STATISTICAL METHODS FOR DYNAMIC NETWORK DATA

by

Eric Alan Vance

Department of Statistical Science
Duke University

Date:

Approved:

Dr. David L. Banks, Supervisor

Dr. Mike West

Dr. Sayan Mukherjee

Dr. Tamraparni Dasu

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University

2008
ABSTRACT

(Statistics)

STATISTICAL METHODS FOR DYNAMIC NETWORK DATA

by

Eric Alan Vance

Department of Statistical Science
Duke University

Date: ____________________________

Approved:

_______________________________
Dr. David L. Banks, Supervisor

_______________________________
Dr. Mike West

_______________________________
Dr. Sayan Mukherjee

_______________________________
Dr. Tamraparni Dasu

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University

2008
Abstract

Motivated by questions involving three examples of dynamic network data, we apply, extend, and develop statistical methodology for the treatment of these data. Ranging from a handful of observations per week on African elephants, to 1,000 telephone calls per day in a calling network, to data streaming every five minutes from a server on a large AT&T network, our data are dynamic and the networks they induce—or are derived from—are constantly evolving, thus posing special challenges to the statisticians analyzing them. In this dissertation we demonstrate the usefulness of our statistical methodology by characterizing these dynamics and pinpointing their causes.

Chapter 1 discusses the evolution of social network models and introduces the themes which guide the rest of the work.

Chapter 2 answers questions about how and why wild African elephants interact through the use of a social network model.

Chapter 3 explains this social network model in detail and develops novel interpretations and uses for it.

Chapter 4 applies concepts from social network theory to discover fraud within a dynamic telephone calling network.

Chapter 5 discusses the challenges faced when analyzing dynamic, streaming data and develops methodology for handling this type of data. The method of DataSpheres is applied to a real data example to show how this methodology can be useful to managers of such streaming data.

Finally, we close, in Chapter 6, with some areas for future research that have opened up as a result of studying these models for dynamic network data.
Acknowledgements

This dissertation would not have been completed without the help and advice from the several people mentioned below and many more individuals not explicitly listed here, but who have helped and guided me throughout my life. To all of you I say, “Thank you.”

I extend my profound gratitude to my thesis advisor, David Banks, whose useful comments, understanding, and encouragement have provided me with the necessary support for this undertaking. Also to Tamraparni Dasu—my mentor and supervisor during my summer at AT&T—and my other committee members Mike West and Sayan Mukherjee, thank you for your help in my endeavors.

My colleagues at ISDS, SAMSI, NISS, AT&T, and Duke’s Department of Statistical Science (DSS) have helped me further my statistical knowledge in many ways. I am grateful to my officemate Jarad Niemi for making a special effort to help me think through difficult problems. Carlos Carvalho, Joe Lucas, James Scott, Ouyang Zhi, Natesh Pillai, Richard Hahn, Dan Merl, Mike Last, Edo Airoldi, Chris Volinsky, Peter Hoff, David Dunson, Michael Lavine, Kristian Lum, and many others have given me useful statistical advice immediately applicable to this dissertation. My collaborators and co-authors Beth Archie, Susan Alberts, and Julie Hollister-Smith have also directly contributed to my work.

I appreciate the advice given to me by Jerry Reiter and Laura Gunn about teaching, by Scotland Leman about writing and getting a job, and by my aunt Joan Mitchell who spurred me to try to write my dissertation in eleven days. I am also thankful to those who provided funding for me including the Duke Graduate School and DSS, SAMSI, NISS, AT&T, David Banks, and Ian Dinwoodie.
Contents

Abstract iv
Acknowledgements v
List of Tables x
List of Figures xi

1 Introduction 1
  1.1 Dynamic Streaming Data .......................... 1
  1.2 Social Network Models ............................ 3

2 Social Networks in Elephants 9
  2.1 Introduction ..................................... 9
     2.1.1 Background on elephant social behavior ....... 9
     2.1.2 Scientific questions ............................ 10
     2.1.3 Previous attempts .............................. 11
     2.1.4 Social network models ......................... 13
  2.2 Elephants, Data, and the Model .................... 14
     2.2.1 Family AA and its data ........................ 16
     2.2.2 The model ...................................... 21
     2.2.3 Model algorithm ............................... 26
  2.3 Analysis ........................................... 26
     2.3.1 Priors and estimation techniques .............. 27
     2.3.2 Results ....................................... 29
2.4 Conclusion .......................................................... 37

3 Social Space for Social Networks .................................. 44
   3.1 Social Space Explained ......................................... 44
      3.1.1 An anecdote .............................................. 44
      3.1.2 Social space for elephants .............................. 46
      3.1.3 Mathematical explanation ............................... 48
   3.2 Problems and Their Solutions ............................... 50
      3.2.1 Rotations and reflections ............................... 50
      3.2.2 Identifiability problems ............................... 52
   3.3 Practical Usefulness of Social Space ....................... 59
      3.3.1 Identifying missing covariates ........................ 59
      3.3.2 Choice of dimension .................................... 65
   3.4 Innovations .................................................... 67
      3.4.1 Analysis of pairwise effects ........................... 67
      3.4.2 Pickiness ............................................... 70

4 Fraud Detection in Calling Networks ............................ 75
   4.1 Introduction .................................................. 75
      4.1.1 Calling networks ........................................ 76
      4.1.2 Dynamic calling networks and Communities of
              Interest ................................................... 78
      4.1.3 Telephone fraud ......................................... 80
   4.2 Data from a Calling Network ................................ 80
      4.2.1 Visualizing data ......................................... 82
5.4.1 Power and the asymptotic relative efficiency of DataSpheres 129
5.4.2 Multivariate normal simulation .......................... 133
5.4.3 Heavy-tailed simulation ................................. 137

5.5 Application of DataSpheres to Network Server Data ........................................... 139
5.5.1 Variable selection and elimination .......................... 141
5.5.2 Univariate EDA ........................................ 142
5.5.3 Application of DataSpheres method ........................ 143

5.6 Conclusion .................................................. 149

6 Future Research ............................................. 150
6.1 Extensions and Modifications ................................. 150
6.1.1 Elephant families ...................................... 150
6.1.2 Calling networks ...................................... 152
6.1.3 DataSpheres ........................................... 152

6.2 Agent-based Modeling of Dynamic Social Networks ............................................. 153
6.2.1 Dynamic social networks ................................ 154
6.2.2 Rules incorporating higher-order effects ..................... 155
6.2.3 Bivariate extension ................................... 159
6.2.4 Discussion ........................................... 160

A Additional Social Space and Pickiness Plots ..................................................... 161

Bibliography ..................................................... 168

Biography ......................................................... 175
List of Tables

2.1 The observations matrix \((Y, N)_{\text{dry}}\) of Family AA affiliations in the dry season. The observations between Alison and Agatha were (142,311). ................................................................. 18

2.2 The observations matrix \((Y, N)_{\text{wet}}\) of Family AA affiliations in the wet season. The observations between Alison and Agatha were (115,186). ................................................................. 18

2.3 Centered matrix of Mother/Daughter kinship indicators, \(K^m\) .... 23

2.4 Centered matrix of Sisters kinship indicators, \(K^s\) .... 23

2.5 Centered matrix of genetic (DNA) relatedness kinship, \(K^r\) .... 24

2.6 Posterior results for the model parameters. ......................... 29

3.1 For the dry and wet seasons, the 95% highest posterior density intervals of the pairwise inner products for mother/daughter, sisters, and unrelated pairs, and for pickiness in the four models: Null, DNA-only, ModaSis, and Full kinship. ......................... 65

3.2 The posterior inner products for the various models generally decreased in size as more covariate information was added to the models. The first rows of the dry and wet season tables show the posterior mean absolute value of the size of these pairwise effects. The second rows show the posterior mean. The third and fourth rows show 95% highest posterior density intervals for the inner products and their associated effects in the model in terms of changes in predicted probabilities. ......................... 68

4.1 The triad census for the May 11th network. ......................... 91

4.2 The interesting triads created by the nine suspicious telephone numbers. .......................... 97

5.1 The 27 variables in the Server data set and their definitions. ... 140
List of Figures

1.1 Left: A three node network with a possible transitive tie between 
    $i$ and $k$. Right: Positions of the three actors in social space—the 
    shorter the distance, the higher the probability of a tie. 6

2.1 Known pedigree of Family AA with birth dates. 19

2.2 Examples in social space, from left to right, of a positive pairwise 
    effect; a negative pairwise effect; and a zero pairwise effect. 26

2.3 An example of transitivity: when two positions in social space $z_i$ 
    and $z_k$ are both close to a third position $z_j$, the two positions $z_i$ 
    and $z_k$ will necessarily be close to each other. 27

2.4 Posterior distributions for the intercept $\beta_0$ in the wet and dry 
    seasons. The intercept gives the baseline log odds of two elephants 
    affiliating. 30

2.5 Posterior distributions for the variance, $\sigma_{soc}^2$, of the sociability 
    random effect $\alpha_i \sim N(0, \sigma_{soc}^2)$. They are nearly indistinguishable 
    between the wet and dry seasons. 31

2.6 Posterior means of the elephants’ individual sociability random 
    effect $\alpha_i \sim N(0, \sigma_{soc}^2)$. Overall variability between seasons is similar, 
    but differences exist in individual sociabilities. 32

2.7 Posterior distributions for the three kinship coefficients $\beta_k^m$, $\beta_k^s$, $\beta_k^r$ 
    in the dry season and the wet season. 34

2.8 Posterior distributions in the dry and wet seasons for the overall 
    kinship effects $\beta_k^{[wet]}$ for mother/daughter pairs, sisters, and 
    unrelated/unknown pairs. 40

2.9 Posterior distributions for the variance, $\sigma_\gamma^2$, of the normal pairwise 
    error random effect $\gamma \sim N(0, \sigma_\gamma^2)$. This pairwise error was small in 
    both the wet and dry seasons. 41
2.10 Posterior distributions for the variances, $\sigma_{z_1}^2$, $\sigma_{z_2}^2$, of the social space position vectors in the wet and dry seasons. 42

2.11 Posterior mean positions $Z^*$ in social space in the dry season and wet season. 43

3.1 Two couples in marriage social space. Couple 1 (blue) has a positive inner product and a corresponding high probability of a successful marriage. Couple 2 (red) has a negative inner product indicating low probability of success in this example. 46

3.2 In a hypothetical elephant social space, elephants $i$ and $j$ share similar terrain preferences and are more likely to be observed together than with elephant $k$. 47

3.3 Examples, from left to right, of a positive inner product; a negative inner product; and an inner product equal to zero. 50

3.4 Examples, from left to right, of three vectors comprising $Z$; a rotation around the origin of $Z$ by $\frac{3}{2}\pi$ radians; and a reflection of $Z$ about the line $x = y$. Each set of vectors has the same inner product matrix $Z'Z$. 51

3.5 Untransformed positions in social space of elephant Amy with mean very near the origin. 53

3.6 Left: bimodality in Amy’s positions in social space when rotated or reflected such that $z_1$ lies on the positive $y$-axis. Right: no bimodality in Amy’s positions is present in when $Z$ is rotated or reflected so that $z_7$ lies on the positive $y$-axis. 54

3.7 A rotation around the point A—which is near the origin—works poorly in practice. The original configuration of points in social space in the plot on the left is rotated around A and reflected about C to produce the nonsensical plot on the right. 55
3.8 Top Left: Amy’s positions in social space after the Procrustean transformation using a poor choice of \( Z_0 \). Top Right: Amy’s positions after the Procrustean transformation using \( Z_0 = \hat{Z} \). Lower Left: Angelina’s positions in social space using the poor choice of \( Z_0 \). Lower Right: Angelina’s positions using \( Z_0 = \hat{Z} \). 

3.9 Left: the plot of social space positions \( \hat{Z} \) in the dry season for the Null model 3.7 shows four clusters of mother/daughter pairs. Right: the boxplots show the size of the posterior inner products for various types of pairs of elephants in Family AA. 

3.10 The plot of \( \hat{Z} \) in social space for the DNA-only model 3.8 still shows four clusters of mother/daughter pairs. The boxplots to the right show that including DNA relatedness did not greatly affect the size of the posterior inner products for the various types of elephant pairs. 

3.11 The plot of \( \hat{Z} \) in social space for the ModaSis Model 3.9 shows the four clusters of mother/daughter pairs much more spread out. The boxplots to the right show that including the Mother/Daughter and Sisters indicators caused the sizes of the posterior inner products for the various types of elephant pairs to be similar. 

3.12 The plot of \( \hat{Z} \) in social space for the Full kinship model 3.10 shows the four clusters of mother/daughter pairs much more spread out. The boxplots to the right show that including all the genetic and kinship covariates caused the sizes of the posterior inner products for the various types of elephant pairs to be similar. 

3.13 Pickiness decreased as more covariate information was added to the model. The elephants were generally more picky in the wet season than in the dry season in the Null (No Gen) and DNA-only model. The elephants were equally picky in both seasons when the Mother/Daughter and Sisters indicators were added to the model. 

3.14 Pickiness of the elephants during the dry season in the Full kinship model. 

4.2 Four successive weeks of streaming call dyads are aggregated into weekly networks. The graphs of these four networks show varying degrees of connectedness with the third week graph (lower left) being completely connected. 83

4.3 Global summary statistics by day for the one day graphs. 86

4.4 A and B form a mutually connected dyad. The pair C and D are an asymmetric dyad. A and C (also A and D, B and C, and B and D) form an “empty” dyad since no link exists between them. 87

4.5 Total number of unique calls, asymmetric dyads, and mutual dyads by day for the one-day networks. 88

4.6 The 16 types of triads, labeled by their number of mutual dyads, asymmetric dyads, null dyads, and a letter to further distinguish the triad if necessary. 90

4.7 Four of the most interesting types of triads. From left to right: the 021D triad, 021C, 111D, and 201. 93

4.8 The addition of a link between A and C to the intransitive 021C triad creates the transitive 030T triad. 93

4.9 Daily counts of four of the most interesting, variable triads. 94

4.10 A five-node network consisting of four asymmetric dyads. This graph has two types of triads: six 021D triads and four 003 empty triads. 95

4.11 The communities of interest of degree two for TN α for May 22nd. The graph shows all the numbers TN α called and all the calls associated with those 41 numbers. 97

4.12 Histogram of the times of the calls made by the nine suspicious TNs vs. the other 973 telephone numbers on May 22nd and 23rd. 99

4.13 Daily count of the four most interesting, variable triads without the 9 telemarketers included in the network. 101
4.14 Summary statistics by day without the 9 telemarketers in the network.


4.16 Daily counts of the seven rare triads: 030T, 030C, 120D, 102U, 120C, 210, and 300. These were unaffected by the nine telemarketers.

5.1 A 2-dimensional example of a DataSphere partition into sixteen bins according to depth layer and direction of greatest (positive or negative) variation.

5.2 Left: Euclidean distance forms circular contours of depth and orders the points A, B, then C. Right: Mahalanobis distance forms elliptical contours of depth and orders the points C, B, then A.

5.3 For various shifts in the mean $\mu_1$, the power of the Hotelling $T^2$ test and the DataSpheres method is plotted. The Hotelling test is more powerful for each of the three covariance matrices.

5.4 For various values of $||\mu_1||$ along the x-axis, the power of the Hotelling test and DataSpheres method is shown on the y-axis. The DataSpheres method is more powerful than the Hotelling $T^2$ test when the data have very heavy tails.

5.5 Daily boxplots of the three variables CPU Used %, Number of Procs, and Number of Threads.

5.6 Daily boxplots of the three variables Ping Latency, Used Swap, and Log Packets.

5.7 The data points from Week 1 were binned according to the DataSpheres method. Darker colors represent higher counts in each bin.

5.8 DataSpheres for Week 2, Week 3, Week 4, and Week 5. Darker colors represent higher counts in each bin.

6.1 Graph of correlated student friendship (red) and respect (blue) relationships.
6.2 Graph of student friendship (red) and respect (blue) relationships with importance placed on the gender covariate. Green numbers indicate male students while lavender numbers indicate females.

A.1 Left: \( \hat{Z} \) positions in social space in the wet season for the Null model 3.7. Four clusters of mother/daughter pairs are apparent. Right: boxplots show the size of the posterior inner products for various types of pairs of elephants in family AA.

