2013

Dot Product Graphs and Their Applications to Ecology

Sean Bailey
Utah State University

DOT PRODUCT GRAPHS AND

THEIR APPLICATIONS TO ECOLOGY

by

Sean Bailey

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Mathematics

Approved:

David Brown Major Professor	LeRoy Beasley Committee Member
David Koons Committee Member	Mark R. McLellan Vice President for Research Dean of the School of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2013

Abstract

Dot Product Graphs and

Their Applications to Ecology

by

Sean Bailey, Master of Science

Utah State University, 2013

Major Professor: Dr. David Brown
Department: Mathematics and Statistics

During the past few decades, examinations of social, biological, and communication networks have taken on increased attention. While numerous models of these networks have arisen, some have lacked visual representations. This is particularly true in ecology, where scientists have often been restricted to at most three dimensions when creating graphical representations of pattern and process. I will introduce an application of dot product representation graphs that allows scientists to view the high dimensional connections in ecological networks. Using actual data, example graphs will be developed and analyzed using key measures of graph theory.

(40 pages)

Public Abstract

Dot Product Graphs and

Their Applications to Ecology

by

Sean Bailey, Master of Science

Utah State University, 2013

Major Professor: Dr. David Brown
Department: Mathematics and Statistics

We will introduce a new tool to visualize the comparison between different birds. This tool will allow users to use any number of measurable traits to see relationships between different birds, both individually and collectively.

(40 pages)

Acknowledgements

I would like to acknowledge Dr Brown for his help and encouragement throughout my studies. He has been my Yoda in my understanding of graph theory. I would also like acknowledge Dr Beasley's guidance and suggestions throughout the process, especially as we have considered the linear algebra aspects. Dr Koons has been instrumental in his proposal of the initial question for this thesis as well his explanations of the various biological aspects. Finally, I need to acknowledge the encouragement and support of my wife, Monica. She has been my most important cheerleader in this process.

Sean Bailey

Contents

List of Figures

Chapter 1

Introduction

1.1 Graphical Representations

During the past few decades, examinations of social, biological, and communication networks have taken on increased attention. During these examinations, graphical representations of these networks and systems have proven to be very useful. Such representations can be used to determine or demonstrate the interconnections or relationships between elements of these networks. A myriad of mathematical tools also allow a system to be analyzed based on its graphical representation.

Yet troubles arise when we attempt to create graphical representations in k-dimensions with $k > 3$. This problem is especially true in biological representations. A method used by biologists is to restrict their representations to 2- or 3-dimensions [4] . This restriction requires that a biologist rank the criteria being used and use only a subset of the information available. In addition, 2-dimensional scatterplots are limited as to the quantity and quality of tools available for analysis.

In order to allow for quantified species data to be analyzed in both quantitative and qualitative ways, we have developed a method of creating a graphical representation for k-dimensions for any $k \in \mathbb{N}$. This method incorporates a type of simple, undirected graph. For clarity, when we refer to a graph throughout the remainder of this thesis, we will be referring to a simple, undirected graph.

1.2 Notation, Terminology, and Concepts

A simple, undirected graph is defined as an ordered pair of sets: a set of vertices and a set of edges. The set of *vertices* is $V = \{v_1, v_2, ..., v_n\}$, and a set of *edges*,$E$, consists of a subset of all non-ordered pairs of vertices. Further we define two vertices as *adjacent* if and only if they comprise an edge. The vertices which comprise an edge may be called

endpoints of the edge. Similarly we define an edge to be *incident* to a vertex if and only if the vertex is one of the endpoints of the edge. Thus if $v_iv_j \in E$, then v_i is adjacent to v_j and both vertices are incident to v_iv_j. For convenience, we denote such a graph simply as G.

G is a *simple graph* if no vertex is adjacent to itself and adjacent vertices have only one edge between them [1]. During this thesis, we will only consider simple, undirected graphs.

There are a number of parameters and structures useful in analyzing a graph, G. To introduce these characteristics, consider the following graph, G (Figure 1.1).

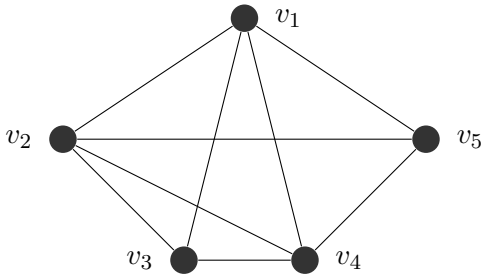


FIGURE 1.1: An Example of a Graph.

The set of vertices is $V = \{v_1, v_2, v_3, v_4, v_5\}$ and the set of edges is $E = \{v_1v_2, v_1v_3, v_1v_4, v_1v_5, v_2v_3, v_2v_4, v_2v_5, v_3v_4, v_4v_5\}$.

Another structure of this graph are *paths*. A *path* is a sequence of unique vertices each of which is adjacent to the preceding and following vertex in the sequence [1]. An example of a path from v_1 to v_5 in Figure 1.1 is $v_1v_4v_3v_2v_5$. It can also be seen that v_1v_5 is also a path from v_1 to v_5. We will define the *length of a path* as the number of edges in a path [1].

A path from a vertex back to itself is called *a cycle* [1]. An example of a cycle in Figure 1.1 is $v_1v_2v_3v_4v_5v_1$. This cycle has length 5. We will refer to a cycle of length k as a k-cycle. Additionally, we define a *chordless cycle* to be a cycle such that non-sequential vertices of the cycle are not adjacent in the graph G. An example of a chordless cycle in Figure 1.1 is $v_1v_2v_5$. Conversely, $v_1v_2v_3v_4v_5v_1$ is not chordless because the edge v_1v_4 exists.

A graph generated from another graph by deleting a set of vertices and the edges

incident to those vertices is called an *induced subgraph*. Some examples of induced subgraphs of Figure 1.1 are shown in Figure 1.2.

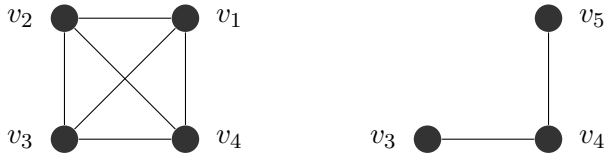


FIGURE 1.2: Induced Subgraphs of Figure 1.1.

1.3 Properties of Graphs

Graphs are used in modeling so that their structures can be analyzed in hopes of yielding insight into the source of our model. The structural parameters of a graph can be considered as either quantitative or qualitative, local or global.

An example of a qualitative parameter of a graph is *connectedness*. A graph is connected if and only if there exists a path between any two vertices. The graph in Figure 1.1 is connected. A connected graph with every vertex adjacent to every other vertex is called *complete*. In addition, an induced subgraph that is complete is called a *clique*. While not every graph is complete, every graph does contain at least one clique. An example of a clique in Figure 1.1 is the induced subgraph on the left in Figure 1.2. The number of vertices in the largest clique of a graph is a global parameter called *the clique number*, $\omega(G)$.

Complete bipartite graphs are graphs such that $V(G) = V_1 \cup V_2$ with $V_1 \cap V_2 = \emptyset$, $uw \notin E(G)$ if $u, w \in V_i$ for $i = 1, 2$, and $xy \in E(G)$ if $x \in V_1$ and $y \in V_2$. We will use the notation K_{n_1,n_2} for a complete bipartite graph, where n_1 and n_2 are the number of vertices in each subset, V_1, V_2, respectively. An example of $K_{3,2}$ is shown in Figure 1.3.

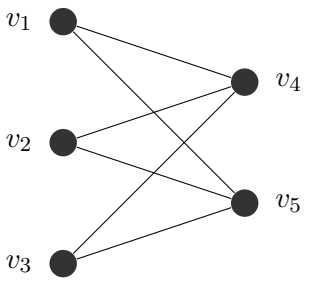


FIGURE 1.3: An Example Bipartite Graph.

A local parameter of a graph is the *degree of its vertices*. The degree of a vertex, written $deg(v_i)$, is the number of vertices adjacent to it. For example, $deg(v_1) = deg(v_2) = deg(v_4) = 4$ and $deg(v_3) = deg(v_5) = 3$. Specifically, we consider the maximum and minimum degrees of a graph. We will denote the maximum degree of a graph, G, by $\Delta(G)$ and the minimum degree by $\delta(G)$. Our example graph in Figure 1.1 has $\Delta(G) = 4$ and $\delta(G) = 3$.

The maximum degree of a graph is related to a global measure of a graph, *the chromatic number*. The chromatic number of a graph, written $\chi(G)$, is the minimal number of colors necessary to assign each vertex in graph G a color such that no two colors are adjacent. For our example graph in Figure 1.1, $\chi(G) = 4$. The possibility of this coloring can be seen in the following labeling of the graph (Figure 1.4).

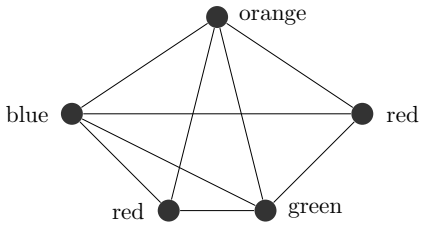


FIGURE 1.4: A Coloring of Figure 1.1.

