

Using the First Three Eigenfunctions of the  
Combinatorial Laplacian to Reduce  
Dimensionality of Higher Dimensional Data Sets

Amna Hyder, under the supervision of Dr. Min-Oo

April 19, 2014

## Abstract

Pattern recognition in machine learning aims to reduce a large dimensional data set to a lower dimensional manifold contained within it in order to eliminate redundant information. For example, a photograph with  $n^2$  pixels, corresponding to an  $n$  by  $n$  image, would have a dimensionality given by  $R^{n^2}$ , however, this is likely higher than the intrinsic dimensionality of the image, which is dependent on the degrees of freedom of the camera. Principle Component Analysis is one of the most popular methods used to reduce the dimensionality of data sets by producing a set of linearly uncorrelated variables in a new coordinate system. However, PCA does not consider the intrinsic structure of the manifold, which motivates us to use another method from graph theory that would take such topology into account.

Graph theory is a field of mathematics that gained popularity in most scientific disciplines in the 1990s. In graph theory any abstract object or collection of objects can be represented by **vertices**,  $V = \{v_1, \dots, v_i, \dots\}$ , and the links between them by **edges (E)**. A **graph**  $G$ , also known as a network, is an ordered pair,  $G = (V, E)$  consisting of these vertices and edges. A graph can be directed or undirected. In a directed graph, the direction of the edges is specified and  $e_{ij} \neq e_{ji}$ . The combinatorial laplacian is a matrix representation of the graph that preserves local information and considers the geometry of the manifold contained within it. The first three eigenfunctions of the combinatorial laplacian for two different data sets, protein-protein interactions and carcinogenicity of hydrocarbons are plotted to determine key features. Roughly twenty thousand different properties of 303 different polyaromatic hydrocarbons are plotted along with carcinogenicity to determine which features play the biggest roles in the toxicity of these compounds. The identification of patterns in this data set through the eigenfunctions of the combinatorial laplacian may

help determine which features can be used as predictors for polycyclic aromatic hydrocarbon carcinogenicity.

Mortgages are the largest asset that every bank has. Accurate predictions of mortgage prepayment rates is a critical factor in determining borrowing rates for banks. Mortgage prepayment rates are dependent on a variety of different factors, and spectral analysis of these factors may allow banks to make more accurate predictions. Part of this study aims to use graph theory to generate a model that allows for more accurate mortgage prepayment rates.

Protein sequences fold into three dimensional structures as a result of long range and short range interactions between base pairs. Although for the most part this structure can be determined by the nuclear magnetic resonance techniques (NMR) or x-ray crystallography, these are very expensive and time consuming processes. Thus, there is a significant drive in computational biology to develop a method that can predict three dimensional structures from a protein sequence. The third part of this thesis aims to see how graph theory plays a role in protein folding.

## 1 History of Graph Theory

Network analysis is a research area that dates back to 1736 with Leonhard Euler's solution of the Königsberg bridge problem [10, 1]. The city of Königsberg was built on two islands that were connected by a series of seven bridges. According to legend, countless hours were spent by the people of this city in trying to determine whether or not there existed a path that would cross each of the bridges exactly once (a popular brain teaser of the time). Euler was the first

to prove that such a path could not exist by representing the land masses as vertices and the bridges connecting them as edges. He said that the bridge problem could be solved by finding the existence of a 'Eulerian path' a path which traverses each edge only once, and showed that no such path existed for the bridges of Königsberg (the elementary proof of this can be found in Newman et al., **The Structure and Dynamics of Networks**). This proof is considered by many the first formal theorem in **graph theory** [10, 1, 5]. Graph theory gained popularity in the late 1940s, when sociologists and anthropologists were trying to gain mathematical tools to interpret data from their studies.

In 1929, the author Frigyes Karinthy published a collection of short stories among which was the story *Chains*. Although this was not a scientific work, the book stirred up one of the deepest questions in graph theory: the small world problem. In this story, he claimed that people become increasingly connected to each other by each degree of their connections, forming a small world network. Without any mathematical or scientific rigour, he stated that any two people globally had five degrees of separation, and that globalization would likely minimize this degree of separation significantly [11].

In 1967, Stanley Milgram looked at the popular *small world problem*, which allowed people to determine their social connectivity, and helped uncover the average path length between people in the world [11]. In the 1990s, mathematicians begin to develop tools to analyze large scale networks, expediting the use of graph theory as a practical tool of data analysis in nearly every field. It has found applications in biology, the internet, chemistry and physics among several others.

## 2 Graph Theory

In graph theory any abstract object or collection of objects can be represented by **vertices**,  $V = \{v_1, \dots, v_i, \dots\}$ , and the links between them by **edges (E)** [11]. A **graph**  $G$ , also known as a network, is an ordered pair,  $G = (V, E)$  consisting of these vertices and edges. Each edge can have a corresponding weight  $w(e)$ , that describes the strength of this connection. These weights can range from values of 0 to 1 where a weight of 1 shows the strongest correlation between the two and that of 0 implies that the two nodes are not connected. However some types of graphs can have weights that do not lie within this range, and this must be taken into account in the analysis. The inverse of the weight  $w(e)^{-1}$  can be thought to represent the distance between the vertices it connects and the sum of these distances as the path length of the graph.

### 2.1 Types of Graphs

A graph can be directed or undirected. In a directed graph, the direction of the edges is specified and  $e_{ij} \neq e_{ji}$ . Directed graphs are common in electrical circuits where current flows in a specific direction.

