ is real-valued and symmetric, it is diagonalizable and all of its eigenvalues are real. Eigenvector decomposition of $\widetilde{M}$ gives us $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ as the eigenvalues and $\{\upsilon_k\}$ their corresponding orthonormal eigenvectors. In matrix form:

$$\widetilde{M} = \Upsilon \Lambda \Upsilon^T \tag{A.1.4}$$

where Λ is the diagonal matrix with $\{\lambda_k\}$ on its diagonal, and Υ is a matrix whose columns are the corresponding eigenvectors $\{\upsilon_k\}$. Since $\widetilde{M}$ and M are similar, they have the same set of eigenvalues $\{\lambda_k\}$. In matrix form, we have:

$$M = D^{-1/2} \widetilde{M} D^{1/2} = D^{-1/2} \Upsilon \Lambda \Upsilon^T D^{1/2} \tag{A.1.5}$$

Thus, if we let

$$\begin{aligned} \phi_k &= D^{1/2} \upsilon_k \\ \psi_k &= D^{-1/2} \upsilon_k \end{aligned} \tag{A.1.6}$$

then

$$M = \Psi \Lambda \Phi^T \tag{A.1.7}$$

which implies that $\{\phi_k\}$ and $\{\psi_k\}$ defined in Eq. A.1.6 are the left and right eigenvectors of M corresponding to eigenvalues $\{\lambda_k\}$. Since $\{\upsilon_k\}$ are orthonormal vectors, it is easily seen from Eq. A.1.6 that ϕ_i and ψ_j are bi-orthonormal:

$$< \phi_i, \psi_j >= \delta_{ij} \tag{A.1.8}$$

where $< \cdot, \cdot >$ is the standard dot product. It is also verifiable that

$$\widetilde{M} d^{1/2} = D^{-1/2} W D^{-1/2} d^{1/2} = D^{-1/2} W \mathbb{1} = D^{-1/2} d = d^{1/2} \tag{A.1.9}$$

Therefore $d^{1/2}$ is an eigenvector of $\widetilde{M}$ with eigenvalue 1. Since the vector $d^{1/2}$ is positive, Perron-Frobenius theorem [107] tells us that its corresponding eigenvalue 1 is the largest eigenvalue of $\widetilde{M}$ and $\forall k \ |\lambda_k| \leq 1$. Thus $\lambda_1 = 1$. In fact, if G is connected (so that M represents an irreducible and aperiodic Markov chain) then $\forall k > 1 \ |\lambda_k| < 1 = \lambda_1$. We also have $\upsilon_1 = \frac{d^{1/2}}{\|d^{1/2}\|}$, which leads to $\phi_1 = \frac{d}{\|d^{1/2}\|}$ and $\psi_1 = \frac{1}{\|d^{1/2}\|}$. That means ψ_1 is a constant vector, while $\phi_1(i) = \frac{d_i}{\sqrt{\sum_k d_k}}$.

Let $p_t(j|i)$ be the probability distribution of the person stays at country j at time t, given that he starts out from country i at time 0. From Eq. A.1.2, we have:

$$p_t(j|i) = e_i^T M^t = e_i^T \Psi \Lambda^t \Phi^T = \sum_k \psi_k(i) \lambda_k^t \phi_k(j) \tag{A.1.10}$$

where e_i is a vector whose entry $e_i(k) = \delta_{ik}$. Hence, if G is connected (so that every eigenvalue other than λ_1 is less than 1), then regardless of the initial starting point:

$$lim_{t \to \infty} p_t(j|i) = \psi_1(i)\phi_1(j) = \frac{d_j}{\|d^{1/2}\|^2} = \frac{d_j}{\sum_k d_k} \tag{A.1.11}$$

The first eigenvector ϕ_1 therefore serves as the stationary distribution of the ran-

dom walk M. Furthermore, it can also be considered a density estimate, which tells us of how frequently our walker passes by a particular country. In social network terminology, it is the centrality vector. It is obvious that ϕ_1 is related to degree centrality. ϕ_1 is also similar to the eigenvector centrality of W [108], in the sense that it satisfies $\phi_1(i) = \sum_j \frac{W_{ij}}{d_j}\phi_1(j)$.