The following theorems demonstrate the relations between some of these parameters.

Theorem 1: The chromatic number of a graph must be greater than or equal to the clique number of the graph.

Theorem 2 (Brooks' Theorem): If a graph G has a maximum vertex degree $\Delta(G)$, then the chromatic number of the graph $\chi(G) \leq \Delta(G)$, unless the graph is complete or an odd cycle. In these latter cases, $\chi(G) = \Delta(G) + 1$ [2].

All of these properties will be used to analyze the graphical representations of biological networks, and we will discuss the implications of the presence or absence of the qualitative parameters and the implications of the values of the quantitative parameters. We will give examples of biological networks which exhibit these properties and suggest interpretations consistent with the presence of these properties.

Chapter 2

Background

2.1 Dot Product Representations of Graphs

Our graphical representation will utilize a specific type of representation of a graph called a *dot product representation.* For a graph G, a *dot product representation* is a function $f : V \to \mathbb{R}^k$ along with a positive constant, t, where each vertex is assigned a vector, $f(u) = \{x_1, x_2, \cdots, x_k\}^T$, in $\mathbb{R}^k$ and $uv \in E(G)$ if and only if $f(u) \cdot f(v) \geq t$ [3].

An important proposition about dot product representations of graphs enables our use of them for any graph.

Proposition 1: Every graph can be represented as a dot product representation.

Proof: Let G be a graph with n vertices and k edges. We arbitrarily label each edge from 1 to k. We can define $f : V \to \mathbb{R}^k$ such that the i^{th} component of $f(u)$ is 1 if u is incident to the edge corresponding to i, with 0 otherwise. Thus $uv \in E(G)$ implies the i^{th} component of $f(v)$ and $f(u)$ is 1. So $f(v) \cdot f(u) \geq 1$. Similarly $uv \notin E(G)$ implies that there is no common component in $f(u)$ and $f(v)$ that is 1. So $f(v) \cdot f(u) = 0 < 1$ in this case. Thus we have a dot product representation of G.

2.1.1 Examples of Dot Product Representations of Graphs

To further understand dot product representations of graphs, consider the graph, H (Figure 2.1).

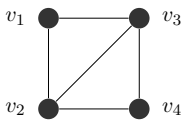


FIGURE 2.1: H: An Example of a Dot Product Representation of Graph.

We can turn this into a dot product representation by assigning each vertex a vector

such that the dot product of adjacent vertices is greater than or equal to 1 and the dot product of nonadjacent vertices is less than 1. Using the assignment designated in the proof of Proposition 1, we get the following assignment:

$$f(v_1) = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \; f(v_2) = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \end{pmatrix} \; f(v_3) = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{pmatrix} \; f(v_4) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix},$$

a dot product representation of dimension 5. But this assignment is not the only one that works. A brief examination of the dot products of the following vectors shows that the following assignment also produces the graph in Figure 2.1:

$$f(v_1) = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} \; f(v_2) = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \; f(v_3) = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \; f(v_4) = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}$$

a dot product representation of dimension 2.

2.1.2 Dot Product Dimension

This possibility of multiple dot product representationss corresponding to the same graph gives rise to another parameter: *the dot product dimension of a graph*. We define *the dot product dimension*, written $\rho(G)$, as the minimal dimension of the vectors necessary to create a dot product representation for G. It can be trivially seen from the proof of Proposition 1 that for every graph G, $\rho(G) \leq |E(G)|$ [3].

To understand this parameter more fully, let us consider $\rho(H)$ for the graph in Figure 2.1. We showed that we could represent each vertex with a vector of dimension 2. But could we use a vector with dimension 1? We can. Here is a possible function: $f(v_1) = \langle \frac{1}{2} \rangle, f(v_2) = \langle 2 \rangle, f(v_3) = \langle 2 \rangle, f(v_4) = \langle \frac{1}{2} \rangle$. Since these 1-dimensional vectors satisfy our representation, $\rho(H) = 1$ for our example.

While we were able to determine the dot product dimension of Figure 2.1, computing dot product dimension is a NP-hard problem [5]. This implies that the problem of finding the dot product dimension is as hard as a problem where the time is takes to find a solution using an algorithm increases quickly as the size of the problem grows. Still the dot product dimension of certain classes is known. For example, $\rho(K_n) = 1$ for all $n \in \mathbb{N}$. Also $\rho(K_{n_1,n_2}) = min\{n_1, n_2\}$ [3].

2.1.3 Applications of Dot Product Representations to Social Networks

A basic example of the application of dot product representations is with social networks. Suppose we have four individuals in a social network who are surveyed on their preference level towards dogs, ice cream, religion, and sports. We can give positive preferences positive values, apathetic preferences zero, and negative preferences negative values with stronger preferences receiving values with greater magnitudes. So the following table could represent the preferences of our four individuals:

Category	Individual A	Individual B	Individual C	Individual D
Dogs	1	-0.5	-1	-1
Ice Cream	0.5	1	0	1
Religion	-1	-1	0	1
Sports	1	0	1	-0.5

This table shows that Individuals A and C have similar preferences towards sports, but their preferences on the other categories differ. If we set our threshold value for dot products at 0.5, then these preference vectors generate graph in Figure 2.2.

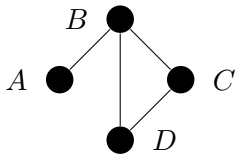


FIGURE 2.2: An Example of a Dot Product Representation of a Social Network.

Looking at this graph, we see that it suggests that B would be friends with everyone. On the other hand, A would have only one friend, B. So while A and C have similar preferences towards sports, they have opposing differences in the other categories that outweigh that similarity.

For this example, we used a threshold of $t = 0.5$. But t could have been any positive number. Therefore our representation would change as we either increase or decrease our threshold. We will further demonstrate how this relation changes in Chapter 3.

Research into the dot product representations of graphs of social networks has been of growing interest [14][12]. Significant work has been done in this area by Schienerman and Young (2009) [13]. This work suggests that biological networks would also be

modeled with dot product representations since biological networks and social networks exhibit similar properties.

2.2 Evolutionary Fitness Gradient

Since dot product representations require vectors, we decided to consider the evolutionary fitness gradients of various species of birds. The dimension of these gradients is 4, allowing us to represent and analyze high dimension models. These gradients are developed from the work of Robertson, Lande, and Leslie.

In 1966, Robertson introduced the secondary theorem of natural selection [11]. This theorem refers to the per generation evolutionary change in trait x as a function of the genetic heritability of trait x and the strength of natural selection of the trait. In agricultural circles, this theorem is known as the "breeder's equation." Heritability proportional to the additive genetic variance of trait x among the individuals of a population. An individual's *fitness* is perhaps best measured as its contribution to the population's Malthusian growth rate, r, which is the rate at which one generation propagates the next generation of a population. The secondary theorem of natural selection states that $\frac{d\bar{x}}{dt} = \sigma^2 \frac{dr}{d\bar{x}}$, where the σ^2 is the genetic heritability and $\frac{dr}{d\bar{x}}$ is the strength of the selection.

The secondary theorem of natural selection was extended to the multivariate case by Lande [7][8]. This extension allows for the examination of the evolutionary change of multiple traits such as brain mass, body mass, and fertility. Suppose these traits are quantified and denoted as $x_1, x_2, ..., x_n$. We first let $\bar{\mathbf{x}} = \langle \bar{x_1}, \bar{x_2}, \bar{x_3}, ..., \bar{x_n} \rangle^T$ and $\nabla r = \langle \frac{\partial r}{d\bar{x_1}}, \frac{\partial r}{d\bar{x_2}}, ..., \frac{\partial r}{d\bar{x_n}} \rangle^T$. Then Lande showed that $\Delta\bar{\mathbf{x}} = \mathbf{G}\nabla r$, where $\mathbf{G}$ is the matrix of additive genetic covariance amongst traits.

The gradient ∇r is known as the *fitness gradient* when considering life history traits. In evolutionary biology, the fitness of an individual organism is defined as the probability that the organism will survive and reproduce. Some of the life history traits that affect fitness include juvenile survival and fertility.

In our application, we will consider the simple 2×2 matrix model developed by Leslie [9][10]. This matrix, called the *Leslie matrix*, helps model fitness in terms of juvenile survival, the onset of first reproduction, fertility, and adult survival. We will define σ_1 as the probability of juvenile survival, γ as the probability of becoming an adult and beginning reproduction, ϕ as per time step fertility, and σ_2 is the probability of adult survival. The Leslie matrix is

$$A = \begin{bmatrix} \sigma_1(1-\gamma) & \phi \\ \sigma_1\gamma & \sigma_2 \end{bmatrix}$$

Let us consider the blue tit, a small European bird, to look at an example of a Leslie matrix. The blue tit has a probability of juvenile survival of $\sigma_1 = 0.16$, a probability of adult survival of $\sigma_2 = 0.40$, a probability of becoming an adult and beginning reproduction of $\gamma = 1$, and a per time step fertility of $\phi = 4.23$. Thus the Leslie matrix for the blue tit is:

$$A = \begin{bmatrix} 0 & 4.23 \\ 0.16 & 0.40 \end{bmatrix}$$

Using this matrix, Leslie showed that $ln(\lambda_1) = r$, where λ_1 in the eigenvalue of A with the largest magnitude. Thus the fitness gradient of a species associated with the Leslie matrix is $\nabla r = \left\langle \frac{\partial ln(\lambda_1)}{d\sigma_1}, \frac{\partial ln(\lambda_1)}{d\gamma}, \frac{\partial ln(\lambda_1)}{d\phi}, \frac{\partial ln(\lambda_1)}{d\sigma_2} \right\rangle^T$.