A.2 Left: \( \hat{Z} \) positions in social space in the wet season for the DNA-only model 3.8. Four clusters of mother/daughter pairs are still apparent. Right: boxplots show that including DNA relatedness did not greatly affect the size of the posterior inner products for the various types of elephant pairs.

A.3 Left: \( \hat{Z} \) positions in social space in the wet season for the ModaSis model 3.9. The four clusters of mother/daughter pairs are much more spread out. Right: boxplots show that including the Mother/Daughter and Sisters indicators caused the sizes of the posterior inner products for the various types of elephant pairs to be similar.

A.4 Left: \( \hat{Z} \) positions in social space in the wet season for the Full kinship model 3.10. The four clusters of mother/daughter pairs are much more spread out. Right: boxplots show that including all the genetic and kinship covariates caused the sizes of the posterior inner products for the various types of elephant pairs to be similar.

A.5 Pickiness of the elephants in the wet and dry seasons in the Null model.

A.6 Pickiness of the elephants in the wet and dry seasons in the DNA-only model.

A.7 Pickiness of the elephants in the wet and dry seasons in the ModaSis model.

A.8 Pickiness of the elephants in the wet and dry seasons in the Full kinship model.
Chapter 1

Introduction

1.1 Dynamic Streaming Data

Motivated by questions involving three examples of dynamic network data, we apply, extend, and develop statistical methodology for the treatment of these data. Ranging from a handful of observations per week on African elephants, to data streaming every five minutes from a server on a large AT&T network, our data are dynamic and the networks they induce—or are derived from—are constantly evolving, thus posing special challenges to the statisticians analyzing them. In this dissertation we aim to characterize these dynamics and pinpoint their causes, thereby demonstrating the usefulness of our statistical methodology.

Large mammal ecologists studying the dynamics of wild elephant behavior observed the social relations of one family of elephants for a period of three years. Every week or so during this period, relational data between pairs of elephants were recorded with the goal of understanding how the social dynamics change from the wet season to the dry season, and what role genetics and kinship play in explaining elephant social behavior. In Chapter 2 we answer these questions
by applying a modified version of the social network model of Hoff (2005) to the
data aggregated by season and by elephant pair.

We treated these aggregated data as binomial random variables for each pair of
elephants, and by doing so, created a social network amongst the elephant family
members. We modified the model in order to handle this form of binomial data and
to accept user-defined prior distributions for the model parameters. In Chapter
3 we provide further detail about the unique feature of this model: its use of
“social space,” a space of unobserved latent characteristics that represent potential
transitive tendencies in the social network. We develop novel interpretations of
aspects of this social space and demonstrate how it can be used effectively for
model selection and for the identification of missing covariates by experts with
domain knowledge of the data.

Telephone calls between residential customers of a large communications com-
pany were the second source of data we encountered. Similar to the observations
on elephants, these were relational data which arrived over time, though at the
much faster rate of approximately 1,000 calls per day. When aggregated, these
data also induced social networks based on the presence or absence of calls be-
tween telephone numbers. We wished to understand if and in what ways the
networks changed over time, and whether any fraud occurred in our sample. In
Chapter 4 we track the frequency of certain types of triads in the networks which
leads to an interesting finding.

Managers and engineers at AT&T wish to understand in which ways the vari-
ous nodes in AT&T networks change over time so that problems can be identified
and fixed before spreading through more nodes of their networks. In Chapter 5
we address the problem of how to detect changes in rapidly accumulating, high-
dimensional data streams by developing the method of DataSpheres in the context of this special type of data. We focus on detecting changes in the streaming data from one server in a large network of e-commerce servers. We aggregate this streaming data by week and apply the method of DataSpheres to detect if and in what way the data stream has changed from week to week.

1.2 Social Network Models

Two of the three data sets we analyze in this dissertation consist of relational data, i.e., measurements or observations on the relationship between two individuals. We aggregate these data to form social networks and apply models and techniques from social network analysis to these data sets. Banks and Carley (1996) give an overview of the development of social network models and provide techniques for analyzing dynamic social networks. In this chapter we review the development of social network models based on their treatment of first, second, and third-order effects in the networks.

One of the first and simplest social network models is the random graph model of Erdős and Rényi (1959) which fixes the probability $\theta$ of an edge existing between any two of the $n$ nodes in the graph. Thus each of the $\binom{n}{2}$ edges has probability $\theta$, and every graph with $k$ edges is equally probable.

In the context of social networks, we consider the nodes of a graph to be actors or individuals in a network and the graph edges to be ties, relationships, or links between these actors. These terms represent the same things and we will use them interchangeably.

The Erdős-Rényi model is a very simple model that does not account for attributes of individual actors (first-order effects) that might affect the propensity
for an actor to send or receive ties. Nor does it allow the existence of a directed tie from one actor to another to affect the probability of a reciprocal tie (a second-order effect). Third-order effects that take into account the other ties in the network—effects such as transitivity, clusterability, and balance often seen in real social network data—are also absent from the random graph model.

In Chapter 2 we model the pattern of social relationships between ten members of a family of elephants in part by allowing each elephant to have her own “sociability,” which is her propensity to be together with other elephants. This sociability effect is a first-order effect unique to each elephant.

The \( p_1 \) model developed by Holland and Leinhardt (1981) uses the first-order effects of differential popularity and the second-order effect of reciprocity to model a social network. Let \( Y = [y_{ij}] \) be the adjacency matrix of the network for \( i, j \in \{1, \ldots, n\} \), where \( y_{ij} = 1 \) if a tie from \( i \) to \( j \) exists and \( y_{ij} = 0 \) otherwise. The likelihood of the network \( Y \) is:

\[
P(Y|\theta, \alpha, \beta, \rho) \propto \exp\left( \theta y_{++} + \sum_{i=1}^{n} \alpha_i y_{i+} + \sum_{j=1}^{n} \beta_j y_{+j} + \rho M \right),
\]

where \( y_{++} \) is the total number of ties in the network, \( y_{i+} = \sum_{i=1}^{n} y_{ij} \) is the out-degree of node \( i \), \( y_{+j} = \sum_{j=1}^{n} y_{ij} \) is the in-degree of node \( j \), and \( M = \sum_{i<j} y_{ij} y_{ji} \) is the number of mutual or reciprocal links.

The parameter \( \rho \) measures the force of reciprocation in the network. This second-order effect of reciprocity is very common in social network data but was not accounted for well by previous social network models. The density of the network is controlled by the parameter \( \theta \), similar to the \( \theta \) parameter in the Erdős-Rényi model. Controlling the first-order effects of differential popularity are the parameters \( \alpha \) and \( \beta \), called the productivity and attractiveness parameters. Each
actor $i$ has his own $\alpha_i$ and $\beta_i$. These parameters sum to zero which means the full specification of a directed network has $2n$ parameters.

The elephant data in Chapter 2 is, by its nature, undirected. This means that our model for this data ignored reciprocity as each tie between pairs of elephants was automatically reciprocated. On the other hand, this second-order effect in the calling network data of Chapter 4 was of great interest to us. The lack of reciprocity in many of the links between telephone numbers provided an important clue about the behavior of certain nodes within the network.

The $p_1$ model treats all dyads independently (except for mutual dyads) and does not allow for third-order effects such as transitivity, another effect common in real social network data. In contrast, Hoff, Raftery, and Handcock (2002) model dyads in a social network as marginally dependent since the probability of a tie between two actors depends on the locations of the actors in social space and thus depends on the other ties in the network. However, conditioned on the latent social space characteristics of each actor, the probability of a tie is independent of all other ties in the network:

$$P(Y | Z, X, \theta, \delta) = \prod_{i \neq j} P(y_{ij} | z_i, z_j, x_{ij}, \theta, \delta),$$

$$\text{logit}(p_{ij}) = \theta + \delta' x_{ij} - |z_i - z_j|,$$

(1.2)

where $p_{ij}$ is the probability that $y_{ij} = 1$, $X = [x'_{ij}]$ are (possibly vector-valued) covariates on all the dyads $(i, j)$, $\delta$ are the regression coefficient parameters on the dyadic covariates, and $Z = [z_i]$ is the $d \times n$ matrix of positions of the individuals in a $d$-dimensional latent social space.

As in the in the Erdős-Rényi and $p_1$ models, the parameter $\theta$ controls the density of ties for the Hoff et al. (2002) model, which we will henceforth call
the distance model. The dyadic covariate terms $\delta'X$ can potentially account for second-order effects. For example, if $x_{ij} = x_{ji}$, i.e., if the covariate information on the link from $i$ to $j$ is the same as that for the link from $j$ to $i$, then the probabilities of each link will be the same, $p_{ij} = p_{ji}$. In this case the coefficients $\delta$ could govern the force of reciprocity. An argument could also be made that under certain conditions the terms $\delta'X$ could account for differential popularity of the nodes, but the governing of first-order effects is not explicitly included in this model as it is in the $p_1$ model.

![Diagram of a three node network with possible transitive tie between $i$ and $k$.](image)

**Figure 1.1:** Left: A three node network with a possible transitive tie between $i$ and $k$. Right: Positions of the three actors in social space—the shorter the distance, the higher the probability of a tie.

However, an appealing feature of the distance model is that it can account for third-order effects via the latent positions $Z$ in social space. The interpretation of latent social space in this model is straightforward. Each individual has a value for $d$ latent characteristics. These values can be plotted in $d$-dimensional space, and the Euclidean distances between these points affect the probability of ties between
individuals. Individuals close together in social space will in general have higher probabilities of ties with each other than with individuals farther away in social space (see Equation 1.2). For example, if $i$ is linked to $j$ and $j$ is linked to $k$, this suggests that the pairs $(i, j)$ and $(j, k)$ are close together in social space. Therefore $i$ and $k$ should also be relatively close together in social space, implying that they are more likely to have the transitive tie from $i$ to $k$. This idea—that positions in social space can induce dependency between the ties and thus account for the higher probability of transitive relations often seen in social network data—was a key development in social network models. This principle is illustrated for a 2-dimensional social space in Figure 1.1. Social space can also induce reciprocity of ties since distance is symmetric.

Transitivity, or the lack thereof, amongst the calling network data in Chapter 4 plays an important role in our discovery of fraud within the system. Also, the ties between elephants in Chapter 2 are necessarily transitive for individual observation times. The data measure when elephants are together so that, for example, if $i$ and $j$ are observed near each other on Day $t$, and $j$ and $k$ are together as well, then $i$ and $k$ must be together. Therefore any model of these data should account for this third-order effect. We do so in Chapters 2 and 3 by using the model of Hoff (2005) which also incorporates a latent social space.

We call this model the Hoff model, as opposed to the distance model. It uses the inner product of two positions $\mathbf{z}_i^T \mathbf{z}_j$ in social space for actors $i$ and $j$ to increase or decrease the probability of a tie between $i$ and $j$. This model will be explained in detail in Chapters 3 and 2 and is shown in Equation 2.1. Briefly, this model is similar to the distance model in that it models the probability $p_{ij}$ of ties between individuals, includes a parameter $\theta$ that governs the density—
or the baseline probability—of ties within the network, and potentially contains measured covariate information on the dyads with the terms $\delta'X$. As opposed to using the distance between individuals in social space, the Hoff model incorporates third-order dependence via the bilinear (linear in both $i$ and $j$) social space effect $z'_iz_j$. In addition, it explicitly models the first-order effect of differential popularity through random effects terms $\alpha$ and $\beta$ which are analogous to the productivity and attractiveness fixed effects in the $p_1$ model (Equation 1.1). For directed networks, the tendency of ties toward reciprocity can be modeled through the correlation structure of the $\gamma_{ij}$ error term. Specifically, $(\gamma_{ij}, \gamma_{ji})' \sim N(0, \begin{bmatrix} \sigma_\gamma^2 & \rho \sigma_\gamma^2 \\ \rho \sigma_\gamma^2 & \sigma_\gamma^2 \end{bmatrix})$. In this way the Hoff model captures first, second, and third-order effects present in a social network.

The Hoff model and its application to elephant social structure is the focus of Chapters 2 and 3. In Chapter 4 we utilize the ideas of triads and third-order effects in a social network to analyze a dynamic telephone calling network. In Chapter 5 we develop the method of DataSpheres to detect changes in streaming network data. We conclude in Chapter 6 by suggesting potential modifications and extensions of our current work, and we introduce the concept of agent-based models of social networks for flexibly accounting for higher-order effects in a dynamic social network.
Chapter 2

Social Networks in Elephants

2.1 Introduction

2.1.1 Background on elephant social behavior

When male African elephants reach the ages of 14-17 years old, they leave their families to live solitary lives and associate with only a few other young, bachelor males until they start experiencing musth at around age 30 (Hollister-Smith et al., 2007). Young adult females, however, typically remain with their mothers and close female relatives for life. Thus, all herds of African elephants consist almost exclusively of adult females and their juvenile offspring. For the six or so days every 4-6 years a mature female is in estrus, male elephants—typically those who are in musth—will be seen lurking around the herd of adult females and juveniles fighting with the other adult males trying to mate. Therefore observed groups of elephants actually consist of adult females, their juvenile offspring, and the occasional adult males looking to mate.

The elephant herd is led by the matriarch, typically the oldest female in the group. Since elephants grow continuously (Hollister-Smith et al., 2007), the oldest
female also tends to be the biggest and strongest in the herd. The role of the matriarch in leading her family of elephants to sources of food and water is not entirely known, but McComb et al. (2001) present evidence that the matriarch serves as a repository of knowledge and memory of these food and water sources.

Generally, family groups consist of 2 to 10 adult females and their immature offspring. However, group size changes over the course of hours, days, or weeks. Female elephants live in a fluid society in which families can divide into units as small as a single adult female and her immature offspring, or entire family groups can fuse with other family groups to form bond groups, or even larger aggregations (Douglas-Hamilton (1972); Moss and Poole (1983); Wittemyer et al. (2005)). Although association patterns of individuals within families are flexible, family groups are characterized by consistent patterns of association, high frequencies of affiliative social behaviors (e.g., social rubbing or greeting), and coordinated movements and activities.

2.1.2 Scientific questions

One of the most important, unanswered questions about the evolution of animal behavior is: Why do some species live alone while others live in social groups? These differences in social organization, both within and across species, are presumed to be a result of adaptive evolution (Alexander (1974); Rubenstein (1978); Wrangham (1980); Pulliam and Caraco (1984)). One way that researchers have tried to answer this question is to relate differences in social organization with potential selective forces. For instance, species in low-resource habitats may live in smaller groups than species in high-resource habitats because the cost of competing for food with group members is high. While most such social animals live in
groups with a stable composition, “fission-fusion” species live in labile societies for which social groups can divide into sub-groups or fuse with other groups over short periods of time. These species respond flexibly to changes in the environment, and are thus ideal for testing hypotheses about the selective forces that have led to the evolution of group-living (Dunbar (1992); Kummer (1995)). Fission-fusion species are some of the most highly social animals on earth—including elephants, dolphins and other cetaceans, chimpanzees, and humans.

In 2004 Dr. Beth Archie, then a Duke graduate student in the Department of Biology, visited the Statistical Consulting Center for help quantifying elephant social structure. The two questions of most interest to her were: “How does the social structure of elephants change in the wet season versus the dry season?” and “What is the role of kinships in explaining elephant social behavior?” The rest of this chapter details our method for answering these questions and reveals the results of our inquiries.

2.1.3 Previous attempts

Few statistical methods have been applied to fission-fusion societies to test hypothesized predictors of social association, such as kinship or resource availability. In Archie et al. (2006) pairwise association indices (AI) between elephants are converted into measures of dissimilarity. These “distances” are then related to genetic relatedness and seasonality through multidimensional scaling, hierarchical cluster analysis, and the non-parametric Mantel test. The association index $AI_{ij}$ for two elephants $i$ and $j$ is the number of times they are seen together, $y_{ij}$, divided by the total number of times at least one of the two elephants is seen, $n_{ij}$. When this association index is subtracted from one, a measurement, $d_{ij}$ is
obtained which describes how far apart elephants $i$ and $j$ are from each other:

$$d_{ij} = 1 - AI_{ij} = 1 - \frac{y_{ij}}{n_{ij}}.$$

This measurement of how far apart two elephants are can be interpreted as the probability that two elephants are seen apart given that at least one of them has been seen. In fact, $d_{ij}$ is the maximum likelihood estimate for $1 - p_{ij}$, where $p_{ij}$ is the probability that elephants $i$ and $j$ are observed together given that at least one of them has been observed. This measurement $d_{ij}$ is not a true mathematical distance since two distinct elephants $i$ and $j$ could be always observed together, thus $y_{ij} = n_{ij}$ and $d_{ij} = 0$, but $i \neq j$. A true distance requires $d_{ij} \neq 0$ iff $i = j$.