A.2 Diffusion Distance & Diffusion Maps

For each country i, we can imagine the diffusion process starts with an initial distribution $p_0(j|i) = \delta_{ij}$. After t steps, this distribution diffuses out to the neighborhood of i, with the landscape described by $p_t(j|i)$. The walker is more likely to end up in countries close to i than those far away. The difference between any 2 countries can be measured by:

$$D_t^2(i,j) = \|p_t(l|i) - p_t(l|j\|_\omega^2 = \sum_l^n (p_t(l|i) - p_t(l|j))^2\omega(l) \tag{A.2.1}$$

where $\omega(l) = \frac{1}{d_l}$ is the weight function, which normalize the distance by the centrality measure of each node. We call $D_t^2(i,j)$ the diffusion distance between i and j at time t. It can be seen as the weighted difference between the two distributions of concentrations after t steps of two random walks starting from nodes i and j. We also define diffusion map Ψ_t as the mapping between the original data space onto the first κ left eigenvectors of M:

$$\Psi_t(i) = (\lambda_1^t\psi_1(i), \lambda_2^t\psi_2(i), \ldots, \lambda_\kappa^t\psi_\kappa(i)) \tag{A.2.2}$$

It is easily verifiable that the diffusion distance in Eq. A.2.1 is equal to Euclidean distance in the diffusion map space:

$$
\begin{aligned}
D_t^2(i,j) &= \sum_{l}^{n} \left(\sum_{k}^{\kappa} \lambda_k^t (\psi_k(i) - \psi_k(j)) \phi_k(l) \right)^2 \frac{1}{d_l} \\
&= \sum_{k_1,k_2}^{\kappa} \lambda_{k_1}^t \Big(\psi_{k_1}(i) - \psi_{k_1}(j) \Big) \lambda_{k_2}^t \Big(\psi_{k_2}(i) - \psi_{k_2}(j) \Big) \sum_{l}^{n} \frac{\phi_{k_1}(l)\phi_{k_2}(l)}{d_l} \\
&= \sum_{k_1,k_2}^{\kappa} \lambda_{k_1}^t \Big(\psi_{k_1}(i) - \psi_{k_1}(j) \Big) \lambda_{k_2}^t \Big(\psi_{k_2}(i) - \psi_{k_2}(j) \Big) \sum_{l}^{n} \frac{d_l \psi_{k_1}(l)\phi_{k_2}(l)}{d_l} \\
&= \sum_{k_1,k_2}^{\kappa} \lambda_{k_1}^t \Big(\psi_{k_1}(i) - \psi_{k_1}(j) \Big) \lambda_{k_2}^t \Big(\psi_{k_2}(i) - \psi_{k_2}(j) \Big) \sum_{l}^{n} \psi_{k_1}(l)\phi_{k_2}(l) \\
&= \sum_{k_1,k_2}^{\kappa} \lambda_{k_1}^t \Big(\psi_{k_1}(i) - \psi_{k_1}(j) \Big) \lambda_{k_2}^t \Big(\psi_{k_2}(i) - \psi_{k_2}(j) \Big) \delta_{k_1 k_2} \\
&= \sum_{k}^{\kappa} \lambda_k^{2t} \Big(\psi_k(i) - \psi_k(j) \Big)^2 \\
&= \| \Psi_t(i) - \Psi_t(j) \|^2
\end{aligned}
\tag{A.2.3}
$$

Practically, only the last $(\kappa - 1)$ coordinates are to be considered because ψ_1 is a constant vector. Additionally, since $\forall k \ |\lambda_k| <= 1$, components $\lambda_k^t \psi_k(i)$ in Eq. A.2.2 corresponding to smaller values of λ_k start to vanish as t increases. The larger the value of t, the more components of $\Psi_t(i)$ are brought down to 0, thus reducing the dimensionality of the embedded space.