To determine this gradient, we first find the eigenvalues of A by using the characteristic equations of A, $det(A - \lambda I) = 0$. By definition of the determinant of a 2×2 matrix, $(\sigma_1 - \sigma_1\gamma - \lambda)(\sigma_2 - \lambda) - \phi\sigma_1\gamma = 0$. This is a quadratic equation so we can use the quadratic formula to find λ:

$$\lambda = \frac{\sigma_1 - \gamma\sigma_1 + \sigma_2 \pm \sqrt{\sigma_1^2 - 2\gamma\sigma_1^2 + \gamma^2\sigma_1^2 + 2\sigma_1\sigma_2 - 2\gamma\sigma_1\sigma_2 + \sigma_2^2 - 4\phi\gamma\sigma_1}}{2}.$$

Both eigenvalues of A are positive. Thus

$$\lambda_1 = \frac{\sigma_1 - \gamma\sigma_1 + \sigma_2 + \sqrt{\sigma_1^2 - 2\gamma\sigma_1^2 + \gamma^2\sigma_1^2 + 2\sigma_1\sigma_2 - 2\gamma\sigma_1\sigma_2 + \sigma_2^2 - 4\phi\gamma\sigma_1}}{2}.$$

We can then take the partial derivative of $ln(\lambda_1)$ to get our fitness gradient:

$$\nabla r = \begin{pmatrix} \dfrac{\frac{1}{2} - \frac{1}{2}\gamma + \frac{2\sigma_1 - 4\sigma_1\gamma - 2\sigma_2 + 2\sigma_1\gamma^2 + 2\gamma\sigma_2 + 4\phi\sigma_2}{4\sqrt{\sigma^2 - 2\sigma^2\gamma - 2\sigma_1\sigma_2 + +\sigma_1^2\gamma^2 + 2\sigma_1\gamma\sigma_2 + \sigma_2^2 + 4\phi\sigma_1\gamma}}}{\frac{1}{2}\sigma_1 - \frac{1}{2}\sigma_1\gamma + \frac{1}{2}\sigma_2 + \frac{1}{2}\sqrt{\sigma^2 - 2\sigma^2\gamma - 2\sigma_1\sigma_2 + \sigma_1^2\gamma^2 + 2\sigma_1\gamma\sigma_2 + \sigma_2^2 + 4\phi\sigma_1\gamma}} \\ \\ \dfrac{-\frac{1}{2}\sigma_1 + \frac{-2\sigma_1^2 + 2\sigma_1^2\gamma + 2\sigma_1\sigma_2 + 4\phi\sigma_1}{4\sqrt{\sigma^2 - 2\sigma^2\gamma - 2\sigma_1\sigma_2 + +\sigma_1^2\gamma^2 + 2\sigma_1\gamma\sigma_2 + \sigma_2^2 + 4\phi\sigma_1\gamma}}}{\frac{1}{2}\sigma_1 - \frac{1}{2}\sigma_1\gamma + \frac{1}{2}\sigma_2 + \frac{1}{2}\sqrt{\sigma^2 - 2\sigma^2\gamma - 2\sigma_1\sigma_2 + \sigma_1^2\gamma^2 + 2\sigma_1\gamma\sigma_2 + \sigma_2^2 + 4\phi\sigma_1\gamma}} \\ \\ \dfrac{\frac{\sigma_1\gamma}{\sqrt{\sigma^2 - 2\sigma^2\gamma - 2\sigma_1\sigma_2 + +\sigma_1^2\gamma^2 + 2\sigma_1\gamma\sigma_2 + \sigma_2^2 + 4\phi\sigma_1\gamma}}}{\frac{1}{2}\sigma_1 - \frac{1}{2}\sigma_1\gamma + \frac{1}{2}\sigma_2 + \frac{1}{2}\sqrt{\sigma^2 - 2\sigma^2\gamma - 2\sigma_1\sigma_2 + \sigma_1^2\gamma^2 + 2\sigma_1\gamma\sigma_2 + \sigma_2^2 + 4\phi\sigma_1\gamma}} \\ \\ \dfrac{\frac{1}{2} + \frac{-2\sigma_1 + 2\sigma_1\gamma + 2\sigma_2}{4\sqrt{\sigma^2 - 2\sigma^2\gamma - 2\sigma_1\sigma_2 + +\sigma_1^2\gamma^2 + 2\sigma_1\gamma\sigma_2 + \sigma_2^2 + 4\phi\sigma_1\gamma}}}{\frac{1}{2}\sigma_1 - \frac{1}{2}\sigma_1\gamma + \frac{1}{2}\sigma_2 + \frac{1}{2}\sqrt{\sigma^2 - 2\sigma^2\gamma - 2\sigma_1\sigma_2 + \sigma_1^2\gamma^2 + 2\sigma_1\gamma\sigma_2 + \sigma_2^2 + 4\phi\sigma_1\gamma}} \end{pmatrix}$$

Critical aspects of the fitness gradient as far as our model is concerned are:

1. The orientation of the gradient vector, which are the angels generated between the vector and the primary axis. The orientation determines which variable of fitness affects the fitness of the species the most.

2. The length of the vector, which we will take to be the Euclidean norm of the gradient. This norm is found for any vector $\bar{v} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix}$ by $||\bar{v}|| = \sqrt{x_1^2 + x_2^2 + \cdots + x_k^2}$.

   This norm of the gradient is the rate at which the fitness of a species is increasing as the various variables increase.

These fitness gradients will be used to create dot product representations of dimension 4. It is these representations that we will analyze in our application.

Chapter 3

Applications of Dot Product Graphs

Recall that the dot product of vectors $\mathbf{u}$ and $\mathbf{v}$ is $\mathbf{u} \cdot \mathbf{v} = ||\mathbf{u}||||\mathbf{v}||cos\theta$, where θ is the angle between the two vectors when originating at the same point and $||\mathbf{u}||$ is the Euclidean length (or 2-norm) of vector $\mathbf{u}$. So we may regard $\mathbf{u} \cdot \mathbf{v} \geq t$ as a measure of the angle between them. We will be using angles between vectors to indicate similarity between corresponding species, like we used $\mathbf{u} \cdot \mathbf{v} \geq t$ to indicate possible compatibility in the social network example.

3.1 Applications to Evolutionary Ecology

Using the fitness gradients mentioned in Chapter 2 and demographic data available in literature, we were able to assign vectors to each species of a set of 10 long-lived seabirds in order to provide an evolutionary ecology example of dot product graphs. To allow us to examine the angles between vectors, we decided to normalize them using the Euclidean norm. Thus, if the vectors are normalized (reduced to length 1), we can redefine connectedness of vertices if and only if the angle between the vectors, θ, is less than or equal to a given angle.

To better understand how this normalization affects vectors and their dot products, consider the vectors $\mathbf{u} = \begin{pmatrix} 5 \\ 5 \end{pmatrix}$ and $\mathbf{v} = \begin{pmatrix} \frac{1}{6} \\ \frac{1}{6} \end{pmatrix}$. It can be noted that $\mathbf{u} \cdot \mathbf{v} = \frac{5\sqrt{2}}{6} < 1$. But $||\mathbf{u}|| = 5\sqrt{2}$ and $||\mathbf{v}|| = \frac{\sqrt{2}}{6}$. The normalized version of the vectors then are $\hat{\mathbf{u}} = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix}$ and $\hat{\mathbf{v}} = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix}$. This implies that $\hat{\mathbf{u}} \cdot \hat{\mathbf{v}} = 1 > \mathbf{u} \cdot \mathbf{v}$. Similarly, it is possible to generate to vectors such that their dot product is more than the dot product of their normalized versions.

This difference between normalized and non-normalized vectors led us to consider the graphical representations generated when the vectors were not normalized. Thus

we can compare the two different models and their implications. In the next chapter we will discuss similarities and differences between models using normalized versus non-normalized vectors.

3.2 Examples of Normalized Graphs

FIgures 3.1-3.7 are a series of graphical representations based on the normalized vectors as the threshold, θ, decreases. We will denote $G_{\theta=\alpha}$ to be the dot product graph where vertices u and v are adjacent if and only if the angle between $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ is less than or equal to α (in radians).

For this sequence of graphs, we begin with $\theta = 0.2$. We then incrementally decrease θ by 0.02 until our final graph, $G_{\theta=0.08}$, only has 2 adjacent vertices. This sequence was chosen as it shows a general sequence of reduced connections, as well as a sequence of isolated species. We will discuss these sequences and implications in Chapter 4.