However, the measurement $d_{ij}$ is symmetric, non-negative, and equals zero when $i = j$ and therefore satisfies the requirements for a dissimilarity coefficient (Chatfield and Collins, 1980). Symmetry can be shown since $y_{ij} = y_{ji}$ and $n_{ij} = n_{ji} \forall i, j$, therefore $d(i, j) = d(j, i)$. Also, $d(i, j) \geq 0 \forall i, j$ since $y_{ij} \leq n_{ij}$. When $i = j$, $y_{ij} = n_{ij}$, therefore $d_{ij} = 0$.

Archie et al. (2006) used the matrix of dissimilarities within a family of elephants to determine hierarchical cluster trees of elephants, also known as association trees. These trees were then used to categorize groups of elephants by how often they associated with the other elephants. For example, elephants with dissimilarities of between 0.9 and 1 were categorized into the most dissimilar aggregation. Their average genetic relatedness was computed and compared to the aggregations of elephants defined by having dissimilarities of between 0.8 and 0.9. The authors found that as the groups became successively less similar, the groups’ average genetic relatedness decreased. For example, the groups of elephants with dissimilarities between 0.4 and 0.5 had a higher average genetic relatedness than
the groups of elephants with dissimilarities between 0.5 and 0.6. This relationship between group dissimilarity and average genetic relatedness monotonically decreased for the ten aggregations of elephants in the Archie et al. study. Each aggregation, from the most dissimilar (0.9 to 1) to the least dissimilar (0.2 to 0.3), had successively higher average genetic relatedness between group members.

Another method used by Archie et al. was to calculate the linear correlation between the matrix of dissimilarities and the matrix of pairwise genetic relatedness by using a Mantel test. This test calculates the standard Pearson correlation coefficient between entries in the matrices, and assesses statistical significance of this correlation by repeatedly permuting rows and columns of one of the matrices and recalculating the correlation coefficient. The significance level of the observed correlation coefficient between the two matrices is the proportion of permutations that lead to a higher correlation coefficient. The results from Archie et al. (2006) were that the correlation between pairwise associations and genetic relatedness were generally strong and statistically significant.

Finally, the dissimilarity matrix can be used to depict the spatial or social relationships between individuals by projecting the dissimilarities onto two dimensions via multidimensional scaling. MDS can be a useful exploratory tool but cannot be used to infer the strength of predictors of elephant social behavior.

2.1.4 Social network models

In order to better quantify elephant social structure, we apply Peter Hoff’s (2005) bilinear mixed effects model for dyadic data in a novel way to relate changes in habitat and genetic relationship to changes in social organization. Social networks are statistical tools for modeling relationships between observed subjects, also
known as actors. In the field of behavioral ecology, social networks have been used to study relationships in a chimpanzee troupe (Sade and Dow, 1994) and for the identification of individuals crucial to social cohesion in dolphin pods (Lusseau and Newman, 2004).

There are a number of social network models, and their sophistication has increased steadily since the introduction of the $p_1$ model by Holland and Leinhardt (1981). Newer models have increased complexity and flexibility such as the distance model developed by Hoff, Raftery, and Handcock (2002) which models the probability of a relation between two actors dependent on their positions in an unobserved social space. Hoff’s bilinear mixed effects model (2005) refines the concept of social space. In this chapter we use this model to quantify certain key components of the pairwise social relationships between elephants by incorporating sender effects, receiver effects, pairwise effects, and exogenous variables. A component of the pairwise effect is the inner product of vectors of unobserved latent characteristics of a pair of actors. These vectors correspond to the positions of the two actors in a latent social space. This is attractive since (1) the inner product term for all pairs of actors can be viewed as a random effect with mean zero, (2) using a latent social space captures transitivity in the data, and (3) positions in this latent social space can be plotted and interpreted visually by researchers who have domain knowledge about elephant behavior.

2.2 Elephants, Data, and the Model

The major objective of this chapter is to demonstrate the novel use of bilinear mixed effects models to isolate factors that predict animal social behavior. These models extend traditional social network models by allowing unmeasured covari-
ates to appear as latent variables. Using these models, we investigate two factors that have previously been thought to be important to elephant social relationships: seasonality and kinship.

As in all fission-fusion species, spatial relationships for elephant herds are flexible. According to Moss and Poole (1983) and Wittemyer et al. (2005), the fission and fusion of elephant groups is associated with seasonality. In the dry season, social groups tend to be less cohesive and smaller; families are often divided into small subgroups of 1-4 adult females, and they rarely fuse with other families to form larger aggregations. In contrast, during the wet season, families often travel in intact groups. Whole families often fuse with other families, and sometimes hundreds of animals can be found together in one continuous aggregation. This relationship between seasonality and sociality is probably driven by resource availability. In the dry season, food is scarce, and groups fission to reduce the costs of resource competition, although groups of elephants are often still too large for the amount of available food. In the wet season, food is more abundant and individuals are able to live in larger groups.

A second factor that influences elephant social relationships is kinship. Female elephants spend most of their lives together in the same group with their first and second order maternal relatives (Archie et al., 2006). As a result of groups being closely related, the costs and benefits of sociality can accumulate both directly and indirectly. Sociality can affect an elephant’s direct fitness, which is the survival advantage given to an individual and her own offspring, and indirect fitness, which is the advantage given to the offspring of close genetic relatives. Hence, one of the major questions about the evolution of sociality in elephants—and other animals that live in kin groups—is to what extent has kin selection influenced
the evolution of social relationships? Our findings in this chapter provide clues towards answering this major question.

Over the past few decades, researchers have used parametric and non-parametric randomization models such as the Mantel test to investigate factors, such as resource availability or kinship, that predict social relationships. In the rest of this chapter, we apply the Hoff (2005) model to fit data collected on one elephant family group.

2.2.1 Family AA and its data

Since 1972, the elephants living in and around Amboseli National Park—at the base of Mt. Kilimanjaro in Kenya—have been studied by researchers at the Amboseli Elephant Research Project. Over the past decade, the Amboseli population has consisted of around 1200 individuals. Of these, the adult females and juveniles live in approximately 50 named family groups.

Dr. Beth Archie and other researchers collected data on these groups opportunistically. Each day from their Land Rover, the researchers searched for elephants on the savanna or in the woodlands, and when elephants were spotted, the researchers identified the families and family members present by sight. Overall size, tusk shape, body shape, head shape, ear shape, color, and body markings, including tears in the ears or scars, can be used to easily identify individual elephants. Elephants were also assigned an age. The ages of adult females born since 1972 are known to within 2 weeks, while the ages of females born prior to the onset of data collection were estimated from body size (Lee and Moss (1995); Moss (2001)).

The data from the “AA” family—one of the best-studied families in Amboseli—
are used to demonstrate the application of bilinear mixed effects models and the novel method to identify factors that predict animal social behavior. Recorded observations of the AA family began in 1972 and extend to the present day (2008).

In this chapter we analyze the spatial association patterns among the members of the AA family observed between 2000 and 2003. During that time, the family consisted of ten adult females ranging in age from 17 to 54. The oldest female, Amy, is the matriarch of the AA family. Data collection on the AA family was restricted to daytime hours, but took place during all times of the year. Whenever researchers were in a position to observe the AA family, they collected “scan samples” of association (Altmann, 1974). During scan sampling, observers recorded the spatial relationships among all visible elephants. Two elephants $i$ and $j$ were considered to have been together if they were part of the same general aggregation of elephants, and the most distant individual in the aggregation was no farther than 100 metres from her nearest neighbor (Archie et al., 2006). In other words, if elephants $i$ and $j$ were within 100 metres of each other, or were both within 100 metres of third elephant, or were otherwise “connected” to each other through a chain of elephants each within 100 metres of the other links in the chain, then $i$ and $j$ were considered to have been together. If the distance between elephants $i$ and $j$ was greater than 100 metres and there were no “connecting” elephants, the two females were considered apart. When one elephant was present and another was completely missing from the field of sight, these two elephants were also recorded as being apart. No inference was made on the affiliation status of two elephants when both were missing, i.e., neither was observed. Using these methods, the researchers collected 637 scans of spatial association within the AA family between 2000 and 2003.
Table 2.1: The observations matrix \((Y, N)^{dry}\) of Family AA affiliations in the dry season. The observations between Alison and Agatha were \((142,311)\).

<table>
<thead>
<tr>
<th></th>
<th>Alt</th>
<th>Amb</th>
<th>Ame</th>
<th>Amy</th>
<th>Ang</th>
<th>Anh</th>
<th>Ast</th>
<th>Aud</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amb</td>
<td>–</td>
<td>–</td>
<td>(168,286)</td>
<td>(226,252)</td>
<td>(205,244)</td>
<td>(169,298)</td>
<td>(135,292)</td>
<td>(204,257)</td>
</tr>
<tr>
<td>Ame</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>(176,300)</td>
<td>(155,292)</td>
<td>(224,241)</td>
<td>(116,309)</td>
<td>(164,295)</td>
</tr>
<tr>
<td>Amy</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>(215,256)</td>
<td>(189,300)</td>
<td>(141,308)</td>
<td>(228,255)</td>
</tr>
<tr>
<td>Ang</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>(165,295)</td>
<td>(136,284)</td>
<td>(199,255)</td>
</tr>
<tr>
<td>Anh</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>(117,321)</td>
<td>(176,296)</td>
</tr>
<tr>
<td>Ast</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>(142,296)</td>
</tr>
</tbody>
</table>

The elephants in the AA family comprise nodes in a social network with the links between elephants being the proportion of times the pair of elephants was observed together. In order to test whether the social structure of Family AA changed with rainfall and food availability, we separated the affiliation observations by season. For the dry season, which runs from April through October, researchers made 432 total observations of the AA family. Table 2.1 shows the observation data \((Y, N)^{dry}\), where \(y_{ij}\) is the number of times elephants \(i\) and \(j\) were observed together and \(n_{ij}\) is the number of times at least one of the elephants was seen.

Table 2.2: The observations matrix \((Y, N)^{wet}\) of Family AA affiliations in the wet season. The observations between Alison and Agatha were \((115,186)\).

<table>
<thead>
<tr>
<th></th>
<th>Alt</th>
<th>Amb</th>
<th>Ame</th>
<th>Amy</th>
<th>Ang</th>
<th>Anh</th>
<th>Ast</th>
<th>Aud</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aga</td>
<td>(156,159)</td>
<td>(132,175)</td>
<td>(135,185)</td>
<td>(144,175)</td>
<td>(145,175)</td>
<td>(132,185)</td>
<td>(110,186)</td>
<td>(145,172)</td>
</tr>
<tr>
<td>Ali</td>
<td>(115,183)</td>
<td>(120,170)</td>
<td>(111,192)</td>
<td>(123,179)</td>
<td>(123,180)</td>
<td>(111,189)</td>
<td>(131,148)</td>
<td>(123,177)</td>
</tr>
<tr>
<td>Alt</td>
<td>–</td>
<td>(132,172)</td>
<td>(135,182)</td>
<td>(141,175)</td>
<td>(142,175)</td>
<td>(132,182)</td>
<td>(110,183)</td>
<td>(144,170)</td>
</tr>
<tr>
<td>Ame</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>(128,193)</td>
<td>(129,193)</td>
<td>(158,161)</td>
<td>(113,185)</td>
<td>(128,191)</td>
</tr>
<tr>
<td>Amy</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>(160,161)</td>
<td>(128,190)</td>
<td>(120,177)</td>
<td>(155,163)</td>
</tr>
<tr>
<td>Ang</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>(128,191)</td>
<td>(120,178)</td>
<td>(156,163)</td>
</tr>
<tr>
<td>Anh</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>(113,182)</td>
<td>(125,191)</td>
</tr>
<tr>
<td>Ast</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>(117,178)</td>
</tr>
</tbody>
</table>

During the wet season, which runs from November through March, 205 sepa-
rate observations were made. The matrix \((\mathbf{Y}, \mathbf{N})^{\text{wet}}\) is shown in Table 2.2. These binomial affiliation data are the response variable in the bilinear mixed effects model detailed in the next subsection (Equation 2.1). The explanatory variables are the kinship relationships. We fit separate models for the wet season and the dry season affiliation data to determine the effects of seasonality on elephant social structure.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{family_tree.png}
\caption{Known pedigree of Family AA with birth dates.}
\end{figure}

From the birth records of the Amboseli elephants collected over the past 35 years we know a partial pedigree of the adult female members of the AA family (Figure 2.1). From direct observation, we know that Amy, the matriarch, has three adult female offspring Angelina, Audrey, and Amber. Alison and Astrid are a mother/daughter pair, as are Amelia and Anghared, and also Agatha and
Althea. Angelina, Audrey, and Amber are the only known half-sisters in the family.

Dr. Beth Archie used genetic analysis to estimate pairwise genetic relatedness among the members of this family where the pedigree relationships (e.g., between Amy and Amelia) were unknown. Archie et al. (2006) and Archie et al. (2003) provide a detailed description of how genetic relatedness was measured between elephants. To summarize, DNA was extracted from fecal samples collected from 236 adult females using the QIAamp DNA Stool Mini Kit (Qiagen, Valencia, CA). DNA samples were genotyped using a modified version of the multiple tubes approach (Taberlet et al., 1996) resulting in complete genotypes for 236 adult female elephants at 11 microsatellite loci. These microsatellites are regions within DNA that are repeated short sequences of nucleotides. The number of nucleotide repeats within the microsatellites varies between individuals, and thus are useful genetic markers for comparing DNA across individuals or populations.

From these microsatellite genotypes, the pairwise genetic relatedness was calculated via the method described in Queller and Goodnight (1989). This method estimates kinship by calculating the proportion of alleles two individuals share as a result of identity by descent (i.e., direct inheritance from a parent) with respect to the overall frequencies of alleles across the entire reference population. As a result, average pairwise genetic relatedness across the population is zero; average pairwise genetic relatedness between mothers and their offspring (i.e., first-order maternal relatives) is 0.5, while average pairwise genetic relatedness between half-siblings is 0.25. Two individuals sharing fewer alleles than expected by chance—given the allele frequencies of the reference population—will have a negative value of pairwise genetic relatedness. Given the sometimes poor quality of the DNA samples
and the inherent variability in the measurement process, the resulting pairwise genetic relatedness values were somewhat noisy. When the kinship between a pair of elephants was known, the theoretical value for genetic relatedness was used instead of the noisy measured value. For example, the mother/daughter pair Amy-Angelina had a noisy measured genetic relatedness value of 0.39, but 0.5 was used instead.

2.2.2 The model

In order to quantify and understand aspects of elephant social structure we model the probability of two elephants in Family AA being seen together. This pairwise probability depends on a family/seasonal baseline effect, individual effects, and dyadic effects. We follow the general framework of the bilinear mixed effects model for dyadic data proposed by Hoff (2005), and then extend it for binomial observations. We also modify the model to allow for user-defined prior distributions.

Using a logistic regression for binomial data we have, for elephants $i$ and $j$,

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \beta_0 + \alpha_i + \alpha_j + \beta_m^m k_{ij}^m + \beta_s^i k_{ij}^s + \beta_r^r k_{ij}^r + \gamma_{ij} + z'_i z_j ,$$  \hspace{1cm} (2.1)

where $y_{ij} \sim \text{Bin}(n_{ij}, p_{ij})$, i.e., $\frac{y_{ij}}{n_{ij}}$ is the proportion of times that elephants $i$ and $j$ are observed together given that at least one of them is observed. The other terms and parameters in the model are discussed below.

This model can be rewritten to show the error structure by separating the fixed effects from the random effects:

$$\log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = \beta_0 + \beta_m^m k_{ij}^m + \beta_s^i k_{ij}^s + \beta_r^r k_{ij}^r + \xi_{ij} ,$$

$$\xi_{ij} = \alpha_i + \alpha_j + \gamma_{ij} + z'_i z_j ,$$  \hspace{1cm} (2.2)
where the vector $z_i$ is elephant $i$’s position in latent social space.