In a simple graph, two vertices can only be connected by at most one edge, and a vertex can not be connected to itself (with what is known as a loop). A multigraph can have loops or multiple edges connecting vertices.

Moreover, a graph can also be weighted or unweighted. In an unweighted graph, the edges are given values of either one or zero in the adjacency matrix depending on if they connect a pair of vertices or not. A graph can be isomorphic to another graph if the weights of the edges are not the same as long as the connectivity between the vertices and edges is the identical.

A bipartite graph is a graph in which the vertices can be divided into two distinct sets such that every edge in the graph directly connects a vertex from one set to one from the other. A quasi bipartite graph shows properties of a bipartite graph and has more use in applications.

A knot in a directed graph such that each vertex in the knot has edges that come out of it that terminate at other vertices in the knot. Thus, it is impossible to leave the knot (as in knot theory) while following along the directions of the edges.

## 2.2 Graph Properties

A subgraph of a graph is one whose vertices and edges are contained within the graph. If there is a set of vertices  $S$  in a subgraph  $G(S)$  such that there is an edge between any two vertices in  $G(S)$ , then the  $S$  is called a *clique*. A maximum clique is a clique which is not a proper subset of another clique.

The distance between two vertices (path length) can be measured in terms of the number of edges connecting them along with their corresponding weights (the inverse of these weights). The eccentricity of a vertex  $v$  in a graph is the distance from  $v$  to the farthest vertex in  $G$  from it. The vertex with the minimum eccentricity is defined in graph theory as the center of the graph. We use several distance measures in the following applications including mahalanobis distance and frobenius distance.

A clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together.

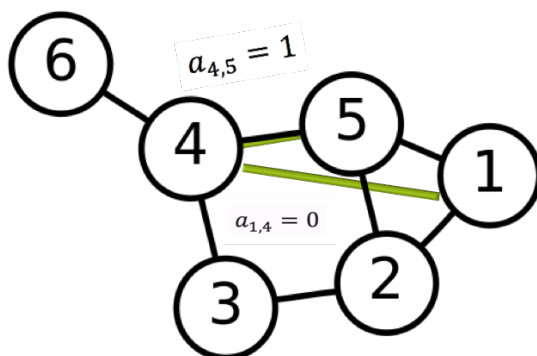


Figure 1: *Constructing the adjacency matrix from a graph.*

### 2.3 Graph Representation

The adjacency matrix shows how the vertices in a graph are connected. Essentially, the adjacency matrix for a graph  $G$  (with  $n$  vertices) is a  $n \times n$  matrix such that Matrix  $A = a_{ij}$

$$a_{ij} = \begin{cases} 1 & \text{For an edge from vertex } i \text{ to vertex } j \\ 0 & \text{otherwise} \end{cases}$$

$A$  is a symmetric matrix with zeros on the diagonal. The **degree** of a vertex describes the number of times an edge terminates at that vertex. The sum of  $i^{\text{th}}$ -row of  $A$ , denoted by  $d_i$  represents the degree of each vertex and the diagonal matrix of these degrees  $(d_1, d_2 \dots d_n)$ , is termed the **degree matrix**,  $D$ .

The edge incidence matrix encapsulates all the information contained in the adjacency matrix, with an additional factor of direction. An  $m \times n$  incidence matrix,  $B$  has columns indexed by edges and rows indexed by vertices. By choosing an arbitrary orientation for each edge, and once for each column, we arrive at the following definition for an incidence matrix  $B$ :

$$b_{ij} = \begin{cases} 1 & \text{For an edge from vertex } i \text{ to vertex } j \\ -1 & \text{For an edge from vertex } j \text{ to } i \\ 0 & \text{otherwise} \end{cases}$$

### 3 The shape of data

Two postulates that we present are that:

1. Data has a shape
2. This shape is important

A simple crude example of shape is how many pieces the data breaks into and the statistical version of this is that it breaks into clusters. This is an important piece of information to have because the clusters often correspond to conceptually different parts of the phenomenon. Another example is the predator prey model which obeys a law that ultimately causes it to loop back on itself and indicates periodic or recurrent behavior.

We immediately create visual representations out of data that we are given. For example, given the problem of determining the relationship between the heights and shoe sizes of a group of people, most would show this in the form of a two dimensional graph with a linear trend. In other words, we would translate the data into a geometric problem. But what happens if instead of two dimensions we have data in  $n$  dimensions. If we were creatures a higher dimensional space, we could represent the points in that space to understand more complex trends. However, as we are confined to three dimensions, we hope that by plotting the eigenfunctions of the combinatorial laplacian in three dimensions we hope to analyze the structure of a higher dimensional data set in a lower dimensional space that still captures and preserves its higher dimensional structure.



Considering an image with  $n^2$  pixels, the dimensionality corresponds to the number of pixels. However, the intrinsic dimensionality of the image depends on the degrees of freedom of the camera, and is likely much lower than the pixel dimension. Pattern recognition in machine learning would argue that there is a lower dimensional manifold embedded within this higher dimensional space that contains all the information about the image. Principle component analysis is one of the most popular methods used to reduce the dimensionality of the data set by producing a set of linearly uncorrelated variables in a new coordinate system, however PCA does not consider the intrinsic structure of this hypothetical manifold. Moreover, when considering the problem of a swiss roll, in which a 2-dimensional data set is rolled up and embedded in two dimensional space, PCA miscalculates the distance between two points and can not recover the two dimensional data set. However we see that our method of plotting the first three eigenfunctions does recover the 2 dimensional data set.