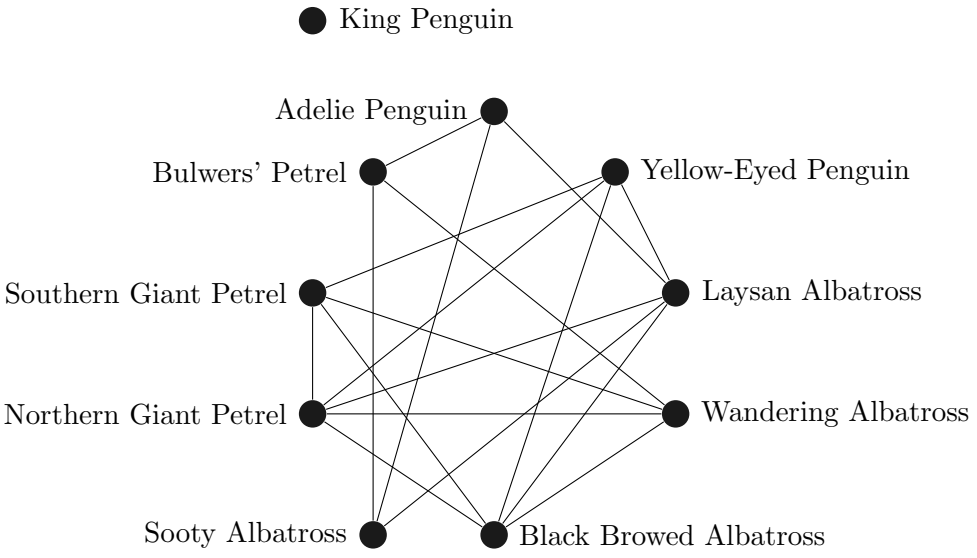


FIGURE 3.1: $G_{\theta=0.2}$.

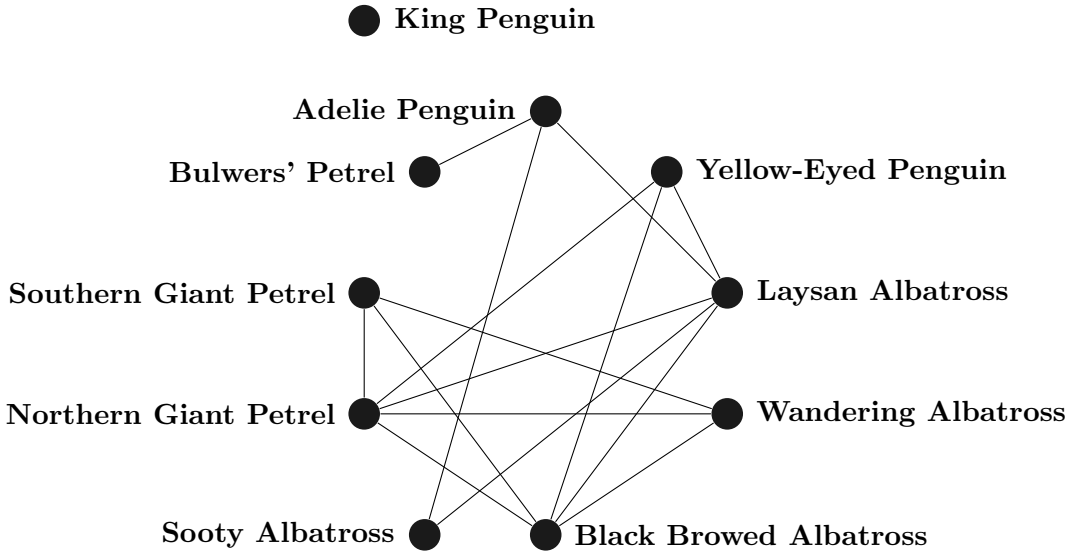


FIGURE 3.2: $G_{\theta=0.18}$.

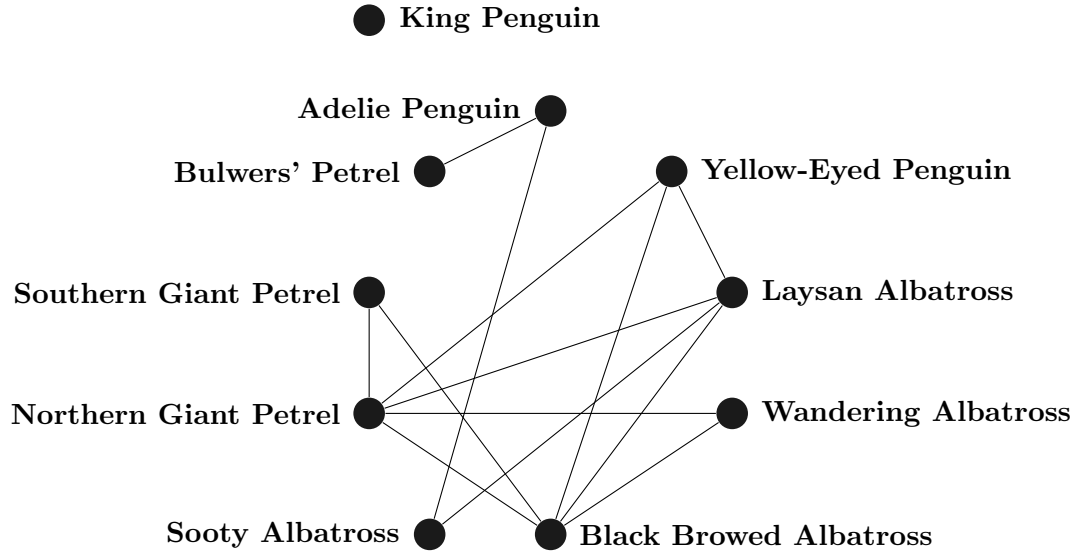


FIGURE 3.3: $G_{\theta=0.16}$.

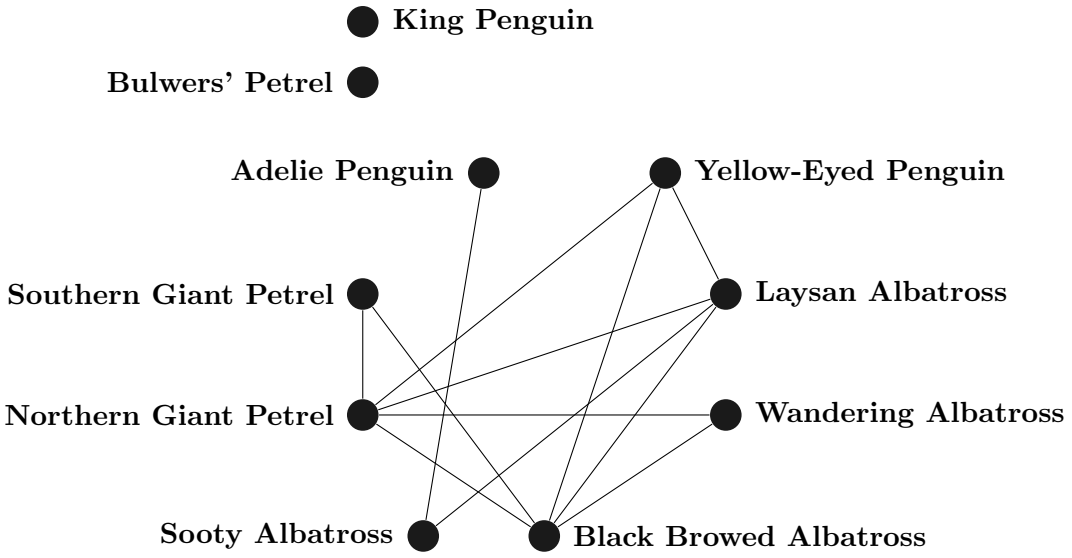


FIGURE 3.4: $G_{\theta=0.14}$.

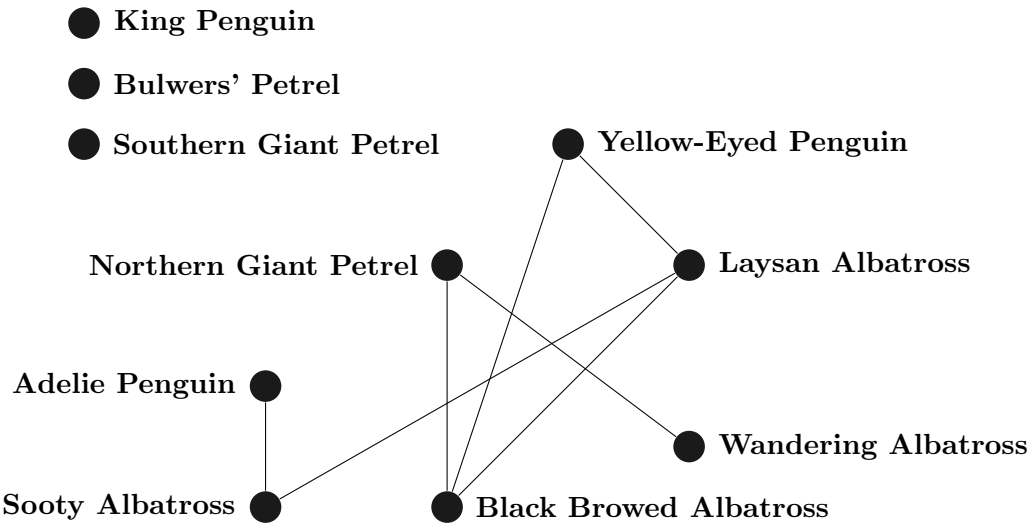


FIGURE 3.5: $G_{\theta=0.12}$.