The intercept $\beta_0$ in Equation 2.1 is a common term for all pairs of elephants in Family AA. Since all other terms in this model are centered or are mean-zero random effects, we interpret this intercept, common to all pairwise observations within the family during a season, as the baseline log odds of any two elephants in the AA family being observed together. If elephants tend to affiliate more or less often with each other in either the wet or dry seasons, this term should capture these seasonal differences.

Each elephant $i$ is assumed to have an innate “sociability” $\alpha_i$. We interpret this as an elephant’s tendency to associate with other elephants. This trait is modeled as a random effect among individuals in the family with $\alpha_i \sim N(0, \sigma_{soc}^2)$. Gregarious elephants should have a large value for their $\alpha$ sociability and would therefore have a higher probability of being observed with other elephants than would a more solitary elephant with a lower sociability. Of course, the labels gregarious and solitary apply to elephants only relative to the other individuals in the family. The average elephant with $\alpha = 0$ should be seen with other elephants an average proportion of times.

The model has five dyadic terms. Three of these are kinship covariates depending on the known or measured relationships between pairs of elephants: indicators for mother/daughter and half-sister relationships, and measured genetic relatedness.

These kinship terms $\beta_k \mathbf{K}$ are included in the model to test the hypothesis that elephants affiliate more often with kin. Our three measures of kinship:

1) Mother/Daughter pair indicator variables $k_{ij}^m$. We have six mother/daughter pairs in the AA family. We have centered this indicator variable so that the
Table 2.3: Centered matrix of Mother/Daughter kinship indicators, $K^m$.

<table>
<thead>
<tr>
<th></th>
<th>Aga</th>
<th>Ali</th>
<th>Alt</th>
<th>Amb</th>
<th>Ame</th>
<th>Amy</th>
<th>Ang</th>
<th>Anh</th>
<th>Ast</th>
<th>Aud</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aga</td>
<td>-</td>
<td>-0.133</td>
<td>0.867</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
</tr>
<tr>
<td>Ali</td>
<td>-0.133</td>
<td>-</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>0.867</td>
<td>-0.133</td>
</tr>
<tr>
<td>Alt</td>
<td>0.867</td>
<td>-0.133</td>
<td>-</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
</tr>
<tr>
<td>Amb</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-</td>
<td>-0.133</td>
<td>0.867</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
</tr>
<tr>
<td>Ame</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-</td>
<td>-0.133</td>
<td>0.867</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
</tr>
<tr>
<td>Amy</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>0.867</td>
<td>-0.133</td>
<td>-</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
</tr>
<tr>
<td>Ang</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>0.867</td>
<td>-</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
</tr>
<tr>
<td>Anh</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>0.867</td>
<td>-0.133</td>
<td>-</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
</tr>
<tr>
<td>Ast</td>
<td>-0.133</td>
<td>0.867</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
</tr>
<tr>
<td>Aud</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>-0.133</td>
<td>0.867</td>
<td>-</td>
<td>-0.133</td>
<td>-0.133</td>
</tr>
</tbody>
</table>

The average value of $K^m$ is 0. If a pair of elephants $i, j$ is a mother/daughter pair, $k^m_{ij} = 0.867$. Otherwise $k^m_{ij} = -0.133$. The values of $K^m$ for the AA family are shown in Table 2.3.

Table 2.4: Centered matrix of Sisters kinship indicators, $K^s$.

<table>
<thead>
<tr>
<th></th>
<th>Aga</th>
<th>Ali</th>
<th>Alt</th>
<th>Amb</th>
<th>Ame</th>
<th>Amy</th>
<th>Ang</th>
<th>Anh</th>
<th>Ast</th>
<th>Aud</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aga</td>
<td>-</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
</tr>
<tr>
<td>Ali</td>
<td>-0.067</td>
<td>-</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
</tr>
<tr>
<td>Alt</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
</tr>
<tr>
<td>Amb</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-</td>
<td>-0.067</td>
<td>0.933</td>
<td>-0.067</td>
<td>-0.067</td>
<td>0.933</td>
<td>-0.067</td>
</tr>
<tr>
<td>Ame</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
</tr>
<tr>
<td>Amy</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
</tr>
<tr>
<td>Ang</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>0.933</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-</td>
<td>-0.067</td>
<td>-0.067</td>
<td>0.933</td>
</tr>
<tr>
<td>Anh</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-</td>
<td>-0.067</td>
<td>-0.067</td>
</tr>
<tr>
<td>Ast</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-</td>
<td>-0.067</td>
</tr>
<tr>
<td>Aud</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-0.067</td>
<td>0.933</td>
<td>-0.067</td>
<td>-0.067</td>
<td>0.933</td>
<td>-0.067</td>
<td>-0.067</td>
<td>-</td>
</tr>
</tbody>
</table>

2) Sisters pair indicator variable $k^s_{ij}$. We have three pairs of half-sisters in the AA family, all are daughters of the matriarch Amy. We have centered this indicator variable so that the average value of $K^s$ is 0, and these values are shown in Table 2.4.

3) Genetic (DNA) relatedness measure $k^r_{ij}$ which measures how closely related elephants $i$ and $j$ are relative to that expected between two elephants in the Amboseli population. Some of these values are known through direct
observation of births within the family; the others are measured from DNA samples. We have centered this variable so that the average value of $K^r$ is 0, as shown in Table 2.5.

<table>
<thead>
<tr>
<th></th>
<th>Aga</th>
<th>Ali</th>
<th>Alt</th>
<th>Amb</th>
<th>Ame</th>
<th>Amy</th>
<th>Ang</th>
<th>Anh</th>
<th>Ast</th>
<th>Aud</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aga</td>
<td>−</td>
<td>0.200</td>
<td>0.320</td>
<td>0.070</td>
<td>−0.080</td>
<td>0.090</td>
<td>−0.100</td>
<td>−0.120</td>
<td>0.160</td>
<td>−0.330</td>
</tr>
<tr>
<td>Ali</td>
<td>0.200</td>
<td>−</td>
<td>−0.030</td>
<td>−0.100</td>
<td>−0.090</td>
<td>0.120</td>
<td>−0.160</td>
<td>−0.240</td>
<td>0.320</td>
<td>−0.160</td>
</tr>
<tr>
<td>Alt</td>
<td>0.320</td>
<td>−0.030</td>
<td>−</td>
<td>−0.060</td>
<td>−0.160</td>
<td>−0.130</td>
<td>−0.130</td>
<td>−0.220</td>
<td>0.030</td>
<td>−0.190</td>
</tr>
<tr>
<td>Amb</td>
<td>0.070</td>
<td>−0.100</td>
<td>−0.060</td>
<td>−</td>
<td>−0.270</td>
<td>0.320</td>
<td>0.070</td>
<td>−0.160</td>
<td>−0.160</td>
<td>0.070</td>
</tr>
<tr>
<td>Ame</td>
<td>−0.080</td>
<td>−0.090</td>
<td>−0.160</td>
<td>−0.270</td>
<td>−</td>
<td>0.170</td>
<td>−0.120</td>
<td>0.320</td>
<td>0.020</td>
<td>−0.170</td>
</tr>
<tr>
<td>Amy</td>
<td>0.090</td>
<td>0.120</td>
<td>−0.130</td>
<td>0.320</td>
<td>0.170</td>
<td>−</td>
<td>0.320</td>
<td>0.070</td>
<td>0.070</td>
<td>0.320</td>
</tr>
<tr>
<td>Ang</td>
<td>−0.100</td>
<td>−0.160</td>
<td>−0.130</td>
<td>0.070</td>
<td>−0.120</td>
<td>0.320</td>
<td>−</td>
<td>0.160</td>
<td>−</td>
<td>0.070</td>
</tr>
<tr>
<td>Anh</td>
<td>−0.120</td>
<td>−0.240</td>
<td>−0.220</td>
<td>−0.160</td>
<td>0.320</td>
<td>0.070</td>
<td>0.160</td>
<td>−</td>
<td>−0.080</td>
<td>0.020</td>
</tr>
<tr>
<td>Ast</td>
<td>0.160</td>
<td>0.320</td>
<td>0.030</td>
<td>−0.160</td>
<td>0.020</td>
<td>0.070</td>
<td>−</td>
<td>−0.080</td>
<td>−</td>
<td>−0.040</td>
</tr>
<tr>
<td>Aud</td>
<td>−0.330</td>
<td>−0.160</td>
<td>−0.190</td>
<td>0.070</td>
<td>−0.170</td>
<td>0.320</td>
<td>0.070</td>
<td>0.020</td>
<td>−0.040</td>
<td>−</td>
</tr>
</tbody>
</table>

The other two dyadic terms are error terms. The unstructured normally distributed error term $\gamma_{ij} \sim N(0, \sigma^2_\gamma)$ can be thought of as white noise or unexplained error. This term is the unexplained pairwise residual. It is the error in the log odds of two elephants $i$ and $j$ being observed together after accounting for their family’s baseline seasonal intercept $\beta_0$, their individual sociabilities $\alpha_i$ and $\alpha_j$, their kinships $k_{ij}$, and the inner product of their positions in social space $z'_i z_j$.

The bilinear term $z'_i z_j$ is structured error derived from a latent social space of unobserved characteristics of the individual elephants. This term is the inner product, or dot product, of the positions $z_i$ and $z_j$ of elephants $i$ and $j$ in the latent social space. Each $z_i \sim N(0, \begin{bmatrix} \sigma^2_{z1} & 0 \\ 0 & \sigma^2_{z2} \end{bmatrix})$.

We suppose that there are many factors besides an elephant family’s seasonal baseline, individual sociabilities, and kinship relationships that contribute to predicting how often two elephants will interact. These factors could include the observable and measurable such as age, dominance rank, and whether the female
has an immature calf; or unmeasurable idiosyncrasies such as habitat or food preferences. Including a latent social space of dimension $d$ to the model introduces the $d$ most important latent factors as explanatory variables. These $d$ latent factors can also reduce variability in the model by accounting for some of the third-order dependence relationships, such as transitivity, observed in social network data (Hoff, 2005).

If two elephants $i$ and $j$ have positions $z_i$ and $z_j$ in social space, the inner product of these positions is $z'_i z_j = \|z_i\| \cdot \|z_j\| \cos(\theta_{ij})$, where $\theta_{ij}$ is the angle between the two vectors. When elephants $i$ and $j$ are close together in angular distance in social space (thus sharing similar values of the $d$ factors as shown in Figure 2.2), then $z'_i z_j > 0$, and the elephants’ probability of affiliation is higher than otherwise predicted from the model. Elephants far apart in social space, and thus having dissimilar latent characteristics as shown in Figure 2.2, will have $z'_i z_j < 0$. If $z'_i z_j = 0$ then elephants $i$ and $j$ have positions in social space at right angles to each other as shown in Figure 2.2, and thus affiliate as often as their family seasonal intercept $\beta_0$, their individual sociabilities $\alpha_i$ and $\alpha_j$, and their kinship relationships $k_{ij}$ predict.

This pairwise effect $z'_i z_j$ captures some of the transitivity in the network of relations between elephants. For example, if elephants $i$ and $j$ affiliate often and therefore have positions $z_i$ and $z_j$ close together in social space, and if elephants $j$ and $k$ also affiliate often and have positions $z_j$ and $z_k$ close in social space as shown in Figure 2.3, then elephants $i$ and $k$ will likely have positions $z_i$ and $z_k$ close in social space, thereby increasing the log odds of affiliation between $i$ and $k$ than otherwise predicted by the model. For an intuitive and more detailed explanation of social space in two dimensions, see Chapter 3.
Figure 2.2: Examples in social space, from left to right, of a positive pairwise effect; a negative pairwise effect; and a zero pairwise effect.

2.2.3 Model algorithm

The parameters in the model (Equation 2.1) were estimated by constructing a Markov chain in the parameters which converged to their joint posterior distribution. Gibbs sampling through the full conditionals was followed at each iteration by a Metropolis-Hastings step in order to sample the dyad-specific parameters. We modified the code used by Hoff (2005) to allow for user-specified priors. Before, only empirical Bayes priors could be used.

2.3 Analysis

In order to compare differences in elephant social structure between the wet season and the dry season, we ran separate models for both seasons. By comparing the results from these two models we were able to determine which qualities of social structure changed with the seasons, and which stayed the same. In these Bayesian bilinear mixed effects models we designated vague, conjugate priors for the parameters and hyper-parameters. We generated draws from the posterior distributions and evaluated the practical significance of these resulting posterior
Figure 2.3: An example of transitivity: when two positions in social space $z_i$ and $z_k$ are both close to a third position $z_j$, the two positions $z_i$ and $z_k$ will necessarily be close to each other.

distributions.

2.3.1 Priors and estimation techniques

For visual interpretability and to avoid over-fitting the model by adding too many extra parameters, we chose the dimension $d$ of social space to be $d = 2$. Besides the effects of the family intercept, each elephant’s individual sociability, and the three kinship factors, of all the remaining various factors influencing how often elephants interact with each other, we include the effects of only the top two by designating a 2-dimensional social space. Using these two unnamed latent factors will allow the model to potentially capture idiosyncratic likes and dislikes between
elephants, and to better incorporate third-order effects such as transitivity.

For the parameters in the model (2.1) and all hyper-parameters we use vague proper priors. Priors for the wet and dry season models are the same.

Intercept: \( \beta_0 \sim N(0, 100) \)

Sociabilities: For \( i = 1 \ldots 10 \), \( \alpha_i \sim N(0, \sigma^2_{soc}) \), \( \sigma^2_{soc} \sim IG(0.5, 0.5) \)

Kinship Coefficients: \( \beta_k \sim N(0, 100 \times I_3) \)

Pairwise error: \( \gamma_{ij} \sim N(0, \sigma^2_\gamma) \), \( \sigma^2_\gamma \sim IG(0.5, 0.5) \)

Social space: For \( i = 1 \ldots 10 \), \( z_i \sim N(0, \begin{bmatrix} \sigma_{z1}^2 & 0 \\ 0 & \sigma_{z2}^2 \end{bmatrix}) \), \( \sigma_{z1}^2, \sigma_{z2}^2 \sim IG(0.5, 0.5) \).

Given this prior specification of the bilinear and random effects, the variance structure of the error term (2.2) is

\[
E(\xi_{ij}^2) = 2\sigma^2_{soc} + \sigma^2_\gamma + \sigma^4_{z1} + \sigma^4_{z2}.
\] (2.3)

The MCMC algorithm consisted of a Gibbs sampler with a Metropolis-Hastings step; see Hoff (2005) for details. Although the MCMC converged very rapidly to stable estimates of the parameters, we ran the chain for 220,000 iterations with a 20,000 iteration burn-in. These first 20,000 samples from the joint posterior distribution were excluded from further analysis.

The inner product operator is invariant under rotations and reflections. Therefore any rotation of the social space position vectors results in the same pairwise effect \( z^t_i z_j \) for elephants \( i \) and \( j \). In order to make posterior inferences on these vectors of latent characteristics, we used a Procrustean transformation to rotate and/or reflect the \( d \times n \) matrix of posterior positions \( Z \) (see Chapter 3 for a detailed treatment). The resulting transformed matrix \( Z^* \) was found using the
area-preserving linear transformation of the matrix of social space positions \( Z \) closest to a chosen starting matrix \( Z_0 \) (Shortreed et al., 2006):

\[
Z^* = Z_0 Z' (Z Z_0 Z')^{-\frac{1}{2}} Z.
\]  

(2.4)

Inferences about social space were made using these transformed positions \( Z^* \).

2.3.2 Results

Table 2.6: Posterior results for the model parameters.