Bibliography

[1] The netflix prize, 2009. URL `http://www.netflixprize.com/`.

[2] G. Linden, B. Smith, and J. York. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, 2003.

[3] S. R. Hathaway and J. C. McKinley. A multiphasic personality schedule (minnesota): I. construction of the schedule. *J. of Clinical Psychology*, 10:249–254, 1940.

[4] J. N. Butcher, J. R. Graham, Y. S. Ben-Porath, A. Tellegen, W. G. Dahlstrom, and B. Kaemmer. *MMPI-2: Minnesota Multiphasic Personality Inventory-2: Manual for administration and scoring (Rev. ed.).* Minneapolis: University of Minnesota Press, 2001.

[5] B. F. Grant and D. A. Dawson. Introduction to the national epidemiologic survey on alcohol and related conditions. *Alcohol Health & Research World*, 29 (2):74–78, 2006.

[6] S. H. Stewart, K. T. Borg, and P. M. Miller. Prevalence of problem drinking and characteristics of a single-question screen. *J. of Emergency Medicine*, 39 (3):291–295, 2010.

[7] M.D. Resnick, P.S. Bearman, R.W. Blum, K.E. Bauman, K.M. Harris, J. Jones, J. Tabor, T. Beuhring, R. E. Sieving, M. Shew, M. Ireland, L. H. Bearinger,

and J. R. Udry. Protecting adolescents from harm: Findings from the national longitudinal study on adolescent health. *J. of Am. Medical Association*, 278 (10):823–832, 1997.

[8] P. S. Bearman, J. Jones, and J. R. Udry. *The National Study of Adolescent Health: research design.* Chapel Hill NC: Carolina Population Center, 1997. URL `www.cpc.unc.edu/addhealth/design`.

[9] A. Strezhnev and E. Voeten. United nations general assembly voting data, 2012. URL `http://hdl.handle.net/1902.1/12379`.

[10] K. S. Gleditsch. Expanded trade and gdp data. *J. of Conflict Resolution*, 46 (5):712–724, 2002.

[11] J. C. Pevehouse, T. Nordstrom, and K. Warnke. The COW-2 international organizations dataset version 2.0. *Conflict Management and Peace Science*, 21 (2):101–119, 2004.

[12] R. Bayer. Diplomatic exchange data set, v2006.1., 2006. URL `http://correlatesofwar.org`.

[13] S. P. Borgatti. 2-mode concepts in social network analysis. In R. A. Meyers, editor, *Encyclopedia of complexity and system science*, pages 8279–8291. Springer, 2009.

[14] L. K. Saul and S. T. Roweis. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *The Journal of Machine Learning Research*, 4: 119–155, 2003.

[15] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[16] M. A. Figueiredo and A. K. Jain. Unsupervised learning of finite mixture models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(3): 381–396, 2002.

[17] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[18] A. K. Menon and C. Elkan. Link prediction via matrix factorization. In *Machine Learning and Knowledge Discovery in Databases*, volume 6912 of *Lecture Notes in Computer Science*, pages 437–452. Springer Berlin Heidelberg, 2011.

[19] S. Y. Kim and B. M. Russett. Voting alignments in the general assembly. In B. M. Russett, editor, *The Once and Future Security Council*, pages 29–57. New York: St. Martin's Press, 1997.

[20] S. Holloway. Forty years of united nations general assembly voting. *Canadian Journal of Political Science*, 23:279–296, 1990.

[21] E. Y. Mun, A. von Eye, M. E. Bates, and E.G. Vaschillo. Finding groups using model-based cluster analysis: Heterogeneous emotional self-regulatory processes and heavy alcohol use risk. *Developmental Psychology*, 44(2):481–495, 2008.

[22] D. Steinley and M. J. Brusco. Choosing the number of clusters in -means clustering. *Psychological Methods*, 16(3):285–297, 2011.