## 4 Introduction of Graph Theory for Data Analysis

The combinatorial laplacian is another matrix representation of the graph that preserves local information and considers the geometry of the manifold contained within it. It is easy to prove that it is the discrete version of the continuous laplacian operator which is defined by the gradient squared and is central to equations in thermodynamics, quantum mechanics and pure mathematics.

$$L = \nabla^2 = \frac{\partial^2}{\partial^2 x} + \frac{\partial^2}{\partial^2 y} + \frac{\partial^2}{\partial^2 z} \quad (1)$$

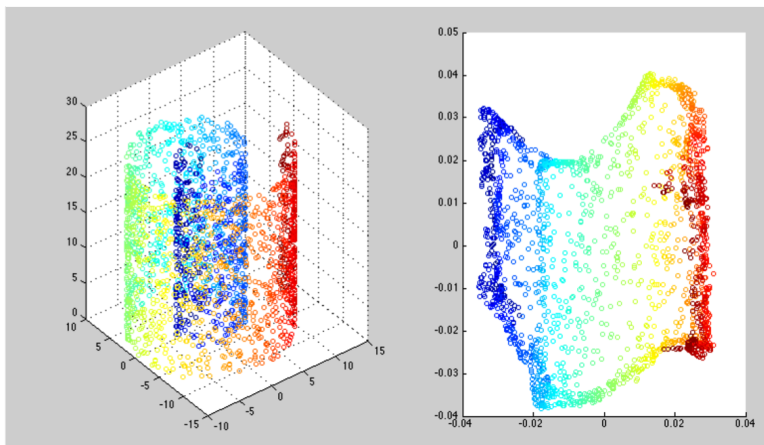


Figure 2: *The plot on the left shows points on a two dimensional plane rolled up into a three dimensional space. The plot on the right shows the eigenfunctions of the laplacian in three dimensions and recovers the structure of the initial data set.*

We develop an understanding of the Laplacian Matrix by the problem of Diffusion. Diffusion is the by which a substance moves from a region of high to low density. It is driven by the relative pressure of different regions. One can also consider diffusion processes on networks.

We consider the case in which an object (with amount  $\psi_i$  at vertex  $i$ ) moves along edges to vertex  $j$  at a rate of  $C(\psi_i - \psi_j)$ , where  $C$  is known as the diffusion constant.

The rate at which  $\psi$  changes, can be given by:

$$\frac{d\psi_i}{dt} = C \sum_j A_{ij} (\psi_j - \psi_i) \quad (2)$$

The adjacency matrix ( $A_{ij}$ ) ensures that the only terms that appear in the sum are ones that correspond to vertex pairs connected by an edge. Splitting the terms in the sum we arrive at the following equation:

$$\frac{d\psi_i}{dt} = C \sum_j A_{ij} \psi_j - C \psi_i \sum_j A_{ij}$$

$$\frac{d\psi_i}{dt} = C \sum_j (A_{ij} - \delta_{ij} D_i) \psi_j$$

The matrix analogue of this equation is:

$$\frac{d\psi}{dt} = C(A - D)\psi \tag{3}$$

where  $\psi$  is the vector with components  $\psi_i$ ,  $A$  is the adjacency matrix, and  $D$  is the degree matrix. The laplacian is defined from this as  $L = D - A$ , which allows for the following:

$$\frac{d\psi}{dt} + CL\psi = 0 \tag{4}$$

$$\frac{d\psi}{dt} + C\nabla^2\psi = 0 \tag{5}$$

Note that this is the same equation as the diffusion equation for the continuous case (equation 6), where the laplacian matrix is analogous to  $\nabla^2$ .

The **normalized Laplacian** is defined as:

$$\widehat{L} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} \tag{6}$$

For pattern recognition, we are concerned with the eigenfunctions and eigenvalues of the combinatorial laplacian. We prove the mathematics of Laplacians using simplicial complexes. In mathematics, a simplicial complex is a topological space of a certain kind, constructed by "gluing together" points, line segments, triangles, and their n-dimensional counterparts; all simplices are oriented (except for vertices). If we let  $S$  be any finite simplicial complex, with simplices. Refer to appendix for proofs.

## 4.1 Dimensionality Reduction

We explore here the mathematics of dimensionality reduction. We argue, that the first three eigenfunctions of the combinatorial laplacian contain most of the information about the data, and hope that plotting them in three dimensions can reveal information about it as well. The adjacency matrix captures the entire structure of a network, however, the eigenvectors and eigenfunctions of the graph laplacian (which is built from the adjacency matrix) allow us to reduce this information into a lower dimension.

### 4.1.1 Historical motivation

So as an prototype to this problem, we look at the question, "can one hear the shape of a drum?". This was an article published by the mathematician Mark Kac in the 1960s.

Formally, the drum can be conceived as an elastic membrane whose boundaries are clamped. By using the continuous laplace operator to represent the

drums surface, Kac was able to deduce the frequencies at which the drumhead can vibrate. Essentially the allowed frequencies were given by the eigenvalues of the laplacian in the region. Since the laplacian is entirely dependent on the shape of this drum we can further say that the geometry is contained within its eigenfunctions or eigenvalues. In a similar manner, facial recognition algorithms frequently use the eigenvalues of the laplacian to create a finger print out of the image. Using the continuous laplacian operator as motivation we can use the discrete version of the laplace operator to preserve geometric structure within the eigenvalues or eigenfunctions.