<table>
<thead>
<tr>
<th></th>
<th>( \beta_0 )</th>
<th>( \sigma_{soc}^2 )</th>
<th>( \beta_m^s )</th>
<th>( \beta_k^s )</th>
<th>( \beta_k^r )</th>
<th>( \sigma_z^2 )</th>
<th>( \sigma_{z1}^2 )</th>
<th>( \sigma_{z2}^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wet Posterior Mean</td>
<td>1.25</td>
<td>0.21</td>
<td>2.48</td>
<td>1.48</td>
<td>-0.91</td>
<td>0.04</td>
<td>0.27</td>
<td>0.27</td>
</tr>
<tr>
<td>Dry Posterior Mean</td>
<td>0.50</td>
<td>0.20</td>
<td>1.85</td>
<td>0.88</td>
<td>0.31</td>
<td>0.03</td>
<td>0.29</td>
<td>0.30</td>
</tr>
<tr>
<td>Posterior Difference</td>
<td>0.76</td>
<td>0.01</td>
<td>0.63</td>
<td>0.59</td>
<td>-1.23</td>
<td>0.01</td>
<td>-0.02</td>
<td>-0.02</td>
</tr>
<tr>
<td>Prob (Wet &gt; Dry)</td>
<td>0.96</td>
<td>0.53</td>
<td>0.94</td>
<td>0.92</td>
<td>0.04</td>
<td>0.65</td>
<td>0.45</td>
<td>0.45</td>
</tr>
</tbody>
</table>

We found that the values of several of the parameters in the model (Equation 2.1) shifted from the wet season to the dry season indicating a change in the elephants’ patterns of social affiliation between the seasons. Table 2.6 summarizes these differences in the posteriors for the model parameters.

The posterior means were calculated from 200,000 draws from the joint posterior distributions of the parameters in the models for the two seasons. The differences in parameter values between seasons were calculated by finding the median difference of random samples from the posterior draws. The posterior probability of a wet season parameter being greater than the dry season parameter was calculated using the frequency of wet season samples greater than dry season samples.

The intercept \( \beta_0 \), or the family effect, was higher for the wet season than for the dry season. Figure 2.4 shows the posterior distributions of the intercepts. The
posterior mean $\hat{\beta}_0^{\text{wet}} = 1.25$, while $\hat{\beta}_0^{\text{dry}} = 0.50$. The posterior median difference $\beta_0^{\text{wet}} - \beta_0^{\text{dry}} = 0.76$. This seasonal change is a difference in the baseline log odds of two individual elephants affiliating with each other. The posterior probability of such a change, $\text{Prob}(\beta_0^{\text{wet}} > \beta_0^{\text{dry}})$, is 0.96. These results confirm the observations that elephants were more gregarious in the wet season than in the dry season.

The sociability random effect, $\alpha_i \sim \text{N}(0, \sigma_{\text{soc}}^2)$, changed little overall between seasons. Figure 2.5 shows the nearly identical posterior distributions of $\sigma_{\text{soc}}^2$ for the wet and dry seasons. These results show that elephants in Family AA varied in their individual sociabilities about the same amount in the wet season as in the dry season as $\hat{\sigma}_{\text{soc}}^2[\text{wet}] = 0.21$ and $\hat{\sigma}_{\text{soc}}^2[\text{dry}] = 0.20$. Figure 2.6 shows the
Figure 2.5: Posterior distributions for the variance, $\sigma_{soc}^2$, of the sociability random effect $\alpha_i \sim N(0, \sigma_{soc}^2)$. They are nearly indistinguishable between the wet and dry seasons.

posterior means of the elephants’ individual sociabilities in the dry season and the wet season.

Gregarious elephants ($\alpha > 0$) in the dry season were gregarious in the wet season too, while the relatively anti-social elephants ($\alpha < 0$) in the dry season remained so in the wet season. However, the sociabilities of three elephants, Amy, Amber, and Angelina, did change from one season to the other. Amy and Angelina had lower sociability in the dry season than in the wet season (with posterior probability 74% and 77% respectively) whereas Amber was relatively more gregarious in the dry season than in the wet season. The posterior Prob
Figure 2.6: Posterior means of the elephants’ individual sociability random effect \( \alpha_i \sim N(0, \sigma_{soc}^2) \). Overall variability between seasons is similar, but differences exist in individual sociabilities.

\[
(\alpha_{Amb[\text{dry}]} > \alpha_{Amb[\text{wet}]}) = 0.75.
\]

In the dry season the coefficient \( \beta^m_k[\text{dry}] \) for the Mother/Daughter pair indicator variable \( K^m \) was relatively large. The posterior mean \( \hat{\beta}^m_k[\text{dry}] = 1.85 \). None of the posterior draws of \( \beta^m_k[\text{dry}] \) were less than zero. Hence, that parameter was certainly positive. The coefficient \( \beta^s_k[\text{dry}] \) for the Sisters indicator \( K^s \) was smaller, but still far from zero. Its posterior mean was \( \hat{\beta}^s_k[\text{dry}] = 0.88 \), and the posterior probability of \( \beta^s_k[\text{dry}] > 0 \) was 99%. The coefficient \( \beta^r_k[\text{dry}] \) for the DNA relatedness measure \( K^r \) had a posterior mean \( \hat{\beta}^r_k[\text{dry}] = 0.31 \), meaning that in the dry season, two elephants sharing more DNA than the average pair in Family AA also had a tendency to affiliate more often than the average pair. However, 0 was within the 60% credible region for the posterior of \( \beta^r_k[\text{dry}] \), so it is possible that the DNA relatedness measure \( K^r \) had no effect on affiliation in the dry season.

Each pair of elephants can be classified into one of three kinship relationships:
mother/daughter, sisters, and unrelated/unknown. In the model each pair has three kinship covariates: Mother/Daughter indicator, Sisters indicator, and DNA relatedness. These three kinship covariates combine for an overall kinship effect on the log-odds of two elephants affiliating with each other. Because the kinship covariates have been centered, a mother/daughter pair has a value of $k^m = 0.867$ for the Mother/Daughter indicator covariate, $k^s = -0.067$ for the Sisters covariate, and a value of $k^r = 0.32$ for DNA relatedness.

Using the posterior means as our estimates of the kinship parameters, a mother-daughter pair $(i, j)$ has an overall posterior mean kinship effect in the dry season of:

$$\hat{\beta}_k[\text{dry}]' k_{ij} = 1.85 \times 0.867 + 0.88 \times -0.067 + 0.31 \times 0.32 = 1.644 .$$

The overall mean kinship effect in the dry season for a sisters pair $(i, j)$ was:

$$\hat{\beta}_k[\text{dry}]' k_{ij} = 1.85 \times -0.133 + 0.88 \times 0.933 + 0.31 \times 0.07 = 0.597 .$$

The overall mean dry season kinship effect for the average unrelated/unknown pair of elephants $(i, j)$ was:

$$\hat{\beta}_k[\text{dry}]' k_{ij} = 1.85 \times -0.133 + 0.88 \times -0.067 + 0.31 \times -0.059 = -0.323 .$$

In the wet season the coefficient $\beta_k^m[\text{wet}]$ for the Mother/Daughter pair indicator $K^m$ was again relatively large. The posterior mean $\hat{\beta}_k^m[\text{wet}] = 2.48$. Every posterior draw of $\beta_k^m[\text{wet}]$ was greater than zero so the parameter was positive. The coefficient $\beta_k^s[\text{wet}]$ for the Sisters indicator $K^s$ was again generally less than the Mother/Daughter coefficient $\beta_k^m[\text{wet}]$, but still far from zero. Its posterior mean $\hat{\beta}_k^s[\text{wet}] = 1.48$, and the posterior probability of $\beta_k^s[\text{wet}] > 0$ was almost 100%. The coefficient $\beta_k^r[\text{wet}]$ for the DNA relatedness measure $K^r$ had a posterior mean $\hat{\beta}_k^r[\text{wet}] = -0.91$. This negative value indicated that a pair of unrelated
elephants (not a mother/daughter or sisters pair) sharing more DNA than their family’s average pair, had a tendency to affiliate less often than the average pair in the wet season. The posterior probability of $\beta_k^{[wet]} < 0$ was 96%. Figure 2.7 shows the posterior distributions for the three kinship coefficients in the wet and dry seasons.

The overall posterior mean kinship effect in the wet season for a mother-daughter pair $(i, j)$ was $\hat{\beta}_k^{[wet]} k_{ij} = 2.48 \times 0.867 + 1.48 \times -0.067 - 0.91 \times 0.32 = $
1.76. The overall mean kinship effect in the wet season for a sisters pair \((i, j)\) was 
\[
\hat{\beta}_{k[^{\text{wet}}]} k_{ij} = 2.48 \times -0.133 + 1.48 \times 0.933 - 0.91 \times 0.07 = 0.987.
\]
The overall mean dry season kinship effect for the average unrelated/unknown pair of elephants \((i, j)\) was 
\[
\hat{\beta}_{k[^{\text{dry}}]} k_{ij} = 2.48 \times -0.133 + 1.48 \times -0.067 - 0.91 \times -0.059 = -0.375.
\]
Figure 2.8 shows the overall kinship effects for the three types of elephant pairs in our analysis.

The Mother/Daughter kinship coefficient was higher in the wet season than in the dry season with probability of \((\hat{\beta}_m[^{\text{wet}}] > \hat{\beta}_m[^{\text{dry}}]) = 0.94\), suggesting that elephants were more likely to be found with their mothers or daughters in the wet season than the dry, even controlling for the fact that the elephants had fewer associations overall in the dry season. This finding coincides with the observations that in the wet season mother/daughter pairs were rarely seen apart, but were frequently observed apart in the dry season. The Sisters kinship coefficient also changed from the wet season to the dry season. The median posterior value of this change was 
\[
(\hat{\beta}_s[^{\text{wet}}] - \hat{\beta}_s[^{\text{dry}}]) = 0.59,
\]
and the probability of \((\hat{\beta}_s[^{\text{wet}}] > \hat{\beta}_s[^{\text{dry}}])\) was 0.92. The DNA relatedness measure changed in the opposite direction from the wet season to the dry season. The posterior value of this difference 
\[
(\hat{\beta}_r[^{\text{wet}}] - \hat{\beta}_r[^{\text{dry}}]) = -1.23.
\]
The probability that \(\hat{\beta}_r[^{\text{wet}}] < \hat{\beta}_r[^{\text{dry}}]\) was 96%.

The normal pairwise error random effect, \(\gamma \sim N(0, \sigma_\gamma^2)\), remained small in the wet and dry seasons, indicating a good fit of the data for both seasons. This pairwise term—the error unexplained by the other terms in the model—varied in size a little more in the wet season than in the dry season with \(\hat{\sigma}_\gamma^2[^{\text{wet}}] = 0.04\) and \(\hat{\sigma}_\gamma^2[^{\text{dry}}] = 0.03\). Figure 2.9 shows the posterior distributions of \(\sigma_\gamma^2\) for the wet and dry seasons. The posterior probability of \(\sigma_\gamma^2[^{\text{wet}}] > \sigma_\gamma^2[^{\text{dry}}]\) was 65%.
The last term in the model (Equation 2.1), for predicting the frequency of affiliation between elephants $i$ and $j$, was the pairwise effect $z'_i z_j$. This pairwise effect was the inner product of the vectors of latent characteristics for $i$ and $j$. Intuitively, $z'_i z_j$ is the similarity in direction of $i$’s and $j$’s position vectors in social space, scaled by the vectors’ magnitudes. This term $z'_i z_j$ can also be interpreted as a mean-zero random effect since, for $l = 1, 2$, $z_l \sim N(0, \sigma^2_{z_l})$.

In both seasons, the posterior distributions of $\sigma^2_{z_1}$ and $\sigma^2_{z_2}$ were nearly identical (see Figure 2.10). The $\sigma^2_{z}$ parameters in the wet season model were about the same size as in the dry season. These very small posterior differences ($\sigma^2_{z_1}[\text{wet}] - \sigma^2_{z_1}[\text{dry}] = -0.02$ and $\sigma^2_{z_2}[\text{wet}] - \sigma^2_{z_2}[\text{dry}] = -0.02$) resulted in pairwise effects that were about the same magnitude in both seasons, and in positions in social space that varied by about the same amount in the wet season and the dry season.

Figure 2.11 shows the posterior means, $Z^*$, of the transformed positions in social space (Equation 2.4). Relative positions in social space describe idiosyncratic “likes and dislikes” between elephants that the other parameters in the model do not capture. For example, in the dry season, Amy and Alison were at opposite ends of social space, as far apart from each other as any pair of elephants. The inner product of their positions, $(z_{Amy} \cdot z_{Ali})[\text{dry}]$, had a posterior mean of $-0.15$, so their log odds of affiliating was $-0.15$ lower than predicted by the other terms in the model. However, in the wet season Amy and Alison actually affiliated more often than the model would have predicted without the bilinear effect. In fact, the posterior mean for $(z_{Amy} \cdot z_{Ali})[\text{wet}] = 0.06$. The posterior probability that $(z_{Amy} \cdot z_{Ali})[\text{wet}] > (z_{Amy} \cdot z_{Ali})[\text{dry}]$ was 0.82.
2.4 Conclusion

Our aims for applying a bilinear mixed effects model to the social networks in elephants were to see what the model would tell us about seasonal differences in patterns of elephant affiliations, and what effect kinship had on these relations. We found that the intercept in the model rose significantly from the dry season to the wet season. This change from an average $\hat{\beta}_0^{\text{dry}} = 0.50$ to $\hat{\beta}_0^{\text{wet}} = 1.25$ in the baseline log odds of affiliation corresponded to a change in the probability of two elephants being in the same place at the same time, assuming the other terms in the model summed to zero, from $\hat{p}^{\text{dry}} = 0.62$ to $\hat{p}^{\text{wet}} = 0.78$. Previous research has suggested that elephants affiliate more often in the wet season due to increased abundance of food. Their environment is able to support larger groups and more pairwise affiliations in the wet season (Moss and Poole, 1983). Our findings support this.

We also found the variances of the sociability random effect, the normal pairwise error, and the bilinear pairwise effects to be very similar in the wet and dry seasons. The expectation of the variance of the error in the model due to these components (Equation 2.3) was nearly the same in the two seasons. For the wet season, the posterior mean error $\hat{\xi}_2^{\text{wet}} = 0.67$. The dry season error $\hat{\xi}_2^{\text{dry}} = 0.68$. The median posterior difference was less than 0.01 between the seasons, and 50% of the posterior samples of $\xi_2^{\text{wet}}$ were larger than the dry season samples of $\xi_2^{\text{dry}}$. Thus, the expected variance in the model unexplained by kinships was the same in both seasons.

An attractive feature of the bilinear mixed effects model is its ability to incorporate exogenous variables measured on individuals or pairs of elephants. We
were able to determine the effect of kinships on the frequency of affiliation between pairs of elephants. These kinship effects were relatively large and did change from the wet season to the dry season. The Mother/Daughter indicator term had the largest effect in both seasons. Mother/daughter pairs had kinship covariates $k^m = 0.867$, $k^s = -0.067$, and $k^r = 0.320$. In the dry season these kinship effects combined to add 1.64 to the log odds, corresponding to an increase of 0.27 over the baseline mean probability $\hat{p}^{dry} = 0.62$ of affiliation. This resulted in a posterior prediction of 0.89 for the frequency of affiliation between mother/daughter pairs in the dry season. This effect was even greater in the wet season when the Mother/Daughter kinship coefficient $\beta^m_k$ increased. The combined kinship effects for mother/daughter pairs in the wet season raised the log odds an average of 1.76, corresponding to an increase in probability from $\hat{p}^{wet} = 0.78$ for the average pair, to 0.95 for mother/daughter pairs. However, though the increase in log odds due to kinships for these pairs was greater in the wet season, the increase in probability was larger in the dry season.

The predicted frequency of affiliation between sisters also increased due to the kinship terms in the model. In the dry season the predicted probability of affiliation increased from $\hat{p}^{dry} = 0.62$ to 0.75 with the combined kinship effects adding 0.59 to the baseline log odds. The increase in probability was similar in the wet season when the total kinship effect for the average sisters pair was 0.99, corresponding to an increase in the probability scale from $\hat{p}^{wet} = 0.78$ to 0.90. In both seasons the kinship terms increased the probability of sisters affiliating by 0.13.

The unrelated elephants in Family AA had average kinship terms $k^m = -0.133$, $k^s = -0.067$, and $k^r = -0.059$. The combined effect of these terms in the dry season
was a decrease in the log odds by $-0.32$. This changed the predicted probability of affiliation from $\hat{p}^{\text{dry}} = 0.62$ to 0.54. In the wet season the average unrelated elephant pair had a similar decrease in probability. The combined effects of the kinship terms were $-0.38$ in the log odds scale resulting in a posterior prediction of 0.71 for the frequency of affiliation. The kinship terms decreased the predicted probability of affiliation for unrelated pairs by 0.08 in the dry season and 0.07 in the wet season.