[23] A. G. Long. Bilateral trade in the shadow of armed conflict. *ISQ*, 52:81–101, 2008.

[24] S. B. Blomberg and G. D. Hess. How much does violence tax trade? *The Review of Economics and Statistics*, 88(4):599–612, 2006.

[25] H. Hegre, J. R. Oneal, and B. Russett. Trade does promote peace: New simultaneous estimates of the reciprocal effects of trade and conflict. *J. of Peace Research*, 47(6):763–774, 2010.

[26] J. R. Oneal, B. M. Russett, and M. L. Berbaum. Causes of peace: Democracy, interdependence, and international organizations, 1885ñ1992. *International Studies Quarterly*, 47:371–393, 2003.

[27] M. S. de Vries. Interdependence, cooperation and conflict: An empirical analysis. *J. of Peace Research*, 27(4):429–444, 1990.

[28] R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *PNAS*, 102(21):7426–7431, 2005.

[29] S. Wasserman and K. Faust. *Social network analysis: methods and applications.* New York: Cambridge University Press, 1994.

[30] V. E. Krebs. Mapping networks of terrorist cells. *Connections*, 24(3):43–52, 2002.

[31] M. E. Newman. The structure of scientific collaboration networks. *PNAS*, 98 (2):404–409, 2001.

[32] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis. Tastes, ties, and time: A new social network dataset using facebook.com. *Social networks*, 30(4):330–342, 2008.

[33] J. F. Padgett and C. K. Ansell. Robust action and the rise of the medici, 1400-1434. *Am. J. of Sociology*, 98(6):1259–1319, 1993.

[34] Z. Maoz. *Networks of Nations: The Evolution, Structure, and Impact of International Networks, 1816-2001*, volume 32 of *Structural Analysis in the Social Sciences*. New York: Cambridge University Press, 2011.

[35] E. M. Hafner-Burton, M. Kahler, and A. H. Montgomery. Network analysis for international relations. *International Organization*, 63(3):559–592, 2009.

[36] D. M. Stinnett, J. Tir, P. Schafer, P. F. Diehl, and C. Gochman. The correlates of war project direct contiguity data, version 3. *Conflict Management and Peace Science*, 19(2):58–66, 2002.

[37] K. S. Gleditsch and M. D. Ward. Measuring space: A minimum-distance database and applications to international studies. *J. of Peace Research*, 38 (6):739–758, 2001.

[38] P. C. Ordeshook. The spatial theory of elections: A review and a critique. In I. Budge, I. Crewe, and D. Farlie, editors, *Party Identification and Beyond*, page 308. Wiley, New York, 1976.

[39] C. Zimmer, G. Schneider, and M. Dobbins. The contested council: Conflict dimensions of an intergovernmental eu institution. *Political Studies*, 53(2):403–422, 2005.

[40] P. Trubowitz. Sectionalism and american foreign policy: The political geography of consensus and conflict. *International Studies Quarterly*, 36(2):173–190, 1992.

[41] A. L. Comrey. A factor analysis of items on the mmpi depression scale. *Educational and Psychological Measurement*, 17:578–585, 1957.

[42] W. W. Deardorff, A. F. Chino, and D. W. Scott. Characteristics of chronic pain patients: Factor analysis of the mmpi-2. *Pain*, 54(2):153–158, 1993.

[43] B. O. Muthén. Should substance use disorders be considered as categorical or dimensional? *Addiction*, 101(s1):6–16, 2006.

[44] P. Scal, M. Ireland, and I. W. Borowsky. Smoking among american adolescents: A risk and protective factor analysis. *J. of Community Health*, 28(2):79–97, 2003.

[45] J. B. Kruskal and M. Wish. *Multidimensional Scaling.* Beverly Hills, CA: Sage, 1978.

[46] H. F. Weisberg and J. G. Rusk. Dimensions of candidate evaluation. *American Political Science Review*, 64:1167–1185, 1970.

[47] K. T. Poole. *Spatial Models of Parliarmentary Voting*, chapter 2, pages 26–28. New York: Cambridge University Press, 2005.