#### 4.1.2 Properties of the Laplacian

The combinatorial laplacian is defined as:

$$L_{ij} = \begin{cases} \text{deg}(v_i) & \text{if } i = j \\ -1 & \text{if } i \neq j \text{ and } v_i \text{ is adjacent to } v_j \\ 0 & \text{otherwise} \end{cases}$$

The laplacian is related the the notion of smoothness as:

$$\begin{aligned} \sum \psi(x)L\psi(x) &= \sum \omega(x,y)(\psi(x) - \psi(y))^2 \\ &= \|\nabla\psi\|^2 \end{aligned}$$

#### 4.1.3 Positive Semi-definite

$$\begin{aligned} \lambda\|x\|^2 &= \langle \lambda x, x \rangle \\ &= \langle BB^T x, x \rangle \\ &= \langle B^T x, B^T x \rangle \\ &= \|B^T x\|^2 \geq 0 \end{aligned}$$

A hermitian matrix is said to be positive semidefinite if:

$$x^T M x \geq 0 \tag{7}$$

for all  $x$  in  $c^n$

For a positive semidefinite matrix, the following statements are all equivalent:

- All eigenvalues of  $A$  are nonnegative
- All the principal minors of  $A$  are nonnegative
- There exists  $B$  such that  $A = B^T B$

There are instances of positive semidefinite matrices which include symmetric dyads. For example, if  $A = uu^T$  for some vector  $u$  then:

$$q_A(x) = x^T B \tag{8}$$

#### 4.1.4 Laplacian Eigenfunctions

Mathematically an eigenfunction of an operator is a (non-zero) function that is returned from The eigenfunctions of the Laplacian can be viewed as an orthonormal basis of global Fourier smooth functions that can be used for approximating any value function on a graph [3] [2] [4] [17].

These basis functions capture large-scale features of the state space, and are particularly sensitive to bottlenecks, a phenomenon widely studied in Riemannian geometry and spectral graph theory. A potential drawback of Laplacian approximation is that it detects only global smoothness, and may poorly approximate a function which is not globally smooth but only piecewise smooth, or with different smoothness in different regions. These drawbacks are addressed

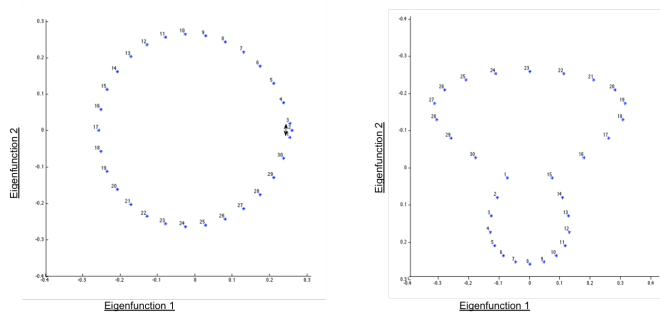


Figure 3: *Plotting the first two eigenfunctions for a graph in which each vertex is only connected to its nearest neighbors and loops around, reproduces a circle. By inducing further connections, the eigenfunctions reflect this change.*

in the context of analysis with diffusion wavelets, and in fact partly motivated their construction.

#### 4.1.5 Heat Kernel

The **Heat Kernel** is defined as:

$$e^{-tL} = F e^{-t\Lambda} F^T \quad (9)$$

and the **zeta function** as:

$$\zeta(s) = \sum_{k=2}^N \lambda_k^{-s} \quad (10)$$

## 5 Applications

### 5.1 Carcinogenicity of PolyAromatic Hydrocarbons

We are given a data set of 303 polyaromatic hydrocarbons with 20,000 corresponding properties, including their chemical carcinogenicity. Using PCA, we may be able to reduce the dimensionality of this data set to identify which

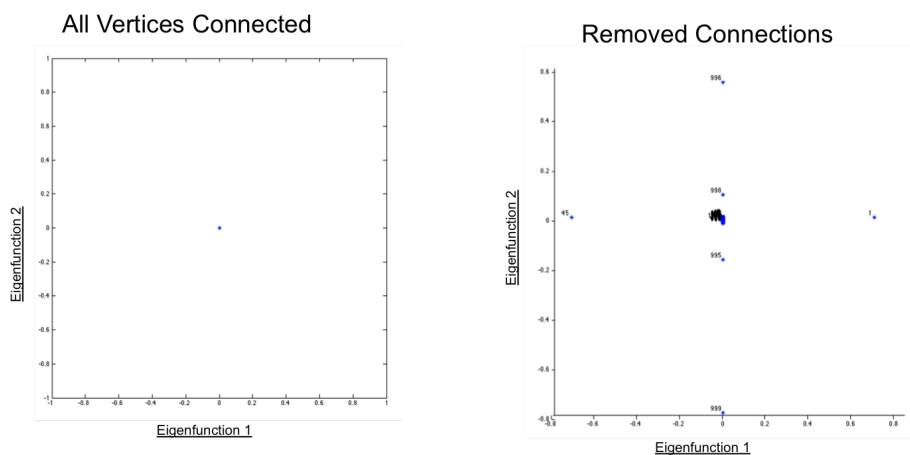


Figure 4: *By plotting the first two eigenfunctions of a complete graph, we see that all the points converge to one point. Removing connections calls out those points as outliers, showing that the eigenfunctions once again reflect this change in the intrinsic structure of the data set.*