We found the effects of kinship, after including in the model the seasonal and family effect (intercept term) and individual elephants’ sociabilities, to be a significant driver of affiliation between elephants. These kinship effects have never before been so clearly demonstrated.
Figure 2.8: Posterior distributions in the dry and wet seasons for the overall kinship effects $\hat{\beta}_k[^{\text{wet}}] k$ for mother/daughter pairs, sisters, and unrelated/unknown pairs.
Figure 2.9: Posterior distributions for the variance, $\sigma^2_{\gamma}$, of the normal pairwise error random effect $\gamma \sim N(0, \sigma^2_{\gamma})$. This pairwise error was small in both the wet and dry seasons.
Figure 2.10: Posterior distributions for the variances, $\sigma_{z1}^2$, $\sigma_{z2}^2$, of the social space position vectors in the wet and dry seasons.
Figure 2.11: Posterior mean positions $Z^*$ in social space in the dry season and wet season.
Chapter 3

Social Space for Social Networks

3.1 Social Space Explained

The interpretation of the inner product social space of Chapter 2 and its implementation can be unintuitive and difficult to use in practice. In addition, little previous work has employed this type of social space for much more than improving model fit. In this chapter we present solutions for interpreting and implementing social space within social network models. We show how using social space within these models can lead to model improvement by indicating when covariates are missing from the model. Furthermore, we invent a new concept derived from social space, called pickiness, and demonstrate how this concept can be useful in practice.

3.1.1 An anecdote

Several years ago one of the author’s aunts told him that the most important factor for predicting whether two people would have a good marriage was the similarity of their attitudes towards money. If one person spent lots of money while the
other was thrifty, arguments and problems would ensue. If both husband and wife were thrifty, or both were equally extravagant, a happy marriage would be more likely.

A few years later while relaying that nugget of advice, the author’s sister overheard him and said, “That’s funny because Aunt Linda told me that the number one factor for predicting if a marriage would be successful was whether or not the couple had a similar sleep schedule. Between family, and careers, and other obligations there’s not much room for quality time with your spouse. If one of you is a morning person and the other is a night owl, that time together is further reduced.” The most successful couples, the claim went, were those that were closest together along the Early Bird to Night Owl continuum.

If we combine the advice and consider that there may be two important factors that predict success in marriage—attitude towards money and sleep schedule—we can create a marriage social space. Individuals have positions in this 2-dimensional space, and the inner product between positions shows the degree of compatibility between the two individuals. The inner product between any two vectors can be thought of as the similarity of their directions in social space, scaled by their magnitudes (the extremeness) of their positions. Individuals with no extreme positions, i.e., having positions near the origin, will have low inner product values with other individuals in social space no matter their direction.

Figure 3.1 shows two couples in marriage social space. According to this example, the blue couple would have a high probability of a successful marriage because they are both early birds and are similarly thrifty with their money. The red couple, on the other hand, would have a lower probability of success because of their extreme positions on opposite ends of the spending spectrum. Both are
Figure 3.1: Two couples in marriage social space. Couple 1 (blue) has a positive inner product and a corresponding high probability of a successful marriage. Couple 2 (red) has a negative inner product indicating low probability of success in this example.

similar in their preference for staying up very late at night, but the difference in their attitudes towards money would likely doom their relationship.

3.1.2 Social space for elephants

As an example of what social space might mean in the context of predicting how often elephants will be observed together, consider two hypothetical factors: preference for hills or flatland, and preference for sun versus shade. If individual elephants had preferences for being in hilly areas over flat areas, these elephants would be more likely to be observed together than with elephants preferring flat
landscapes. Inferring an elephant’s preference for terrain on the hilly to flat spectrum might help predict which elephants would be observed together more often. Likewise, inferring which elephants preferred sunny landscapes to shady ones, and how extreme these preferences were, might also contribute to predicting with which elephants they would be more likely to affiliate.

![Figure 3.2](image-url)

**Figure 3.2**: In a hypothetical elephant social space, elephants $i$ and $j$ share similar terrain preferences and are more likely to be observed together than with elephant $k$.

Combining these preferences to form a 2-dimensional social space with axes along the continuums of shade to sun and hills to flat would create a social space for elephants that was potentially useful in predicting how often they would be observed together. Figure 3.2 shows how elephants $i$ and $j$ might be more likely to be observed together than with elephant $k$ based on $i$ and $j$’s similar preferences.
for hilly, sunny terrain.

The above example draws an analogy between the physical spaces occupied by elephants and their positions in an unobserved social space of terrain preferences. When elephants are physically in the same space they will be observed as being together and that will be recorded in the data.

Despite the parallels between physical space and social space in this example, social space remains a mathematical construct. In practice its axes are unnamed and unknown. Positions in social space are unmeasured and unobserved. However, these positions are estimated from the data, and social space can be a useful addition to social network models. We use it to improve model fit of the data and to identify covariates that should be included in the model, as will be discussed in Section 3.3.1 below.

### 3.1.3 Mathematical explanation

Our general social network model (see Hoff, 2005) for the probability \( p_{ij} \) of a relationship directed from actor \( i \) to actor \( j \) is defined in terms of the linear predictor \( \theta_{ij} = \log \left( \frac{p_{ij}}{1-p_{ij}} \right) \). Here \( \theta_{ij} \) is the sum of an intercept term \( \beta_0 \), a sender-specific term \( S_i \) which might include covariates and a sender-specific random effect \( a_i^s \), an analogous receiver-specific term \( R_j \), a pairwise error term \( \gamma_{ij} \), and \( z_i'z_j \), which is the inner product of the positions in social space \( z_i \) and \( z_j \) of the actors \( i \) and \( j \).

\[
\begin{align*}
\theta_{ij} &= \beta_0 + S_i + R_j + \beta_d'X_{ij}^d + \gamma_{ij} + z_i'z_j, \\
S_i &= \beta_s'X_i^s + \alpha_i^s, \\
R_j &= \beta_r'X_j^r + \alpha_j^s.
\end{align*}
\] (3.1)
The term $z_i z_j$ is the inner product of the two vectors $z_i$ and $z_j$, and is sometimes written as $(z_i, z_j)$. It is also known as the dot product, written as $z_i \cdot z_j$, or the scalar product since the inner product maps two vectors to a scalar. For vectors $a$ and $b$ in $\mathbb{R}^d$,

$$\langle a, b \rangle \equiv a' b = \sum_{i}^{d} a_i b_j = a_1 b_1 + a_2 b_2 + \ldots + a_d b_d .$$

Immediately following from this definition is that, for real-valued vectors, the inner product is symmetric: $\langle a, b \rangle = \langle b, a \rangle$.

Let $\theta : 0 \leq \theta \leq \pi$ be the angle (in radians) between two vectors $a$ and $b$. This angle is defined in terms of the inner product:

$$\theta \equiv \arccos \frac{\langle a, b \rangle}{\|a\| \|b\|} .$$

From this definition we have an alternative expression for the inner product:

$$\langle a, b \rangle = a' b = \cos \theta \|a\| \|b\| . \quad (3.2)$$

In words, the inner product of two vectors is the cosine of the angle between them multiplied by their lengths.

Figure 3.3 shows that when the angle between two vectors is less than $\frac{\pi}{2}$, the inner product will be positive. When the angle is between them is greater than $\frac{\pi}{2}$, the inner product will be negative. The inner product will be 0 when the angle between the two vectors is $\frac{\pi}{2}$.

For $n$ actors in a $d$-dimensional social space, let

$$Z \equiv (z_1, z_2, \ldots, z_n) = \begin{pmatrix}
  z_{11} & z_{21} & \cdots & z_{n1} \\
  z_{12} & z_{22} & \cdots & z_{n2} \\
  \vdots & \vdots & \ddots & \vdots \\
  z_{1d} & z_{2d} & \cdots & z_{nd}
\end{pmatrix} ,$$

49
where $z_i$ is actor $i$’s position in social space, for $i = 1, \ldots, n$.

The symmetric $n \times n$ matrix of inner products between each of the $n$ positions in social space results from premultiplying $Z$ by its transpose $Z'$:

$$Z'Z = \begin{pmatrix}
    z'_1 z_1 & z'_1 z_2 & \cdots & z'_1 z_n \\
    \vdots & \vdots & \ddots & \vdots \\
    z'_n z_1 & z'_n z_2 & \cdots & z'_n z_n
\end{pmatrix}.$$

In the general model (Equation 3.1), the linear predictor $\theta_{ij}$ depends on $z_i$ and $z_j$ only through their inner product $z'_i z_j$. Therefore the probability model depends on $Z$ only through the matrix of inner products $Z'Z$. This poses an identifiability problem for $Z$ as will be discussed below in Section 3.2.2.

### 3.2 Problems and Their Solutions

#### 3.2.1 Rotations and reflections

As we have shown above in Equation 3.2, the inner product between two vectors depends only on the lengths of the vectors and the angle between them. Figure 3.4 illustrates how a rotation or reflection of $Z$ preserves:
1) the angles between the vectors of $\mathbf{Z}$,

2) the lengths of the vectors in $\mathbf{Z}$, and

3) all pairwise inner products in $\mathbf{Z}'\mathbf{Z}$.

Figure 3.4: Examples, from left to right, of three vectors comprising $\mathbf{Z}$; a rotation around the origin of $\mathbf{Z}$ by $\frac{3\pi}{2}$ radians; and a reflection of $\mathbf{Z}$ about the line $x = y$. Each set of vectors has the same inner product matrix $\mathbf{Z}'\mathbf{Z}$.

A rotation matrix $\mathbf{A}$ is defined to be any real, $d \times d$ orthogonal matrix with $\det(\mathbf{A}) = +1$. If one premultiplies a $d \times n$ matrix $\mathbf{B}$ by $\mathbf{A}$, the resulting matrix is a rotation of $\mathbf{B}$. When $d = 2$, $\mathbf{A}$ can be written as

$$\mathbf{A} = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix},$$

where $\theta$ is the angle of rotation around the origin measured from the positive $x$-axis.

From the definition above it follows that a rotation of $\mathbf{Z}$ will not change the matrix of inner products since

$$(\mathbf{A}\mathbf{Z})'(\mathbf{A}\mathbf{Z}) = (\mathbf{Z}'\mathbf{A}')\mathbf{A}\mathbf{Z} = \mathbf{Z}'\mathbf{I}\mathbf{Z} = \mathbf{Z}'\mathbf{Z},$$

51
where the second equality holds because $A$ is an orthogonal matrix, i.e., $A'A = I$.

Likewise, a reflection matrix $B$—defined to be any real, $d \times d$ orthogonal matrix with $\det(B) = -1$ that premultiplies a $d \times n$ matrix—will not change the matrix of inner products: $(BZ)'(BZ) = ZZ$.

In two dimensions a reflection matrix $B$ can be written as a reflection about the $y$-axis followed by a rotation:

$$B = A \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$ 

A rotation of a rotation matrix is a rotation matrix since, if $A_1$ and $A_2$ are rotation matrices, $\det(A_2A_1) = \det(A_2)\det(A_1) = (1)(1) = +1$, and $(A_2A_1)'A_2A_1 = A_1'A_2'A_2A_1 = A_1'I = A_1'A_1 = I$.

A rotation of a reflection matrix is a reflection matrix since its determinant is $-1$. Likewise, a reflection of a rotation matrix is also a reflection matrix. However, a reflection of a reflection matrix is a rotation matrix since its determinant is $+1$.

### 3.2.2 Identifiability problems

If one is content to make inferences on the pairwise inner products $Z'Z$ and the other parameters in the model (Equation 3.1) then estimating the individual positions $Z$ of actors in social space is unnecessary. However, in order to plot individual positions in social space so that experts with prior domain knowledge might visually assess the inadequacies of the model and identify potential covariates missing from the model, as will be discussed in Section 3.3.1, one must

Plotting the posterior draws of $Z$ is essentially meaningless because, as shown above, any rotation or reflection of $Z$ will leave unchanged the angles between
positions in social space and the magnitudes of these positions, and therefore the pairwise inner products \( ZZ \) will be the same for any rotation or reflection of \( Z \).

\[
\begin{array}{c}
\text{Figure 3.5: Untransformed positions in social space of elephant Amy with mean very near the origin.}
\end{array}
\]

Figure 3.5 demonstrates by example how plotting the posterior draws of \( z_i \) is pointless. Analyzing these untransformed positions is also meaningless. Only the magnitude of the position vectors of elephant Amy in this social space are identified. The angle \( z_{Amy} \) makes with the positive \( x \)-axis is arbitrary, as any rotation or reflection of \( z_{Amy} \) and the other positions is equivalent under the model. Not surprisingly, the point estimate \( \mathbf{z}_{Amy} \) of these untransformed posterior positions in social space is \((0.016, 0.013)\)—essentially the origin.

One possible strategy to combat the identifiability issues raised by rotations
and reflections of $Z$ in a 2-dimensional social space is to orient each posterior draw of $Z$ such that $z_i$ is on the positive $y$-axis ($z_{ix} = 0$ and $z_{iy} > 0$) and $z_j$ is to the left of the $y$-axis ($z_{jx} < 0$), where $i, j \in \{1, \ldots, n\}$ and $i \neq j$. The position $z_i$ will be a reference and will determine the angle of the rotation or reflection of $Z$. The second reference point $z_j$ will determine whether the operation is a rotation or a reflection.

![Figure 3.6](image)

**Figure 3.6:** Left: bimodality in Amy’s positions in social space when rotated or reflected such that $z_1$ lies on the positive $y$-axis. Right: no bimodality in Amy’s positions is present when $Z$ is rotated or reflected so that $z_7$ lies on the positive $y$-axis.

However, this strategy can fail in practice as demonstrated in Figure 3.6. The positions of Amy in this social space exhibit a false bimodality. The bimodality in the first figure is an artifact of the arbitrary choice of $z_1$ to be the reference position such that each posterior draw of $Z$ is rotated or reflected to force $z_1$ to lie on the positive $y$-axis and $z_2$ to be positioned to the left of the $y$-axis.

The right plot in Figure 3.6 shows Amy’s posterior positions in the social space.

54
when each $Z$ is rotated or reflected to force the reference positions $z_7$ to lie on the positive $y$-axis and $z_2$ to be on the left side of the $y$-axis (in quadrants II or III of social space). There is no apparent bimodality because $z_7$ is a better choice to use as a reference for rotation/reflection as it is typically farther from the origin than $z_1$. The posterior probability of $\|z_7\| > \|z_1\| = 0.88$.

![Figure 3.7: A rotation around the point A—which is near the origin—works poorly in practice. The original configuration of points in social space in the plot on the left is rotated around A and reflected about C to produce the nonsensical plot on the right.](image)

If the position used as a reference for the angle of rotation or reflection is near the origin, the inner products between this position and the other positions will necessarily be small no matter which quadrant of social space the reference position lies in. However, though variation between quadrants in social space for that reference position might not affect the set of inner products $Z^T Z$ much, the angle of rotation/reflection is very sensitive to small variations in the reference position’s location, and thus the rotation/reflection of $Z$ could have high variance.
The bimodality exhibited in Figure 3.6 is an artifact of slight variations in position \( z_1 \) which are magnified when \( Z \) is rotated/reflected. This behavior makes this method ill-suited for estimating \( Z \), as illustrated in Figure 3.7.

**Procrustean transformation**

In Hoff et al. (2002) and Hoff (2005) the authors compare posterior draws of \( Z \) only after transforming these positions using the Procrustean transformation operator which rotates (according to Hoff, 2005) or also potentially reflects (according to Hoff et al., 2002) \( Z \) so that it is as close as possible to an arbitrary fixed reference matrix \( Z_0 \) in terms of squared positional differences. The Procrustean transformation is given in Sibson (1978) as the solution to:

\[
\Omega = \arg\min_A (A Z - Z_0)'(A Z - Z_0) ,
\]

\[
\Omega = Z_0 Z'(Z Z'_0 Z_0 Z')^{-\frac{1}{2}} .
\] (3.3)

We verify that \( Z_0 Z'(Z Z'_0 Z_0 Z')^{-\frac{1}{2}} \) is indeed either a rotation or a reflection matrix by showing that it is orthogonal and that its determinant is \( \pm 1 \).