[48] Research Triangle Institue. *Software for Survey Data Analyses (SUDAAN) Version 10.* Research Triangle Park, NC: RTI, 2008. URL `http://www.rti.org/sudaan/`.

[49] L. K. Muthén and B. O. Muthén. Mplus. *The comprehensive modelling program for applied researchers: User's guide*, 5, 2012.

[50] H. Abdi and L. J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010.

[51] M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *J. of the Royal Statistical Society*, 21(3):611–622, 1999.

[52] S. Roweis. Em algorithms for pca and spca. In M. I. Jordan, M. J. Kearns, and S. A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10, pages 626–632. MIT Press, 1998.

[53] Basilevsky. *Statistical Factor Analysis and Related Methods: Theory and Applications.* Wiley, 1994.

[54] D. B. Rubin and D. T. Thayer. Em algorithms for ml factor analysis. *Psychometrika*, 47(1):69–76, 1982.

[55] L. Guttman. Order analysis of correlation matrices. In *Handbook of multivariate experimental psychology*, pages 439–458. Chicago: Rand McNally, 1966.

[56] J. C. Lingoes. Some boundary conditions for a monotone analysis of symmetric matrices. *Psychometrika*, 36(2):195–203, 1971.

[57] M. L. Davison. Multidimensional scaling versus component analysis of test intercorrelations. *Psychological Bulletin*, 97(1):94–105, 1985.

[58] T. J. Brazill and B. Grofman. Factor analysis versus multi-dimensional scaling: binary choice roll-call voting and the us supreme court. *Social Networks*, 24(3): 201–229, 2002.

[59] J. B. Kruskal. Multidimensional scaling by optimizing a goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.

[60] J. B. Kruskal. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29:115–129, 1964.

[61] K. T. Poole and H. Rosenthal. A spatial model for legislative roll call analysis. *Am. J. of Political Science*, 29(2):357–384, 1985.

[62] R. Kannan, S. Vempala, and A. Veta. On clusterings: Good, bad and spectral. *J. of the ACM*, 51(3):497–515, 2004.

[63] C. Coombs. *A Theory of Data.* New York: Wiley, 1964.

[64] S. Lafon. *Diffusion Maps and Geometric Harmonics.* PhD thesis, Yale University, May 2004.

[65] R. R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

[66] S. Lafon and A. B. Lee. Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(9): 13934–1403, 2006.

[67] A. Singer and R. R. Coifman. Non-linear independent component analysis with diffusion maps. *Applied and Computational Harmonic Analysis*, 25(2):226–239, 2008.

[68] R. R. Coifman and M. Gavish. Harmonic analysis of digital databases. In J. Cohen and A. Zayed, editors, *Wavelets and Multiscale Analysis: Theory and Applications*, pages 161–197. Birkhauser Basel, 2011.

[69] R. R. Coifman and M. Gavish. Harmonic anaylsis of databases and matrices. In T. D. Andrews, R. Balan, J. J. Benedetto, W. Czaja, and K. A. Okoudjou, editors, *Excursions in Harmonic Analysis*, volume 1, pages 297–310. Springer New York, 2013.

[70] S. Lafon, Y. Keller, and R. R. Coifman. Data fusion and multicue data matching by diffusion maps. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(11):1784–1797, 2006.

[71] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis. Diffusion maps, spectral clustering and reaction coordinates of dynamical systems. *Applied and Computational Harmonic Analysis*, 21(1):113–127, 2005.

[72] R. R. Coifman, Y. Shkolnisky, F. J. Sigworth, and A. Singer. Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. *Multiscale Modeling & Simulation*, 7(2):842–864, 2008.

[73] Y. Keller, S. Lafon, R. R. Coifman, and S. W. Zucker. Audio-visual group recognition using diffusion maps. *IEEE Trans. on Signal Processing*, 58(1): 403–413, 2010.

[74] A. Singer and R. R. Coifman. Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps. *PNAS*, 106(38):16090–16095, 2009.