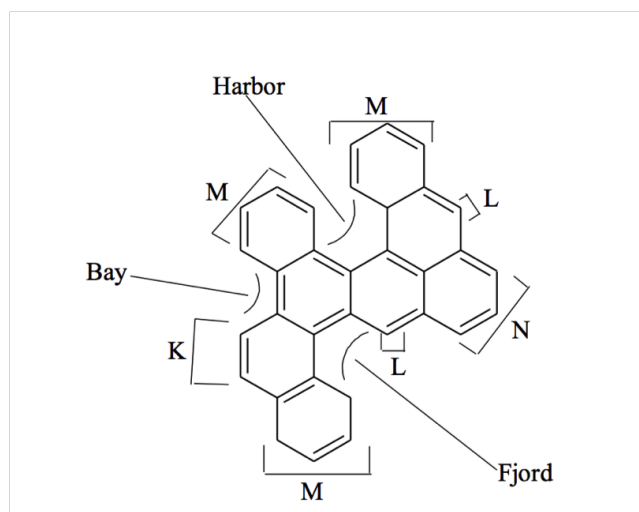


Figure 5: *Geometric Regions That are Thought to Contribute to Carcinogenicity.*



properties overlap (by creating orthogonal dimensions). However, this will not allow us to determine which properties contribute to the chemical carcinogenicity. Statistical analysis can individually determine which properties are most correlated with the carcinogenicity. However, graph theory allows this problem to be approached as a whole, and will allow for clusters of properties that contribute to the chemical carcinogenicity.

Several papers have shown that topological regions such as bays, fjords and harbors which you can see in this figure, have been implicated in the carcinogenicity of the hydrocarbons, however the evidence is not entirely conclusive.

A number of other properties such as the number of deactivating and activating regions were also implicated.

This analysis was conducted in a two step method. First the geometric structures were used to construct a laplacian matrix.

## 5.2 Geometric Connectivity

Initially the euclidean geometry was used to construct laplacian matrices out of each of the individual molecules. The hydrogens and carbons were both included. This created 303 laplacian matrices of different sizes corresponding to the molecule. Figure 6 shows that the first two eigenfunctions of the laplacian matrices contain information about the geometric structure, motivating their use in this application. We can see for example that the molecule naphthalene is very symmetrical in both, anthracene is a mirror image and that 3-methyl benz[a]anthracene is skewed. A graph similarity measure was then constructed to determine how similar two molecules were by getting the distances between their individual laplacian matrices. A minimum connectivity parameter was set that captured roughly 10 percent of the connections. The graph similarity

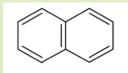
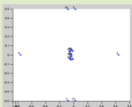
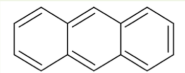
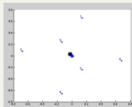
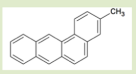
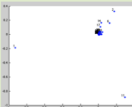
Molecule	Structure	Eigenfunctions of the Laplacian matrix (using connectivity)
Naphthalene		
Anthracene		
3-methyl benz[a]anthracene		

Figure 6: *Eigenfunctions of laplacian matrices made from the geometric connectivity of hydrogens and carbons in each of the individual atoms were plotted. Three are shown here to highlight how geometry is preserved in the eigenfunctions.*

measure used was the Frobenius Distance:

$$F_{A,B} = \sqrt{\text{trace}[(A - B)(A - B)^T]} \quad (11)$$

An adjacency matrix was conducted if two matrices corresponding to a molecule’s geometry were more similar than the minimum parameter. The first three eigenfunctions of this matrix are plotted in three dimensions in Figure 7.

### 5.3 Connectivity From Properties

The second step was to use the chemical properties to create this matrix and see if they converge with the geometries, to determine if the geometries reflect the chemical properties. To analyze the data set, PCA was used to reduce the dimensions to a smaller number of key uncorrelated properties. However, PCA showed that the two highest principle axes had an equal contribution from nearly all 1690 chemical properties. Thus, only 27 properties were considered. The 27