This matrix is orthogonal since:

\[
Z_0 Z'(Z Z'_0 Z_0 Z')^{-\frac{1}{2}} \times (Z_0 Z'(Z Z'_0 Z_0 Z')^{-\frac{1}{2}})' = Z_0 Z'(Z Z'_0 Z_0 Z')^{-1} ZZ'_0
\]

\[
= Z_0 Z'(Z Z'_0 Z_0 Z')^{-1} (Z Z'_0)^{-1} ZZ'_0
\]

\[
= I .
\] (3.4)

The inverses \((Z_0 Z')^{-1}\) and \((Z Z'_0)^{-1}\) exist because the rank of both \( Z_0 \) and \( Z \) is assumed to be \( d \).
The determinant of $Z_0 Z (Z Z_0 Z_0 Z')^{-\frac{1}{2}}$ is either $+1$ or $-1$:

$$
\det Z_0 Z (Z Z_0 Z_0 Z')^{-\frac{1}{2}} = \frac{\det(Z_0 Z')}{\det(Z Z_0 Z_0 Z')^{\frac{1}{2}}}
$$

$$
= \frac{\det(Z_0 Z')}{\sqrt{\det(Z_0 Z') \det(Z_0 Z')}}
$$

$$
= \frac{\det(Z_0 Z')}{\pm \det(Z_0 Z')}
$$

$$
= \pm 1.
$$

Therefore $Z_0 Z (Z Z_0 Z_0 Z')^{-\frac{1}{2}}$ is either a rotation or reflection matrix.

However, as we have seen in Figure 3.6, rotating or reflecting $Z$ according to some reference positions can be a bad strategy. Rotating or reflecting $Z$ to have the minimum difference from $Z_0$ in terms of Frobenius norm, i.e., $\sqrt{\text{trace}(Z - Z_0)'(Z - Z_0)}$, might not work ideally in practice, even if $Z_0$ is chosen carefully. Figure 3.8 shows Amy’s and Angelina’s positions in social space for a poorly chosen $Z_0$, and for a well chosen $Z_0$. The poor choice for the reference matrix $Z_0$ can cause extra variation in the positions in social space or, in some cases, multimodality.

The social network model (Equation 3.1) of Hoff (2005) only depends on $Z Z'$ which, as has been shown above, is unchanged by any rotation or reflection of the posterior draws of $Z$. Therefore, we compute the posterior mean of $Z Z'$ and obtain an estimate $\hat{Z}$ from the eigendecomposition of $\overline{Z Z}$. This estimate $\hat{Z}$ of the positions in $d$-dimensional social space is the first $d$ eigenvectors of $\overline{Z Z}$ multiplied by the square root of their corresponding eigenvalues:

$$
\overline{Z Z} = Q \Lambda Q'
$$

$$
\hat{Z} = \left[ q_1 \sqrt{\lambda_1}, q_2 \sqrt{\lambda_2}, \ldots, q_d \sqrt{\lambda_d} \right] \quad (3.6)
$$
Figure 3.8: Top Left: Amy’s positions in social space after the Procrustean transformation using a poor choice of $Z_0$. Top Right: Amy’s positions after the Procrustean transformation using $Z_0 = \hat{Z}$. Lower Left: Angelina’s positions in social space using the poor choice of $Z_0$. Lower Right: Angelina’s positions using $Z_0 = \hat{Z}$. 

58
In general, \( \hat{Z}'\hat{Z} \neq \overline{Z}\overline{Z} \) unless \( \lambda_{d+1}, \ldots, \lambda_n = 0 \).

These point estimates \( \hat{Z} \) can be plotted for all the actors in the social network model. A visual inspection of these plots, combined with domain knowledge, can give insight into the processes underlying the model, as will be discussed in Section 3.3.1.

### 3.3 Practical Usefulness of Social Space

#### 3.3.1 Identifying missing covariates

In the absence of any covariate information, a basic model for an undirected relationship between actors \( i \) and \( j \) would include an intercept, “sociability” random effects for \( i \) and \( j \), an error term, and the inner product between \( i \)’s and \( j \)’s positions in social space. This Null model is given by:

\[
\begin{align*}
\theta_{ij} &= \beta_0 + S_i + R_j + \beta_d' X_{ij}^d + \gamma_{ij} + z_i' z_j, \\
S_i &= \alpha_i, \\
R_j &= \alpha_j, \\
X_{ij}^d &= 0,
\end{align*}
\]

(3.7)

This Null model 3.7 was run for the AA family of elephants in the wet and dry seasons to see how the social space would look without using any genetic or kinship information. The matrix of social space positions \( \hat{Z} \) for the dry season is plotted in Figure 3.9. The social space for the wet season looks similar. See Figures A.1-A.4 in Appendix A for the \( \hat{Z} \) plots and the boxplots of inner products for all the models in the wet season.
Figure 3.9: Left: the plot of social space positions $\tilde{Z}$ in the dry season for the Null model 3.7 shows four clusters of mother/daughter pairs. Right: the boxplots show the size of the posterior inner products for various types of pairs of elephants in Family AA.

Based on knowledge of the pedigree chart for the AA family (Figure 2.1), the interpretation of this social space plot is immediately clear: Figure 3.9 shows four clusters of mother/daughter pairs. The largest cluster is of Amy and her three daughters. The other three clusters are of the other three mother/daughter pairs in the AA family. Since the angle between the positions of these mother/daughter pairs is very small, the pairwise inner product is high. This indicates that the pairs are seen together more often than the Null model 3.7 (intercept and sociability effects) would otherwise predict.

The other plot in Figure 3.9 shows boxplots of the inner products for mother-daughter pairs, sisters, and unrelated pairs. As we can also see from the social space plot, the posterior inner products between mother/daughter pairs and sisters were higher than for unrelated pairs of elephants. The last boxplot shows the
posterior inner products of the positions with themselves, e.g. \((z_{\text{Amy}})'z_{\text{Amy}}\). These inner products are the squares of the lengths of positions in social space and will be discussed in Section 3.4.2.

An immediate consequence of recognizing this visual pattern of mother/daughter pairs positioned very close together in social space is the realization that genetic or kinship information should be incorporated into the model. Adding the measured DNA relatedness (see Section 2.2.1) to the model might help explain why mother/daughter pairs or sisters are observed together more often than unrelated pairs. The DNA-only model 3.8 below is similar to the “No genetics” Null model except that \(k_{ij}^r\), the pairwise measure of similarity in DNA, is added. The DNA-only model is:

\[
X_{ij}^d = k_{ij}^r ,
\]

\[
\log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \theta_{ij} = \beta_0 + \alpha_i + \alpha_j + \beta_r^r k_{ij}^r + \gamma_{ij} + z_i'z_j. \quad (3.8)
\]

A simple check for someone with domain knowledge of the elephants in Family AA to see if adding DNA relatedness to the model was effective in capturing the tendency of highly related pairs (e.g. mother/daughters and/or sisters) to be observed together more often than otherwise predicted is to look at their positions in social space (Figure 3.10).

Figure 3.10 shows that mothers and their daughters were still tightly clustered around each other. In fact, adding DNA relatedness to the model did not fully account for the positive residual pairwise effect of mother/daughter pairs and sisters. The boxplots of the various types of pairs show that the posterior inner products for mother/daughter pairs and sisters were still significantly higher than for unrelated pairs.
Adding an indicator for a mother/daughter relationship \(k_{ij}^{m}\) and one for sisters \(k_{ij}^{s}\) for the pair of elephants \(i\) and \(j\) is an alternative to using measured DNA relatedness to account for a kinship effect in the social interactions between elephants. Using these known kinships (Moss, 2001), instead of the measured DNA relatedness, gives us the ModaSis model:

\[
\begin{align*}
X_{ij}^d &= (k_{ij}^m, k_{ij}^s)^T, \\
\theta_{ij} &= \beta_0 + \alpha_i + \alpha_j + \beta_m k_{ij}^m + \beta_s k_{ij}^s + \gamma_{ij} + \mathbf{z}_i \mathbf{z}_j.
\end{align*}
\]  

Figure 3.10 shows that mothers and their daughters were much less tightly clustered around each other, and the sisters were also more spread out in the social space of the ModaSis model 3.9. In fact, adding the Mother/Daughter indicator and the Sisters indicator resulted in one elephant, Angelina, moving far away in
The plot of $\hat{Z}$ in social space for the ModaSis Model 3.9 shows the four clusters of mother/daughter pairs much more spread out. The boxplots to the right show that including the Mother/Daughter and Sisters indicators caused the sizes of the posterior inner products for the various types of elephant pairs to be similar.

The model analyzed in Chapter 2 was the full model incorporating all of the genetic and kinship information in the form of the DNA relatedness measure and the Mother/Daughter and Sisters indicators. This Full kinship model is given by:

$$
\begin{align*}
X_{ij}^d &= (k_{rij}, k_{mij}, k_{sij})', \\
\theta_{ij} &= \beta_0 + \alpha_i + \alpha_j + \beta_r' k_{rij} + \beta_m' k_{mij} + \beta_s' k_{sij} + \gamma_{ij} + z_i z_j.
\end{align*}
$$

(3.10)
Figure 3.12: The plot of $\hat{Z}$ in social space for the Full kinship model 3.10 shows the four clusters of mother/daughter pairs much more spread out. The boxplots to the right show that including all the genetic and kinship covariates caused the sizes of the posterior inner products for the various types of elephant pairs to be similar.

As was shown in Chapter 2, when the Mother/Daughter and Sisters indicators were in the model, the DNA relatedness measure was largely irrelevant. The same conclusion can be reached merely by comparing the plots of $\hat{Z}$ in social space for the Full kinship model (3.10) and the ModaSis model (3.9). Figure 3.12 shows that $\hat{Z}$ in the Full kinship model was very similar to $\hat{Z}$ in the ModaSis model. The boxplots also show that the inner products of the social space positions of mother/daughter pairs remained similar to the inner products between unrelated pairs.

Table 3.1 shows the 95% credible regions (highest posterior density) for the inner products of mother/daughter pairs, sisters, and unrelated pairs. As more genetic and kinship covariate information was added to the model, the posterior inner products shrunk towards zero. This pattern was consistent for both the dry
Table 3.1: For the dry and wet seasons, the 95% highest posterior density intervals of the pairwise inner products for mother/daughter, sisters, and unrelated pairs, and for pickiness in the four models: Null, DNA-only, ModaSis, and Full kinship.

<table>
<thead>
<tr>
<th></th>
<th>Dry Season</th>
<th>Null</th>
<th>DNA-only</th>
<th>ModaSis</th>
<th>Full kinship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mother/Daughters</td>
<td>(-0.18, 1.81)</td>
<td>(-0.24, 1.29)</td>
<td>(-0.64, 0.74)</td>
<td>(-0.60, 0.77)</td>
<td></td>
</tr>
<tr>
<td>Sisters</td>
<td>(-0.22, 1.29)</td>
<td>(-0.31, 0.65)</td>
<td>(-0.25, 0.59)</td>
<td>(-0.25, 0.58)</td>
<td></td>
</tr>
<tr>
<td>Unrelated pairs</td>
<td>(-1.10, 0.74)</td>
<td>(-0.90, 0.81)</td>
<td>(-0.42, 0.47)</td>
<td>(-0.46, 0.47)</td>
<td></td>
</tr>
<tr>
<td>Pickiness</td>
<td>(0, 2.24)</td>
<td>(0, 1.74)</td>
<td>(0, 0.98)</td>
<td>(0, 0.98)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Wet Season</th>
<th>Null</th>
<th>DNA-only</th>
<th>ModaSis</th>
<th>Full kinship</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mother/Daughters</td>
<td>(-0.14, 2.39)</td>
<td>(-1.37, 1.85)</td>
<td>(-0.34, 0.77)</td>
<td>(-0.37, 0.63)</td>
<td></td>
</tr>
<tr>
<td>Sisters</td>
<td>(-0.07, 2.01)</td>
<td>(-0.34, 1.62)</td>
<td>(-0.43, 0.73)</td>
<td>(-0.38, 0.61)</td>
<td></td>
</tr>
<tr>
<td>Unrelated pairs</td>
<td>(-1.25, 0.81)</td>
<td>(-1.21, 0.91)</td>
<td>(-0.45, 0.47)</td>
<td>(-0.39, 0.42)</td>
<td></td>
</tr>
<tr>
<td>Pickiness</td>
<td>(0, 2.79)</td>
<td>(0, 2.29)</td>
<td>(0, 1.00)</td>
<td>(0, 0.88)</td>
<td></td>
</tr>
</tbody>
</table>

season and the wet season.

Further examination of the social space plots and the boxplots of inner products, combined with expert domain knowledge, could lead to the identification and inclusion of additional covariates. Perhaps the clusters remaining in the social space of the Full kinship model were due to the elephants being in similar age cohorts or to the presence/absence of infant or juvenile elephants. At this time, such speculations have not been discussed with the biologists who have the expert domain knowledge on these issues. However, as shown above with the examples of kinship covariates, the visual inspection of social space can lead to identification of useful covariates in the model.

3.3.2 Choice of dimension

The method of choosing the dimension $d$ of social space depends on the goal of the analysis (Hoff, 2005). The goal in this chapter for analyzing the elephants’ social
structure was to be able to interpret social space visually so as to easily identify clusters of elephants and other qualities such as the size of the inner products of positions with themselves. The visual identification of meaningful clusters (e.g. mother/daughter pairs) was best accomplished with a 2-dimensional social space, hence the choice of $d = 2$.

In Chapter 2 the goal of the analysis was to quantify aspects of elephant social behavior and identify changes in social structure from season to season. A secondary goal was to present the social space plots to the biologists with expert domain knowledge so they might be able to guide the modeling process and the interpretation of the results. We ran models for $d = 0, d = 1,$ and $d = 3$ to test whether adding or subtracting dimensions of social space significantly altered the conclusions reached in the $d = 2$ models. It did not. The choice of dimension did not affect the conclusions of Chapter 2.

If the goal of the analysis were to predict social ties, perhaps a different choice for the dimension of social space would be appropriate. One method, as suggested by Hoff (2005), is to assess the effect of dimension on the model by cross validation. Splitting the data set into subsets and then choosing the dimension $d$ which maximizes the predictive performance of the model on the subsets left out of the analysis might achieve better predictions than merely choosing a dimension that made social space easily interpretable visually.

Another option would be to use AIC or BIC to assess which dimension $d$ gives rise to parameter estimates that maximize the likelihood of the data subject to a penalty on the number of parameters in the model. The deviance information criterion (DIC) is another model assessment tool (Spiegelhalter et al., 2002) that could be used to select the dimension of social space giving rise to the best fit of
the existing data. The DIC is attractive because it uses the posterior draws from the MCMC instead of parameter point estimates which are used to assess model fit in the AIC and BIC methods.

A stochastic search variable selection (George and McCulloch, 1993) algorithm could be used with a mixture hyperprior on the the variances associated with the positions in each of the potential dimensions \(d = 1, 2, 3, \ldots\) of social space.

The priors currently used on the positions for each dimension (see Section 2.3.1) are independent normal: \(z_i \sim N(0, \sigma^2_{z_i})\) with \(\sigma^2_{z_i} \sim IG\left(\frac{1}{2}, \frac{1}{2}\right)\) for \(i = 1, \ldots, d\).

If, for example, \(\sigma^2_{z_3} = 0\) and \(\sigma^2_{z_1}, \sigma^2_{z_2} \neq 0\), then the dimension of social space would necessarily be \(d = 2\). By changing the hyperprior for \(\sigma^2_{z_i}\) to be a mixture distribution such that a point-mass of probability \(p_i\) was placed at 0 with the rest of the weight \((1 - p_i)\) on a proper prior distribution, the data could then be used to effectively “choose” the dimension \(d\) of social space. The hyperpriors would need to be constructed so that if \(\sigma^2_{z_i} = 0\), then \(\sigma^2_{z_{i+1}}, \sigma^2_{z_{i+2}}, \ldots = 0\). For this algorithm to give sensible results, the priors for the point-mass weights \(p_i\) would have to be carefully chosen, as well as the non-zero component of the hyperprior distribution for \(\sigma^2_{z_i}\).