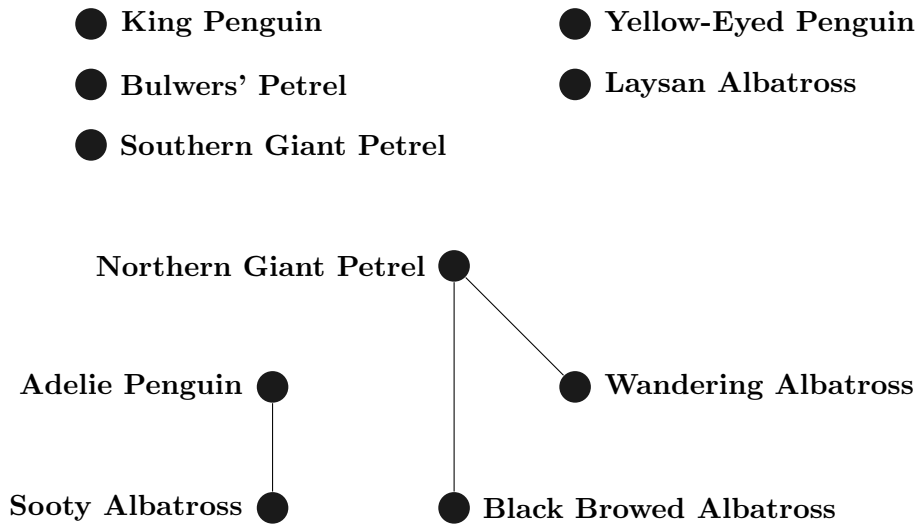


FIGURE 3.6: $G_{\theta=0.1}$.

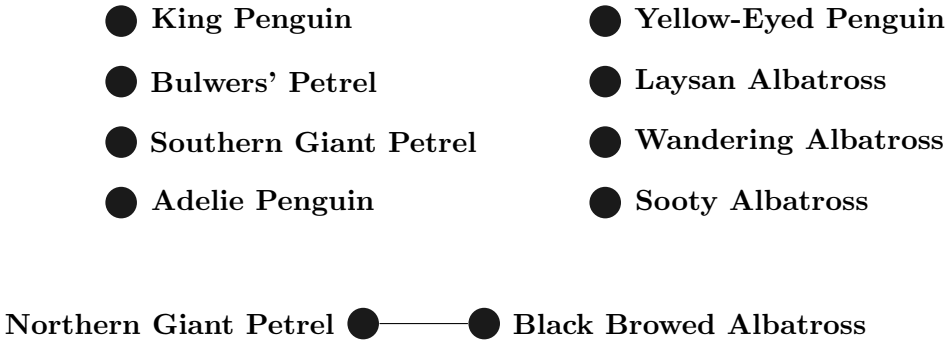


FIGURE 3.7: $G_{\theta=0.08}$.

3.3 Example of Non-normalized Graphs

Figures 3.8-3.13 are a sequence of graphical representations based on the non-normalized vectors. We will denote G_t to be the dot product graph where vertices u and v are adjacent if and only if the dot product of $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ is greater than or equal to t.

Our sequence will begin with $G_{t=0.85}$, which will provide a connected graph that is almost complete. For each graph in the sequence, t will incrementally increase by 0.05 from the preceding graph, except for the final graph. For the final graph, $G_{t=1.07}$, the t increased by on 0.02 as a step of 0.05 would have produced an empty graph.

As with the normalized graph, this sequence of graphs shows a sequence of isolated species as well as a sequence of connected species. We will discuss those sequences and implications in Chapter 4.

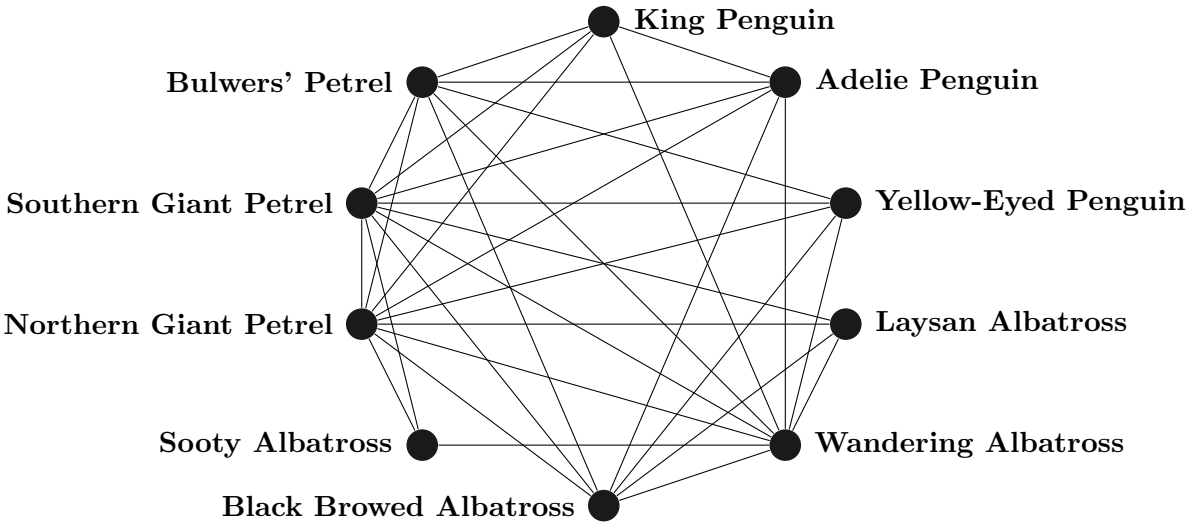


FIGURE 3.8: $G_{t=0.85}$.

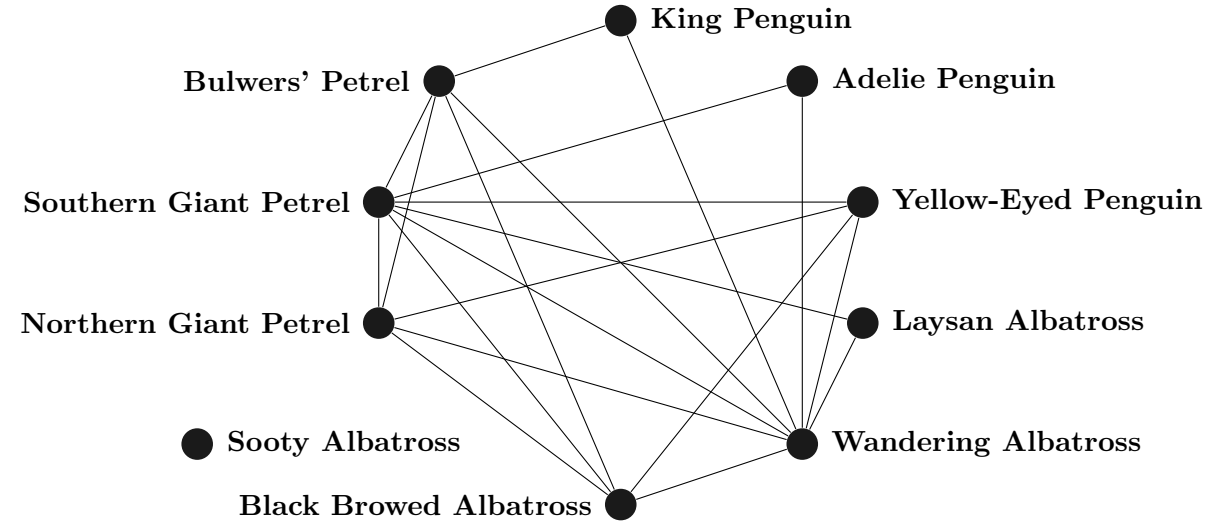


FIGURE 3.9: $G_{t=0.9}$.

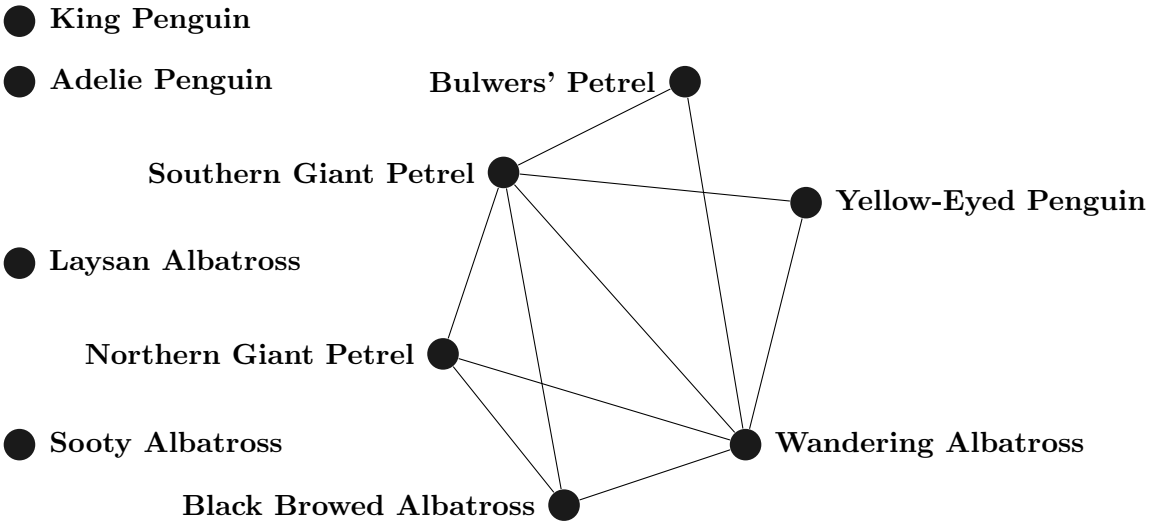


FIGURE 3.10: $G_{t=0.95}$.