[75] R. Xu, S. Damelin, B. Nadler, and D. C. Wunsch II. Clustering of high-dimensional gene expression data with feature filtering methods and diffusion maps. *Artificial Intelligence in Medicine*, 48(2):91–98, 2010.

[76] A. Schclar, A. Averbuch, N. Rabin, V. Zheludev, and K. Hochman. A diffusion framework for detection of moving vehicles. *Digital Signal Processing*, 20(1): 111–122, 2010.

[77] D. Kushnir, A. Haddad, and R. R. Coifman. Anisotropic diffusion on sub-manifolds with application to earth structure classification. *Applied and Computational Harmonic Analysis*, 32(2):280–294, 2012.

[78] R. Talmon, D. Kushnir, R. R. Coifman, I. Cohen, and S. Gannot. Parametrization of linear systems using diffusion kernels. *IEEE Trans. on Signal Processing*, 60(3):1159–1173, 2012.

[79] R. Talmon, S. Cohen, I.and Gannot, and R. R. Coifman. Supervised graph-based processing for sequential transient interference suppression. *IEEE Trans. on Audio Speech and Language Processing*, 20(9):2528–2538, 2012.

[80] A. Bermanis, A. Averbuch, and R. R. Coifman. Multiscale data sampling and function extension. *Applied and Computational Harmonic Analysis*, 34(1):15–29, 2013.

[81] L. C. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1:215–239, 1978.

[82] B. J. Kinne. *Beyond the Dyad: How Networks of Economic Interdependence and Political Integration Reduce Interstate Conflict.* PhD thesis, Yale University, 2009.

[83] M. A. Porter, P. J. Mucha, M. E. J. Newman, and C. M. Warmbrand. A network analysis of committees in the u.s. house of representatives. *PNAS*, 102 (20):7057–7062, 2005.

[84] H. R. Alker. Dimensions of conflict in the general assembly. *The American Political Science Review*, 58(3):642–657, 1964.

[85] R. Rogers, K. W. Sewell, K. S. Harrison, and M. J. Jordan. The mmpi-2 restructured clinical scales: A paradigmatic shift in scale development. *J. of Personality Assessment*, 87(2):139–147, 2006.

[86] A. Tellegen, Y. S. Ben-Porath, J. L. McNulty, P. A. Arbisi, J. R. Graham, and B. Kaemmer. *The MMPI-2 Restructured Clinical (RC) scales: Development, validation, and interpretation.* Minneapolis: University of Minnesota Press, 2003.

[87] D. C. Watson and B. K. Sinha. Dimensional structure of personality disorder inventories: A comparison of normal and clinical populations. *Personality and Individual Differences*, 19(6):817–826, 1995.

[88] A. L. Comrey. A factor analysis of items on the mmpi hysteria scale. *Educational and Psychological Measurement*, 17:586–592, 1957.

[89] J. B. Hoelzle and G. J. Meyer. The factor structure of mmpi-2 restructured clinical (rc) scales. *J. of Personality Assessment*, 90(5):443–455, 2008.

[90] D. A. Dawson, B. F. Grant, and J. Ruan. The association between stress and drinking: Modifying effects of gender and vulnerability. *Alcohol and Alcoholism*, 40(5):453–460, 2005.

[91] A. Agrawal, M. T. Lynskey, P. A. F. Madden, K. K. Bucholz, and A. C. Heath. A latent class analysis of illicit drug abuse/dependence: results from the national epidemiological survey on alcohol and related conditions. *Addiction*, 102(1): 94–104, 2007.

[92] B. F. Grant, T. C. Harford, B. O. Muthén, H. Yi, D. S. Hasin, and F. S. Stinson. Dsm-iv alcohol dependence and abuse: Further evidence of validity in the general population. *Drug and Alcohol Dependence*, 86(2–3):154–166, 2007.

[93] K. Markon and R. F. Krueger. Categorical and continuous models of liability to externalizing disorders: A direct comparison in nesarc. *Archives of General Psychiatry*, 62(12):1352–1359, 2005.