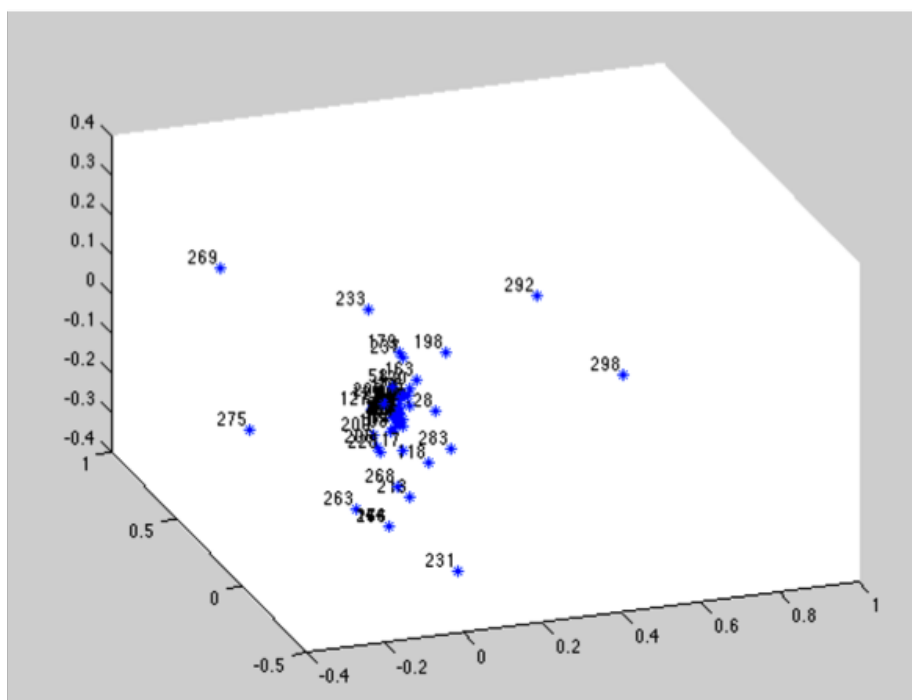


Figure 7: A plot of the first three eigenfunctions of the combinatorial Laplacian made from the geometric structures of the PAHs.

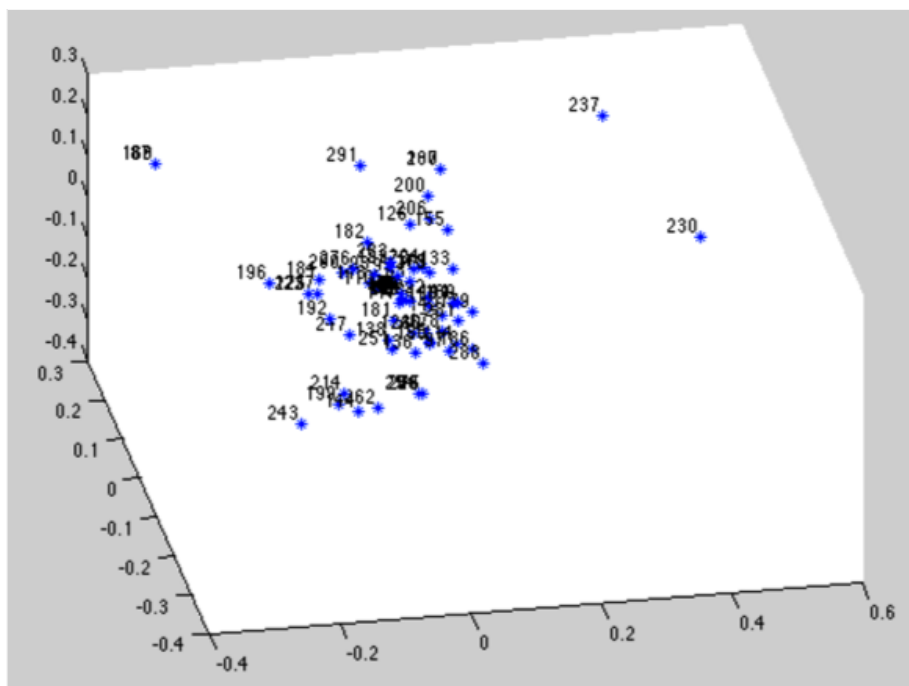


Figure 8: *A plot of the first three eigenfunctions of the combinatorial Laplacian made up of the chemical properties of the PAHs. This should converge with the previous graph, but does not appear to here.*

properties considered were ones that were implicated in the carcinogenicity of the molecules.

The distance measure used was based on first determining the normalized variation from the mean of each of the properties, and then getting a euclidean distance for each molecule in a 27 dimensional 'chemical property' space. After

## 5.4 Discussion

### 5.4.1 Carcinogenicities

We look into the mahalanobis distance as a euclidean analogue for a distance measure. It is used when, gaussian distributed, correlated variables have different scales and gives the relative measure of a points distance from a common point . The mahalanobis distance  $D_m$  between two nodes is defined as:

$$D_m(x, y) = \sqrt{(X_i - \mu)^T C^{-1} (X_i - \mu)} \quad (12)$$

Each molecule was assigned a carcinogenicity value from 1 to 9. The mahalanobis distance between each carcinogenicity value was calculated. The mahalanobis distance between each of the carcinogenicities was computed.

## 5.5 Economic Model of Mortgage

Mortgages are the largest asset that every bank has and most of Canadian banks have about 150 billion dollars invested in mortgages. A critical input to determination of price profitability, hedging instruments, hedge policy and funding policy is prepayment speed of mortgages.

Determination of prepayment speed is critical for measuring optimal funding strategy and pricing. For example, banks may borrowing at very high interest rates assuming that very few customers will prepay. However, if the prepayment speed is higher banks will be stuck with high cost.

An accurate prediction of prepayment rate can save every bank 100s of millions of dollars. However these prepayment rates depend on a variety of different factors, including the media, customer age and other demographics such as income