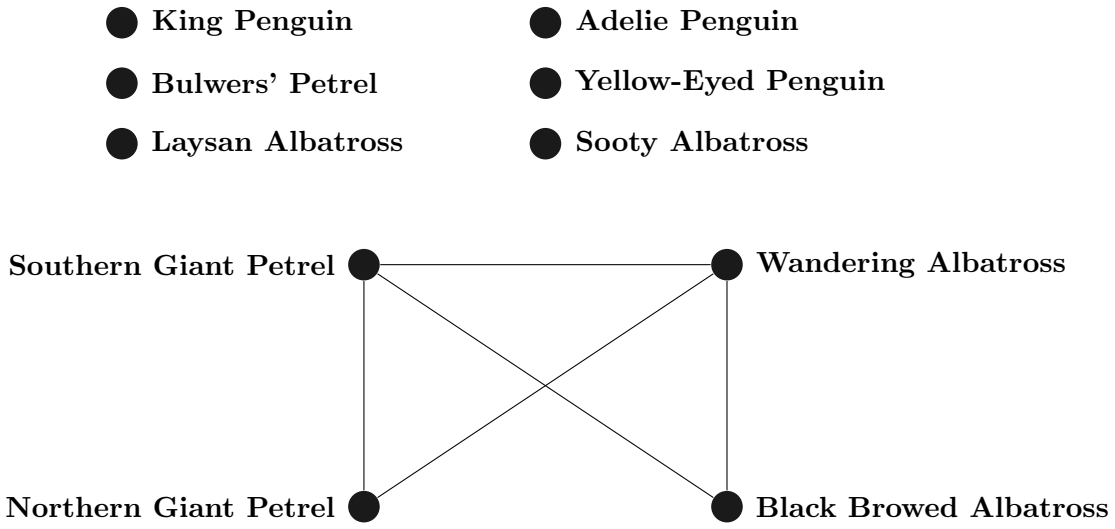


FIGURE 3.11: $G_{t=1}$.

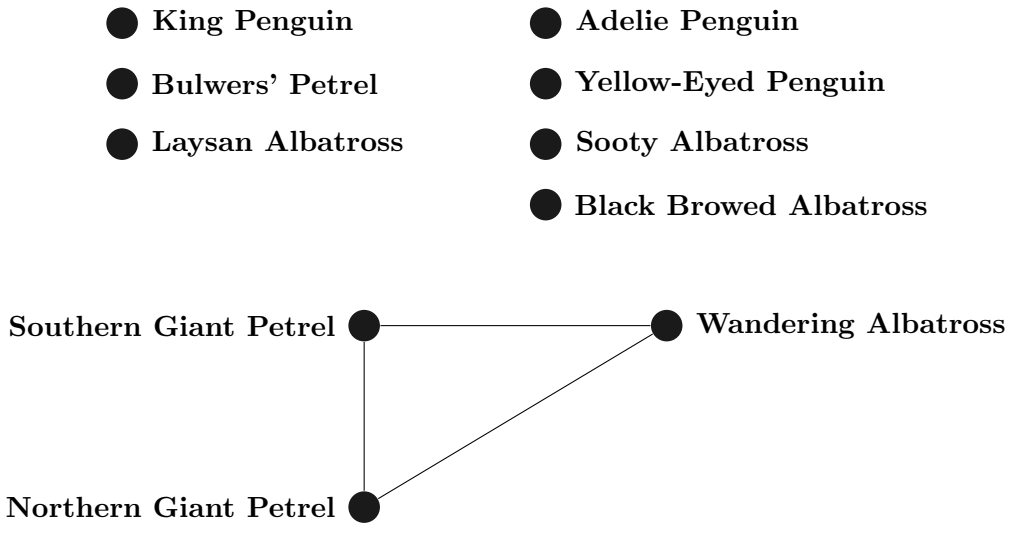


FIGURE 3.12: $G_{t=1.05}$.

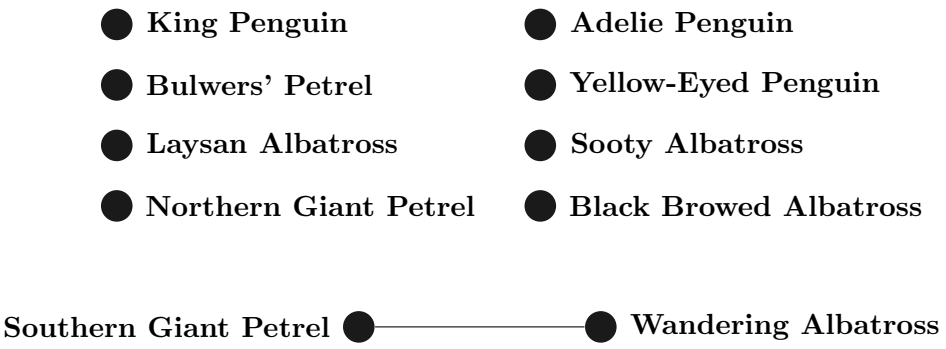


FIGURE 3.13: $G_{t=1.07}$.

Figures 3.1-3.13 show that as t increases and θ decreases, the number of edges decreases. In the next chapter, we will discuss our analysis of these graphs using structural properties of graphs mentioned in the Introduction.

Chapter 4

Analysis of Application of Dot Product Graph Models

4.1 Analysis of the Normalized Vector Models

The first thing one might notice upon examining $G_{\theta=0.2}$ is that our graph is connected except for the King penguin. The fact that this is the first species to be isolated suggests that the fitness gradient of this species is significantly different than the other species in this model. This might strongly suggest that the life cycle of the King penguin is different than all the other species.

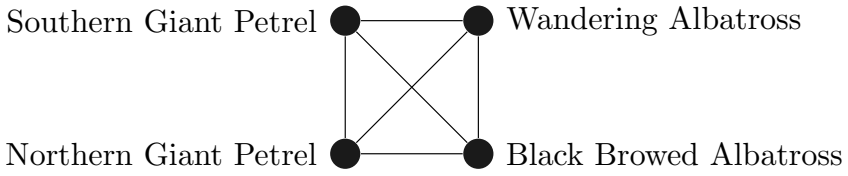
It can be further noted that even within the connected portion of $G_{\theta=0.2}$ that the other two species of penguins, the Adelie penguin and the yellow-eyed penguin, are not connected despite the fact that they are closer phylogenetic relatives than they are to procellerariformes (albatrosses and shearwaters). Similarly some of the other species that might be closely related are not connected within our sequence of graphs.

As we further examine the connections among $G_{\theta=0.2}$ through $G_{\theta=0.08}$, it can be noticed that in $G_{\theta=0.1}$ we have an instance of a disconnected graph with two components. This graph is of importances as it suggests that the life cycles of the Adelie penguin and the sooty albatross are more similar to one another than any other species. This discovery was intriguing as both birds might seem dissimilar to one another in terms of their taxonomical relatedness, but are highly connected with respect to their life histories. Understanding the determinants of tight connections like these could help us better understand life history evolutions. This understanding might even help guide conservation biology in terms of basic knowledge of the ecological determinants of life history evolution.

Consider now the vertex degrees in our sequence of graphs. In $G_{\theta=0.2}$, the northern giant petrel and the Laysan albatross have the maximum vertex degrees with 5. But the Laysan albatross was isolated before the Adelie penguin, which was among those vertices with the minimum degree of 3. On the other hand, it should be noticed that

$deg(NorthernGiantPetrel) = 5$ through $G_{\theta=0.14}$. Thus the northern giant petrel is a species that has more of an average of life cycles in comparison to the other species.

The northern giant petrel is also part of both of the maximal cliques in $G_{\theta=0.2}$. There are actually 2 cliques of size 4 in $G_{\theta=0.2}$ (Figure 4.1)



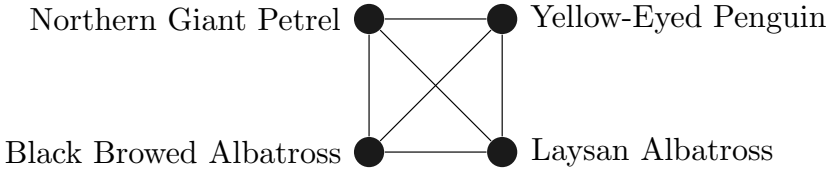


FIGURE 4.1: Cliques of Size 4 in $G_{\theta=0.2}$.