[94] D. S. Hasin and C. L. Beseler. Dimensionality of lifetime alcohol abuse, dependence and binge drinking. *Drug and Alcohol Dependence*, 101(1–2):53–61, 2009.

[95] N. Carragher and L. A. McWilliams. A latent class analysis of dsm-iv criteria for pathological gambling: Results from the national epidemiologic survey on alcohol and related conditions. *Psychiatry Research*, 187(1–2):185–192, 2011.

[96] D. D. Hallfors, M. W. Waller, C. A. Ford, C. T. Halpern, P. H. Brodish, and B. Iritani. Adolescent depression and suicide risk: Association with sex and drug behavior. *Am. J. of Preventive Medicine*, 27(3):224–231, 2004.

[97] P. D. Quinn and K. P. Harden. Behind the wheel and on the map: Genetic and environmental associations between drunk driving and other externalizing behaviors. *J. of Abnormal Psychology*, 122(4):1166–1178, 2013.

[98] L. T. Hoyt, P. L. Chase-Lansdale, T. W. McDade, and E. K. Adam. Positive youth, healthy adults: Does positive well-being in adolescence predict better perceived health and fewer risky health behaviors in young adulthood? *J. of Adolescent Health*, 50(1):66–73, 2012.

[99] Inderjit S Dhillon, Edward M Marcotte, and Usman Roshan. Diametrical clustering for identifying anti-correlated gene clusters. *Bioinformatics*, 19(13):1612–1619, 2003.

[100] K. Fukumizu, F. R. Bach, and I. Jordan, M. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *J. of Machine Learning Research*, 5:73–99, 2004.

[101] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *Algorithmic Learning Theory*, Lectures Notes in Artificial Intelligence 374, pages 63–77. Springer Berlin Heidelberg, 2005.

[102] R. A. Zucker, M. M. Heitzeg, and J. T. Nigg. Parsing the undercontroldisinhibition pathway to substance use disorders: A multilevel developmental problem. *Child Development Perspectives*, 5(4):248–255, 2011.

[103] R. Whelan, R. Watts, C. A. Orr, R. R. Althoff, E. Artiges, T. Banaschewski, G. J. Barker, A. L. W. Bokde, C. Buchel, F. M. Carvalho, P. J. Con-

rod, H. Flor, M. Fauth-Buhler, V. Frouin, J. Gallinat, G. Gan, P. Gowland, A. Heinz, B. Ittermann, C. Lawrence, K. Mann, J.-L. Martinot, F. Nees, N. Ortiz, M.-L. Paillere-Martinot, T. Paus, Z. Pausova, M. Rietschel, T. W. Robbins, M. N. Smolka, A. Strohle, G. Schumann, H. Garavan, and the IMAGEN Consortium. Neuropsychosocial profiles of current and future adolescent alcohol misusers. *Nature*, advance online publication:–, 07 2014. URL `http://dx.doi.org/10.1038/nature13402`.

[104] M. A. Cyders and G. T. Smith. Emotion-based dispositions to rash action: positive and negative urgency. *Psychological Bulletin*, 134(6):807–828, 2008.

[105] B. Nadler, S. Lafon, R. Coifman, and I. G. Kevrekidis. Diffusion maps - a probabilistic interpretation for spectral embedding and clustering algorithms. In *Principal Manifolds for Data Visualization and Dimension Reduction*, volume 58 of *Lecture Notes in Computer Science*, pages 69–96. Springer, 2008.

[106] J. Q. Stewart. Empirical mathematical rules concerning the distribution and equilibrium of population. *American Geographical Society*, 37(3):461–485, 1947.

[107] C. D. Mayer. *Matrix Analysis and Applied Linear Algebra*, chapter 8, pages 661–704. SIAM, 2000.

[108] P. Bonacich. Factoring and weighting approaches to status scores and clique identification. *J. of Mathematical Sociology*, 2:113–120, 1972.