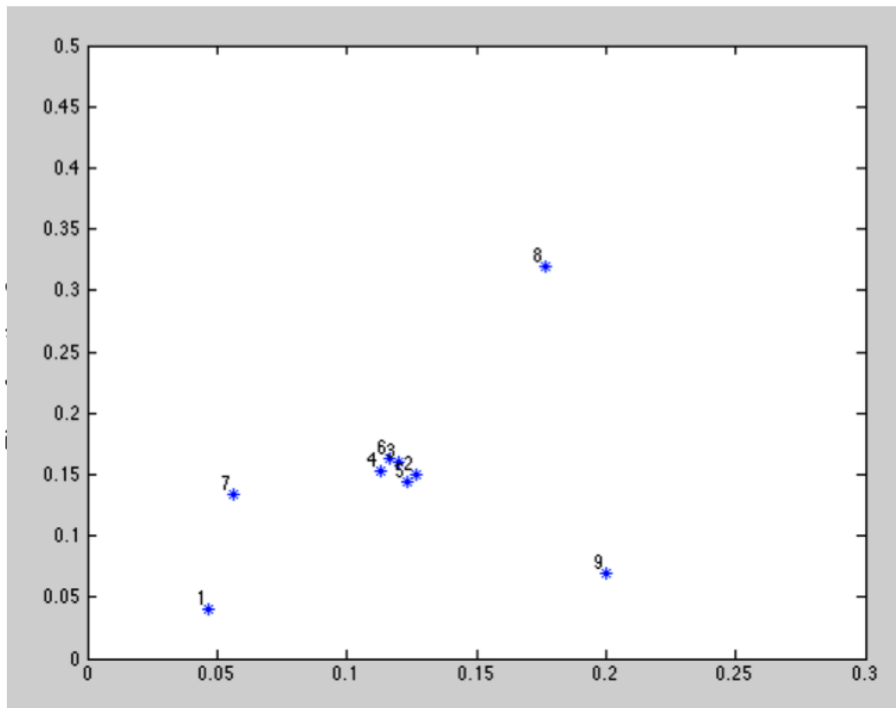


Figure 9: A plot of first two eigenfunctions of the combinatorial Laplacian constructed from the similarity of the carcinogenicity groups.

and financial stability. Each of the factors can be interpreted as another vertex in graph theory, allowing this to become a multidimensional graph-theory problem. Dimensionality reduction might allow us to visualize trends that are not otherwise obvious.

## 5.6 Protein Protein Interactions

Protein-protein interactions can be modelled using graph theory, where each protein is a vertex and the interactions between them are edges. Bu et al., 2003 tried to uncover topological features of protein protein interactions by isolating quasi-clique and quasi-bipartite groups. The analysis allowed uncharacterized proteins to have assigned functions based on the cliques that they fell in. Essentially they were able to use topological structures of the graph to biologically classify groups of proteins in terms of their possible functions.

In this application this involves a nearest neighbor search, and gives the ratio of existing links between nearest neighbors of a vertex and all possible links between nearest neighbors. Contact Order on the other hand is the average sequence distance between residues that form native contacts, over the total length of the protein. The graph properties that were looked at by

## 5.7 Protein Folding

From a chemical perspective proteins are linear heteropolymers and can draw on a combination of 20 different monomers units [13, 8]. However perhaps one of the most remarkable features of proteins is that despite the large degrees of freedoms, proteins consistently and quickly fold into their native states [7]. The three dimensional structure is critical in the role that the protein plays

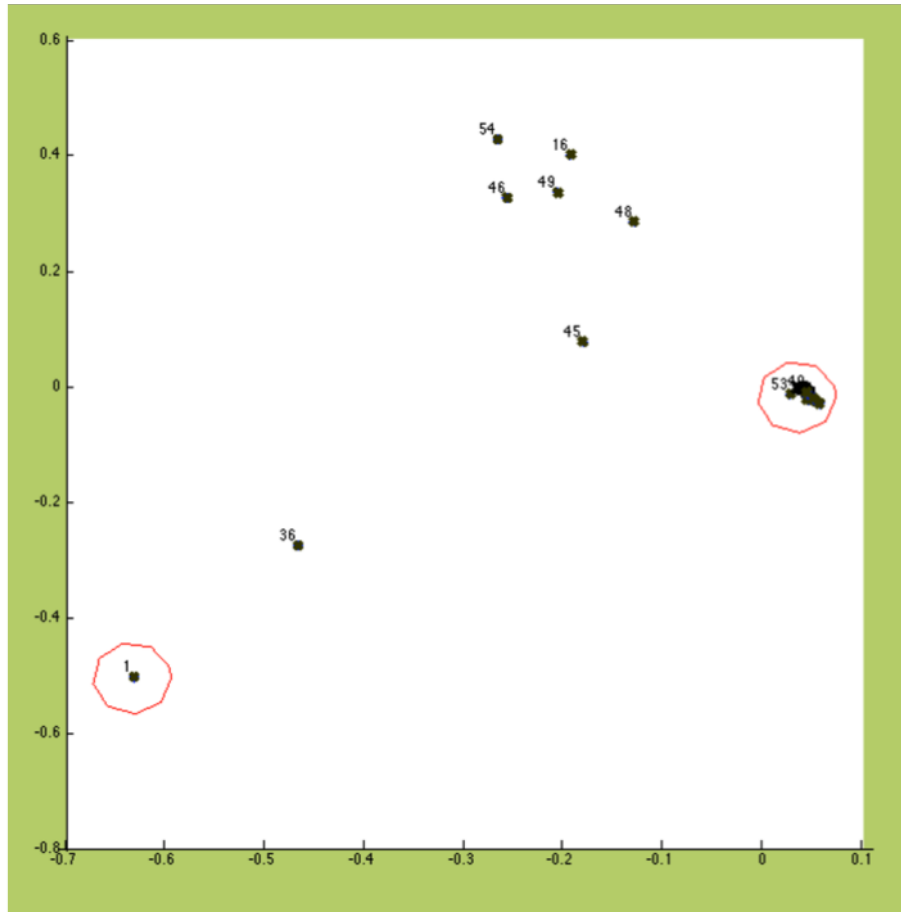


Figure 10: *The first two eigenfunctions of the graph made from protein protein interactions of carbonic anhydrase four. The left circle shows a cluster of several proteins that are similar in function. The outlier has identified a protein that is found in a different part of the body.*



in the body [14, 6]. In their native state protein sequences fold into unique three dimensional structures as a result of long range and short range interactions between base pairs [16, 4]. Although for the most part this structure can be determined by the nuclear magnetic resonance techniques (NMR) or x-ray crystallography, these are very expensive and time consuming processes [9] [15]. Thus, there is a significant drive in computational biology to develop a method that can predict three dimensional structures from a protein sequence.