This latter clique (with the Laysan albatross) remains intact through $G_{\theta=0.14}$ while the former only through $G_{\theta=0.18}$. Thus there seems to be a strong life history connection between the four species of this latter clique (yellow-eyed penguin, Laysan albatross, northern giant petrel, and black browed albatross). This is further re-enforced by the fact that the only adjacent pair of species in $G_{\theta=0.08}$ are from this clique.

Since $\omega(G_{\theta=0.02}) = 4$, $\chi(G_{\theta=0.02}) \geq 4$. A four coloring of $G_{\theta=0.2}$ can be found by coloring the four vertices of one of the maximal cliques first. That coloring can be extended to the remainder of the graph by a process of elimination. Thus $\chi(G_{\theta=0.02}) = 4$. Since $\omega(G_{\theta=0.2}) = \omega(G_{\theta=0.18}) = \omega(G_{\theta=0.16}) = \omega(G_{\theta=0.14})$, the chromatic numbers of these graphs are also the same. One possible implication of this property of the graphs is that the species in $G_{\theta=0.2}$ can be divided into four distinct subsets such that no pair of species in the same subset have similar life histories.

Additionally, it can be seen that no chordless cycles of length greater than 3 or

complete bipartite graphs on subsets of size greater than 2 are contained in these normalized vector representations. The lack of complete bipartite graphs with $n_1, n_2 \geq 2$ in the normalized models can be proven to be a limitation of dot product graphs on normalized vectors. Consider the following lemma:

Lemma 1: *If G is a dot product graph with $V(G)$ represented by vectors of unit length, then K_{n_1,n_2}, where $n_1, n_2 \geq 2$, cannot be an induced subgraph.*

Proof: Initially, it can be noted that the property of being a dot product representation is hereditary. This means that any induced subgraph of G must be a dot product graph if G is. This heredity for dot product graphs was proven [3]. Thus we only need to consider the case where $G = K_{n_1,n_2}$ with $n_1 = n_2 = 2$

We will prove this by contradiction. Suppose that G can be represented by vectors of unit length. Let U and V be the bipartite sets of $V(G)$. Let $\hat{U} = \{\mathbf{u_1}, \mathbf{u_2}\}$ and $\hat{V} = \{\mathbf{v_1}, \mathbf{v_2}\}$ be the set of unit vectors associated U and V, respectively. So given some $t > 0$, $\mathbf{u_i v_j} > t$, $\mathbf{u_i u_j} < t$, and $\mathbf{v_i v_j} < t$ for any $\mathbf{u_i}, \mathbf{u_j} \in \hat{U}$ and $\mathbf{v_i}, \mathbf{v_j} \in \hat{V}$. But since the vectors are normalized, our threshold can also be seen as an angle θ where the angle between adjacent vectors is less that θ and the angle between nonadjacent vectors is greater that θ. Additionally, all of the vectors project onto a n-dimensional sphere. So the set of vectors within θ of a given vector $\mathbf{w}$ would be contained within an ellipse centered around $\mathbf{w}$. Thus $\mathbf{v_1}$ would to be contained within the ellipse centered around $\mathbf{v_2}$ and vice verse. But $\mathbf{u_1}$ would be contained in both ellipses centered around $\mathbf{v_1}$ and $\mathbf{v_2}$. $\mathbf{u_2}$ would also need to be contained on those two ellipses. But the intersection of the ellipse centered around $\mathbf{v_1}$ and $\mathbf{v_2}$ would be contained within the the ellipse centered $\mathbf{u_1}$. Thus $\mathbf{u_2}$ would be contained within the ellipse centered around $\mathbf{u_1}$, which would imply that the angle between $\mathbf{u_1}$ and $\mathbf{u_2}$ is less than θ which is a contradiction.

Thus $K_{2,2}$ cannot be generated by vectors of unit length. Thus it cannot be an induced subgraph of graph generated by vectors of unit length.

Unfortunately, the characterization of the classes of graphs in general that can be generated as dot product graphs on a set of normalized vectors is not known. Thus the validity of using the normalized models is uncertain. But there do appear to be some implications generated by this model that can be examined further form the biological model.

4.2 Analysis of the Non-Normalized Vector Models

Unlike the normalized representations, the first isolated species in the non-normalized model is the sooty albatross. This may be in part because of the length of the fitness gradient for the sooty albatross. It has the second smallest norm of the 10 vectors considered. The impact though of the norm of the vectors is seen as the first three isolated species (sooty albatross, Laysan albatross, and Adelie penguin) are those with the three smallest norms. Conversely, the final pair of adjacent vertices in $G_{t=1.07}$ (wandering albatross and southern giant petrel) are the vectors with the greatest norms of those considered.

Those vertices with largest norms are also those with the maximum vertex degree in $G_{t=0.85}$, namely wandering albatross, northern giant petrel, and southern giant petrel. All of these vertices have degree 9. It is also interesting to note that the first isolated vertex, Sooty albatross, is also the vertex with minimum vertex degree.

This trend continues in $G_{t=0.9}$ where the maximum degree vertices are again the two longest vectors. Similarly the three shortest vectors not isolated in $G_{t=0.9}$ correspond to the vertices with the minimum vertex degree.

We can prove this correlation based on the following lemma.

Lemma 2: *Let* $\mathbf{u}, \mathbf{v}$ *be vectors in* $\mathbb{R}^n$ *whose components are strictly positive. Then for any* $t > 0$*, if* $\mathbf{u} \cdot \mathbf{v} \geq t$ *then* $||\mathbf{u}||||\mathbf{v}|| \geq t$.

Proof: Let $t > 0$ be given. Suppose that $\mathbf{u} \cdot \mathbf{v} \geq t$. Then $||\mathbf{u}||||\mathbf{v}|| \geq ||\mathbf{u}||||\mathbf{v}||cos(\theta) = \mathbf{u} \cdot \mathbf{v} \geq t$.

The converse of this lemma suggests if $||\mathbf{u}|| < \frac{t}{||\mathbf{v}||}$, then $\mathbf{u} \cdot \mathbf{v} < t$. So the vectors of shortest length will be the first to be be isolated since their dot products with the other vectors fail to obtain the threshold first.

We would then expect the maximal clique of $G_{t=0.85}$ to contain those vertices with the greatest norm. To check the validity of this concept we first need to observe that $\omega(G_{t=0.85}) = 6$. There are two cliques of this size (Figure 4.2).

While these cliques each include the three longest vectors, they are both missing the vertices whose vector is the fourth longest, namely the king penguin. In addition, the first clique includes the Adelie penguin, which is the third shortest vector. Thus these maximal cliques seem to be based on more on ω than the norm of the vectors.

A pattern of decreasing maximal cliques can be seen as $\omega(G_{t=0.9}) = 5$, $\omega(G_{t=0.95}) = 4$, and $\omega(G_{t=1}) = 3$. This steady decent suggests that that dot product representation

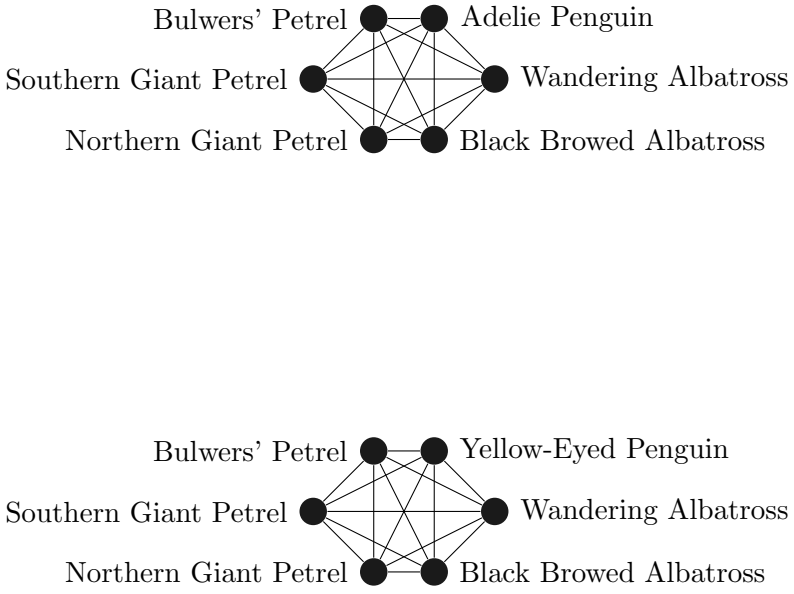


FIGURE 4.2: The Maximal Cliques of $G_{t=0.85}$.

allows for refinement of adjacency as the threshold t increases. Thus the longer a pair of vertices remains adjacent in the sequence of non-normalized graphs, the stronger the similarities between the life histories of the corresponding species.

As a six coloring of $G_{t=0.85}$ is possible, $\chi G_{t=0.85} = 6$. Similarly, $\chi G_{t=0.9} = 5$, $\chi G_{t=0.95} = 4$, and $\chi G_{t=1} = \chi G_{t=1.05} = 3$.

Finally, it can be seen that as with the normalized graphical representations, there are no chordless cycles of length greater than 3 or complete bipartite graphs with $n_1, n_2 \geq 2$ contained in these graphical representations either.

4.3 Comparisons Between The Different Model Types

These types of models yielded some interesting similarities and differences. The first difference that appears to occur is the connectedness of graphs. These differences are seen in the maximal clique size and maximum vertex degree. But these differences may be based more on our choice of threshold values. For example, both $G_{\theta=0.2}$ and $G_{t=0.9}$ are graphs where nine of the ten vertices are connected. While $\omega(G_{t=0.9}) > \omega(G_{\theta=0.2})$, it is possible that a minor increase in the threshold for $G_{t=0.9}$ would yield a graph on nine vertices with maximal clique of size 4.

Another noticeable difference is the first isolated vertex in each model. As we already mentioned, the first isolated vertex for the normalized models was the king

penguin, but the first isolated vertex for the non-normalized models was the Sooty albatross. While these species are different, it can be noted that the king penguin was still the fourth isolated species in the non-normalized models. This is important as it had the fourth longest vector. This suggests that it does have a life history that is significantly different than the other species, as we hypothesized in the analysis of the normalized models.

Additionally, it should be noted while the sooty albatross and Adelie penguin were among the first species isolated in the non-normalized models, they were among the last isolated among the normalized models. These contrasts may only re-enforce the similarities in life histories of these species. Both have very short vectors, and from the normalized model both vectors have a very small angle between them. Thus their life histories are likely to be very similar.

Another similarity illuminated was that existing between the northern giant petrel and the wandering albatross. Both were adjacent until the penultimate graphs in both series. Thus it strongly suggests a connection between the life histories of these species.

4.4 Limitations of Our Application

Both models are limited by their dot product dimension as we are using vectors of dimension 4. Thus it is impossible to create any graph with dot product dimension greater than four using our fitness gradients. While it has been conjectured that $\rho(G) \leq \lfloor \frac{n}{2} \rfloor$, where $n = |V(G)|$, this conjecture is still one of the open problems in dot product graphs [3]. Thus we cannot be sure whether this limitation would be eliminated if we reduced our models to nine vertices. On the other hand, it is known that there are several classes with dot product dimension of four or less. Some of these classes of graphs include trees, interval graphs, planar, and outer planar graphs [3][6]. Therefore, this limitation may not be a concern, particularly as our threshold changes to reduce the number of edges in the model.

4.5 Connection Depth

Our analysis also leads to a new measurement within our graphs, θ connection depth. *θ connection depth*, denoted $cd_\theta(xy)$, is the minimal θ for $x, y \in V(G)$ such that $xy \in E(G_\theta)$. We will similarly define *t connection depth*, denoted $cd_t(xy)$, is the maximal t for $x, y \in V(G)$ such that $xy \in E(G_t)$. Thus θ connection depth is related to the graphs with normalized vectors and t connection depth is related to the graphs

with non- normalized vectors. For example, with x=Northern Giant Petrel and y=Black Browed Albatross, $cd_\theta(xy) \leq 0.08$ since these vertices are adjacent in $G_{\theta=0.08}$. Since this angle is based on the gradient vectors for these species, this value can be calculated to be $cd_\theta(xy) = 0.0451$. The tables for both connection depths can be found in the Appendix.

Chapter 5

Conclusions

5.1 Conclusions

If our model has limitations, we believe these can be compensated for by attention to examination of both normalized and non-normalized models and threshold variation. For example, if a vertex is isolated early in the sequence of normalized graphs, examine the sequence of non-normalized graph. If in this second sequence the vertex is eliminated relatively early, then it can be interpreted that that vertex is significantly different than the other species.

Another combination that would provide significant information is a pair of vertices that are adjacent late in the sequence of normalized graphs and both vertices are isolated early in the sequence of non-normalized graphs. These combinations imply that both vertices are affected predominantly by the same variables of fitness and at similar rates.

The examinations of these networks was further augmented by the graph theory tools we used. But our selection of vertex degree, chromatic number, clique number, and cycle size are only a few of the parameters used in graph theory. Yet these parameters allowed us to more confidently identify connections between different species. Additional parameters such as clique cover number, boxicity, and chordality are a few other parameters that might be used to find further connections as well as strengthen the ones already determined.

This mixture of dot product graphs of both kinds of models (normalized and non-normalized) and graph theory tools provide the opportunity to analyze and exhibit graphical representations of biological networks. In this way, they allow us to continue to examine biological networks without eliminating parameters. This will give biologists more effective ways to examine these networks.

Thus we would encourage biologists to consider using these models. Mathematical

software such as MATLAB and Maple were used to generate our graphs. Therefore we would encourage others to use similar software to develop dot product graphs of the various biological networks they are working with.

The following are also some open problems left to be solved:

- Characterize the class of graphs represented by unit-length vectors.
- Does the dot product dimension change if the parameters used change? For example, if six different quantifiable parameters for a collection of species is known and it is known which species are similar, which of the six parameters are required to create a dot product graph with those known similarities?
- Given a set of fixed vectors and a fixed threshold, what is the probability that a particular graph is obtained?

Bibliography

[1] J.A. Bondy and U.S.R. Murty. *Graph Theory with Applications*. The Macmillan Press Ltd, 1976.

[2] R.L. Brooks. On colouring the nodes of a network. *Mathematical Proceedings of the Cambridge Philosophical Society*, 37:194–197, April 1941.

[3] Charles Fiduccia, Edward Scheinerman, Ann Trenk, and Jennifer Zito. Dot product representation of graphs. *Discrete Mathematics*, 181:113–138, 1998.

[4] S.S. Heppell, H. Caswell, and L.B. Crowder. Life histories and elasticity patterns: perturbation analysis for species with minimal demographic data. *Ecology*, 81:654–665, 2000.

[5] Ross Kang and Tobias Muller. Sphere and dot product representations of graphs. *Discrete Computational Geometry*, 47:548–568, 2012.

[6] Ross J. Kang, Laszlo Lovasz, Tobias Muller, and Edward R. Scheinerman. Dot product representations of planar graphs. *The Electronic Journal of Combinatorics*, 16, 2009.

[7] Russell Lande. Quantitative genetic analysis of multivariate evolution. *Evolution*, 33(1):402–416, March 1979.

[8] Russell Lande. A quantitative genetic theory of life history evolution. *Ecology*, 63:607–615, June 1982.

[9] Patrick H. Leslie. On the use of matrices in certain population mathematics. *Biometrika*, 33:183–212, 1945.

[10] Patrick H. Leslie. Some further notes on the use of matrices in population mathematics. *Biometrika*, 35:183–212, 1948.

[11] Alan Robertson. A mathematical model of the culling process in dairy cattle. *Animal Production*, 8:93–108, 1966.

[12] Stephen Young. *Random Dot Product Graphs: A Flexible Model for Complex Networks*. PhD thesis, Georgia Institute of Technology, 2008.

[13] Stephen J. Young and Edward Scheinerman. Directed random dot product graphs, 2009.

[14] Stephen J. Young and Edward R. Scheinerman. Random dot product graph models for social network. In *Of Lecture Notes In Computer Science*, pages 138–149. Springer, 2007.

APPENDIX

Table 5.1: Table of θ Connection Depths

	King Penguin	Adelie Penguin	Yellow -Eyed Penguin	Laysan Albatross	Wandering Albatross	Black Browed Albatross	Sooty Albatross	Northern Giant Petrel	Southern Giant Petrel	Bulwer's Petrel
King Penguin		0.2688	0.5094	0.4405	0.4597	0.5	0.3392	0.4855	0.6072	0.2792
Adelie Penguin			0.2481	0.1788	0.2725	0.254	0.0878	0.2546	0.3733	0.1401
Yellow-Eyed Penguin				0.1125	0.2108	0.1043	0.2122	0.1342	0.1855	0.2692
Laysan Albatross					0.2045	0.1014	0.1169	0.129	0.2206	0.2199
Wandering Albatross						0.1395	0.2642	0.0946	0.1729	0.1826
Black Browed Albatross							0.2082	0.0451	0.1222	0.2403
Sooty Albatross								0.2204	0.3296	0.1865
Northern Giant Petrel									0.1222	0.2159
Southern Giant Petrel										0.3343
Bulwer's Petrel										

TABLE 5.2: Table of t Connection Depths

	King Penguin	Adelie Penguin	Yellow -Eyed Penguin	Laysan Albatross	Wandering Albatross	Black Browed Albatross	Sooty Albatross	Northern Giant Petrel	Southern Giant Petrel	Bulwer's Petrel
King Penguin		0.8667	0.8079	0.7863	0.9176	0.8367	0.8247	0.879	0.852	0.9007
Adelie Penguin			0.8378	0.799	0.921	0.8619	0.8136	0.8985	0.9031	0.8665
Yellow-Eyed Penguin				0.8306	0.9627	0.9117	0.7845	0.9473	0.9814	0.8685
Laysan Albatross					0.9055	0.8567	0.7845	0.8905	0.9153	0.826
Wandering Albatross						1.0044	0.8981	1.0528	1.0883	0.9802
Black Browed Albatross							0.8476	0.9838	1.021	0.9015
Sooty Albatross								0.8814	0.8928	0.8367
Northern Giant Petrel									1.0646	0.9453
Southern Giant Petrel										0.955
Bulwer's Petrel										