The geometry of tertiary folding of proteins has a certain topology that can be drawn diagrammatically. These diagrams can be two dimensional schematic representations with symbols for helices and strands. The support for the ki-

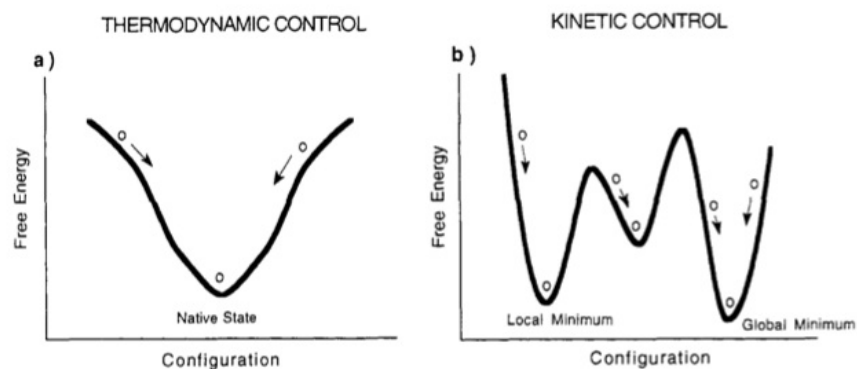


Figure 11: *This figure shows the possible gibbs free energies across a range of configurations for a protein ([12]).*

netic theory is vast. Looking at figure 2.1(b) we can see that if a protein has a number of local minimums in the protein folding pathway then the protein may not be able to be stuck in a state that is not the absolute energy minimum calculated by using the thermodynamic method.

A protein structure can be visualized as a network of side chain interactions, with a graph that can be constructed by considering only interacting residues. The beta carbon atoms of interacting residues may be considered vertices, with edges dependent on the strength of their interaction. Interactions between side chains can be taken into account when considering the graph. The purpose of using the laplacian matrix in this problem may be three fold. First of all, it could help detect hydrophobic residues.

## References

- [1] Edited By and Falk Schreiber. *Analysis of Biological Networks*. John Wiley & Sons, Inc., New Jersey, 2008.
- [2] Geoff Dougherty. *Pattern Recognition and Classification*. Springer Science, New York, 2013.
- [3] Joel Friedman. Some geometric aspects of graphs and their eigenfunctions. *Duke Math. J.*, 69(3):487–525, March 1993.
- [4] Yun Fu and Yunqian Ma, editors. *Graph Embedding for Pattern Analysis*. Springer Science, New York, 2013.
- [5] Timothy E Goldberg. Combinatorial Laplacians of Simplicial Complexes A Senior Project submitted to by. 2002.
- [6] Lesley H Greene. Protein structure networks. *Brief. Funct. Genomics*, 11(6):469–78, November 2012.
- [7] N Kannan and S Vishveshwara. Identification of side-chain clusters in protein structures by a graph spectral method. *J. Mol. Biol.*, 292(2):441–64, September 1999.

- [8] Arun Krishnan, Joseph P Zbilut, Masaru Tomita, and Alessandro Giuliani. Proteins as Networks: Usefulness of Graph Theory in Protein Science. *Bentham Sci. Publ.*, pages 28–38, 2008.
- [9] Hai-Yan Li. Folding rate prediction using complex network analysis for proteins with two- and three-state folding kinetics. *J. Biomed. Sci. Eng.*, 02(08):644–650, 2009.
- [10] Mark E. J. Newman. *Networks: An Introduction*. Oxford University Press, New York, 2010.
- [11] Mark E. J. Newman, Albert-Laszlo Barabasi, and Duncan Watts. *The Structure and Dynamics of Networks*. Princeton University Press, New Jersey, 2006.
- [12] Stefan Pinkert, Jörg Schultz, and Jörg Reichardt. Protein interaction networks—more than mere modules. *PLoS Comput. Biol.*, 6(1):e1000659, January 2010.
- [13] William R Taylor, Alex C W May, and Nigel P Brown. Protein structure : geometry , topology and classification. *Inst. Phys. Publ.*, 517(64), 2001.
- [14] Saraswathi Vishveshwara, K. V. Brinda, and N. Kannan. Protein Structure: Insights From Graph Theory. *J. Theor. Comput. Chem.*, 01(01):187–211, July 2002.
- [15] Liping Wei, Jeffrey Chang, and Russ Altman. *Computational Methods in Molecular Biology*. Elsevier, statistica edition, 1998.
- [16] Yan Yan, Shenggui Zhang, and Fang-Xiang Wu. Applications of graph theory in protein structure identification. *Proteome Sci.*, 9 Suppl 1(Suppl 1):S17, January 2011.

- [17] Jun Zhang, Partha Niyogi, and Mary Sara McPeck. Laplacian eigenfunctions learn population structure. *PLoS One*, 4(12):e7928, January 2009.