### 3.4 Innovations

#### 3.4.1 Analysis of pairwise effects

The addition of the unobserved social space to a social network model for predicting ties between actors allows for third-order dependence patterns (see section 2.2.2) such as transitivity and clusterability (Wasserman and Faust, 1994) to be captured in the model. It also allows for the modeling of idiosyncratic “likes and
dislikes” of the actors not captured by the other terms in the model. The inclusion of social space in the social network model allows for a small adjustment in the predicted probability of being together for each pair in the model. On the log-odds scale these adjustments are exactly the pairwise inner products of the positions in social space. For example, if an elephant $i$ had a special affinity for elephant $j$ that could not be explained by a kinship covariate, this affinity would appear in their positions in social space. The angle between these two positions would be small which would induce a positive inner product/pairwise effect.

Table 3.2: The posterior inner products for the various models generally decreased in size as more covariate information was added to the models. The first rows of the dry and wet season tables show the posterior mean absolute value of the size of these pairwise effects. The second rows show the posterior mean. The third and fourth rows show 95% highest posterior density intervals for the inner products and their associated effects in the model in terms of changes in predicted probabilities.

<table>
<thead>
<tr>
<th></th>
<th>Null</th>
<th>DNA-only</th>
<th>ModaSis</th>
<th>Full kinship</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dry Season</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post mean abs val</td>
<td>0.40</td>
<td>0.32</td>
<td>0.16</td>
<td>0.17</td>
</tr>
<tr>
<td>Posterior mean</td>
<td>0.01</td>
<td>0.03</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>95% interval</td>
<td>(-1.11, 1.14)</td>
<td>(-0.88, 0.93)</td>
<td>(-0.44, 0.52)</td>
<td>(-0.46, 0.53)</td>
</tr>
<tr>
<td>Prob $\Delta$</td>
<td>(-0.27, 0.21)</td>
<td>(-0.22, 0.19)</td>
<td>(-0.11, 0.11)</td>
<td>(-0.11, 0.12)</td>
</tr>
<tr>
<td><strong>Wet Season</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post mean abs val</td>
<td>0.52</td>
<td>0.47</td>
<td>0.17</td>
<td>0.14</td>
</tr>
<tr>
<td>Posterior mean</td>
<td>0.04</td>
<td>0.02</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>95% interval</td>
<td>(-1.31, 1.56)</td>
<td>(-1.21, 1.32)</td>
<td>(-0.45, 0.54)</td>
<td>(-0.38, 0.47)</td>
</tr>
<tr>
<td>Prob $\Delta$</td>
<td>(-0.29, 0.17)</td>
<td>(-0.27, 0.15)</td>
<td>(-0.09, 0.08)</td>
<td>(-0.07, 0.07)</td>
</tr>
</tbody>
</table>

The general model (Equation 3.1) is basically a logit model for a social network with both fixed and random effects. The addition of social space and its associated $z_i'z_j$ term gives structure to the residuals in the model. By analyzing plots of social space we are able to visually assess the pairwise residuals for our binary data. This visual assessment can lead to the identification of covariates (see section 3.3.1) to
add to the model. Table 3.2 is a numerical assessment of these residuals between pairs of elephants.

The inner products between positions in social space for all pairs of elephants averaged nearly zero for both the wet and dry seasons in all four models we tested (Models 3.7-3.10). However, the mean size in absolute value of the inner products ranged in the wet season from a high of 0.52 in the Null model (which lacked any covariate information) to a lowest absolute size of 0.14 in the Full kinship model. In general—and as expected—the more covariate information added to the model, the more the absolute size of the pairwise effect decreased.

Highest posterior density intervals (95%) were created for these pairwise effects for each season and model. In the dry season Null model, 95% of the posterior inner products were between $-1.11$ and 1.14. These intervals showed a trend of shrinking towards zero in both seasons as more covariate information was added to the models as the 95% interval for the Full kinship dry season model was $(-0.11, 0.12)$.

The practical effect on the models of the pairwise effects is shown in the last rows of Table 3.2. The 95% regions of inner products were converted from their original log-odds scale into changes in the probability of a pair of elephants being together as predicted by the model with and without these effects from social space. For example, in the Null model in the dry season, the probability of the “average” pair of elephants being together was 0.62 whereas an extreme negative pair (with an inner product of -1.11 in log-odds) had a predicted probability of 0.35 of being together, and an extreme positive pair (with an inner product of 1.14) had a predicted probability of 0.83 of being together. When all the kinship information was incorporated into the model, the pairwise social space effect caused a change
in probability of between −0.11 and 0.12. This practical effect was even smaller in the wet season (between −0.07 and 0.07, down from between −0.29 and 0.17 in the Null model), indicating that the kinship covariates did a good job in both seasons in modeling elephant social behavior. In sum, social space was less important or influential when more covariates were included.

### 3.4.2 Pickiness

Returning to the example in section 3.1.1 of the 2-dimensional “marriage” social space, if an individual were an extreme night-owl and an especially extravagant spender, his or her position in social space would be in quadrant I far from the origin. If a potential mate also had a position in quadrant I of the social space, the inner product between the pair would likely be a high, positive number. Similarly, the inner product between an individual in quadrant III and the extreme night-owl spender would be a large, negative number. If the potential mate were in quadrant II or IV, any deviation from 90° in the angle between the positions of the pair in social space would be magnified by the extremeness of the extravagant night-owl’s position, causing their inner product be far from zero, either positive or negative. In general, an individual with an extreme position in social space far from the origin would typically have either very high positive or very negative inner products with other individuals, and would therefore have rather high probabilities or low probabilities for marriage compatibility.

An individual whose position in the social space was near the origin would be somewhere in the middle of the Early Bird-Night Owl continuum and would have mid-range, moderate spending habits. The inner product of this individual’s position in social space with another individual would typically be close to zero. This
type of individual would have no strong preferences or issues with compatibility in marriage and would typically have mid-range probabilities of marriage in the hypothetical model.

We call the propensity to have very high positive or very negative inner products with other actors’ positions in social space pickiness. In our example, (see Figure 3.1), Paris Hilton, as an extreme night-owl spender, is picky. The farther an actor’s position in social space is from the origin, the higher that actor’s pickiness is.

For an actor \( i \) in the social network model, we define \( i \)'s pickiness, \( \varphi \), to be the inner product of its position with itself. In other words, an actor’s pickiness is its position vector’s magnitude squared:

\[
\varphi_i \equiv z_i' z_i = \|z_i\|^2.
\]  

(3.11)

For the set of actors \( \{1, 2, \ldots, n\} \) in a social network, their pickiness is the vector:

\[
\Phi = (z_1' z_1, z_2' z_2, \ldots, z_n' z_n) = \text{diag}(Z' Z).
\]

(3.12)

**Elephant pickiness**

The more picky an elephant was, the larger (in absolute value) were the inner products of her position in social space with the other elephants’ positions. This means that the parameters in the model did a poor job of predicting a picky elephant’s interactions with other elephants. If an elephant had a position in social space near the origin, we could infer from the model that she had no idiosyncratic “likes or dislikes”, i.e., her interactions with the other elephants were very close to what the other parameters in the model predicted for her.
Figure 3.13: Pickiness decreased as more covariate information was added to the model. The elephants were generally more picky in the wet season than in the dry season in the Null (No Gen) and DNA-only model. The elephants were equally picky in both seasons when the Mother/Daughter and Sisters indicators were added to the model.

Figure 3.13 shows boxplots of the pickiness for the Family AA (Equation 3.12) in both the wet and dry seasons for all four models. Family pickiness decreased in both seasons as kinship information was added to the models. In the Null model and the DNA-only model, which did not include the Mother/Daughter indicator or the Sisters indicator, pickiness was generally higher in the wet season. One reason for this was that the Mother/Daughter and Sisters indicators were more significant covariates in the wet season than in the dry season (see Figure 2.7 in
Chapter 2). Without these covariates in the model, larger social space pairwise effects were needed to predict the elephants’ interactions.

In the ModaSis and the Full kinship model, both of which included the Mother-Daughter and Sisters indicators, elephant pickiness was small during both the wet and dry seasons. These models did a good job in modeling the elephants’ social interactions. In general, the lower the pickiness the better the model fit.

![Box plot showing pickiness of elephants during the dry season in the Full kinship model.](image)

**Figure 3.14:** Pickiness of the elephants during the dry season in the Full kinship model.

Figure 3.14 shows the pickiness of the individual elephants of Family AA during the dry season for the Full kinship model. Figure A.8 in Appendix A shows individual pickiness in the wet season for this model. In both the wet and dry seasons...
there were no especially picky elephants. No elephant required an extreme position in social space to generate large (both positive and negative) social space pairwise effects to compensate for not being well-predicted by the other parameters in the model. All elephants in Family AA generally conformed to the predictions of the Full kinship model. We can conclude from these plots that the kinship covariate information predicts the probabilities of elephant interactions about equally well for all the elephants in Family AA.

Visual inspection of the social space plots (Figures 3.9-3.12 and Figures A.1-A.4 in Appendix A) can lead to the identification of clusters of outlier pairs and can provide guidance for adding covariates to the model. A statistician or expert researcher might ask, “What do those pairs have in common, and how will knowledge of this help us in our modeling?”

Similarly, an examination of individuals’ pickiness can lead to the identification of outlier individuals. Even though all our data in this social network model are on pairs of elephants, we can visually detect outlier individuals by their pickiness in social space. Asking, “Why does the model not fit this particular individual?” might lead to further insight into the processes governing the social network.

These innovations in the practical use of social space in social network models make the fitting of these models much more attractive and intuitive, and could potentially lead to their wider application in fields beyond statistics.
Chapter 4

Fraud Detection in Calling Networks

4.1 Introduction

Local residential customers of a large telecommunications company create links with other local customers by making telephone calls. These telephone numbers and the calls between them constitute a social network. We use social network analysis to study the pattern of links between telephone numbers. However, these links change over time while the statistical techniques used for analyzing these networks were designed for static networks. In this chapter we address the problems of which statistics are useful in describing a dynamic network, whether these statistics show dynamic behavior, and whether or not changes in these statistics can successfully identify anomalous behavior in the network.

According to Cox et al. (1997), the telecommunications industry losses to fraud in the United States alone were as much as $1 billion per year. Since fraud is such a large problem in calling networks, we apply social network analysis towards detecting fraudulent behavior.
### 4.1.1 Calling networks

A telephone call between two residential telephone customers forms a dyad. The nodes of the dyad are the two telephone numbers. The link between the nodes is the call from one node to the other. This link is directed from the telephone number (TN) originating the call to the TN receiving the call. This dyad is connected for the length of the call. Once the phone call is over, the relationship is over. The end of the telephone call between the two customers literally ends their dyadic connection.

Our data set under study consists of many transitory dyads. For example, TN A called TN B and formed a dyad. Once TN B hung up, the dyad was terminated. Later the same day, TN B called TN A and formed a new dyad in the opposite direction. The next day TN A called TN B and reformed the dyad that existed briefly the day before. The data set is, in effect, a stream of dyads labeled with their time, date, duration, and direction.

A network can be induced from these dyads by defining a window of time and aggregating all dyads that have existed within the time window. For example, the smallest network we could create is a one second network based on all dyads extant during one specific second. Another network could be created by considering all nodes and links active during a particular minute. The May 11, 11:11:11 one second network and the May 11, 11:11 one minute network described above have identical graphs (not shown). Only one telephone call was made during those time windows, so the graph is of two telephone numbers and the telephone call linking them.

The network for all calls initiated between 11:00 and 11:59 on May 11, 2006 consists of 52 telephone numbers and the 32 unique calls between them. Figure 4.1
Figure 4.1: Graphs of four successively larger calling networks. Top left: the one-hour network. Top right: the one-day network. Bottom left: the one-week network. Bottom right: the one-month network.
shows the graph of this network which consists of the 29 telephone numbers initiating a call during the one-hour time window, the 28 telephone numbers receiving such a call, and the 32 directed links between these telephone numbers/nodes (some numbers both initiated and received calls to/from multiple numbers).

The one-day network for May 11th is of course bigger, consisting of 672 telephone numbers and the 547 unique calls between them. Figure 4.1 shows this network with its longer chains and more connected components. The one-week network of the aggregation of all dyads existing from May 11 through May 17 has much larger connected components and fewer unconnected dyads. Just one call during the week can connect a previously unconnected dyad with part of the larger network. The graph of this network of 1289 telephone numbers and 1705 calls between them is shown in Figure 4.1. The month-long network (Figure 4.1) is entirely connected and consists of 1933 telephone numbers and 3919 unique calls between them.

Many calls during a time window are repeated. For instance, a call from one telephone number to another might be repeated several times in a day. Cortes et al. (2003) use repetitions of calls and call duration to define weights for the links between nodes in the graph. We, however, ignore repeated calls during the same time window to focus instead on the presence or absence of a link between nodes in the creation of networks for a given time window.

4.1.2 Dynamic calling networks and Communities of Interest

Dynamic networks are networks in which the nodes and the links between them appear and disappear over time, and a framework for studying the evolution of
these networks over time was developed by Cortes et al. (2003). This framework was extended and made more rigorous by Hill et al. (2006). Cortes et al. (2003) represent the graph of a dynamic network as an exponentially weighted average of new network data with the graph from the previous time step. Exponential updating creates a smoothed view of the network and allows for real-time updating as new data becomes available. Edges with low weights are trimmed from the graph so that the graph does not grow too large and computationally burdensome. This representation of a dynamic network is an effective summary of the current state of the network, but does not have much use in our application in this chapter.

The subgraph of a node and all the nodes linked to or from that node define a basic structure called Communities of Interest (COI), though the definition of COI is application dependent (see Cortes et al., 2002). A graph can be represented by a union of COI. When used by Cortes et al. (2003) to represent a dynamic graph, the COI are trimmed to keep only the highest weighted edges and are sometimes augmented with pseudo-links to capture network activity inferred to be missing from the data set. COI can also be defined by its degree of recursion. For example, a COI of degree two would include the original central node \( X \) and all the COI of the nodes linked to or from \( X \). The COI of degree \( \infty \) would include all nodes connected to nodes connected to nodes, \( \ldots \), connected to \( X \), which is the maximal connected component including node \( X \).

Agarwal and Pregibon (2004) extend the use of COI by fitting separate models for each of the COI. This allows for the scaling to large networks of methods previously applicable only to small networks. It also better accounts for heterogeneity in a large network by flexibly applying many different models to the various, more homogeneous subgroups of the network.
4.1.3 Telephone fraud

Cortes et al. (2002) discuss two variations of fraud within a calling network. The first is when a subscriber opens an account, accumulates lots of telephone charges, then refuses to pay the bill. This often occurs with toll-free numbers. The fraudsters and their friends and family use the toll-free number to make free calls until the telecommunications company cancels the service because of a delinquent account. The fraudsters then open a new account and resume their fraudulent behavior. To identify this type of fraud as early as possible, each new account is assessed a score for risk of fraud based on its pattern of connectivity to the rest of the network. Because of the tendency of fraudsters to share their discovered vulnerabilities with their friends and family, clusters of fraudulent telephone numbers are often seen in networks.

The second variation of fraud is identity theft, where fraudsters use an otherwise legitimate telephone number for fraudulent activity. The new calling patterns on such accounts can be matched with patterns of known, discontinued fraudulent accounts by comparing overlaps in the COI for the suspicious account and the COI for known fraudsters. Accounts with COI that have many overlapping nodes with fraudulent COI are flagged for further investigation. In this chapter we discover a new variation of fraud within the calling network using a different method.

4.2 Data from a Calling Network

Our data set consists of about 2,000 telephone numbers (TN) from an area code in the northeastern United States and all the calls made between these numbers
for a one month period in May 2006. However, data from May 4, 2006 are completely missing from the data set, i.e., there are no data for any calls on May 4th, but complete data exist for all the telephone numbers in the data set for each of the other days in May 2006. These 2,000 telephone numbers belong to residential telephone customers of a large telecommunications company and represent the largest connected component of the directed graph during the one month time period under study. A total of about 30,000 calls were made between these customers during May 2006.

A connected component is a subgraph for which there is a path between any two nodes of the subgraph, ignoring direction of the link. Abney et al. (2005) provide motivation for studying connected components within large calling networks, namely it restricts investigative focus to COI (of degree $\infty$) and makes computation easier. Since these 2,000 telephone numbers form a connected component, the adjacency matrix describing the one-month network has no empty rows or columns, i.e., every node has at least one link to another telephone number in the data set.

Many of the calls in the network are calls from a telephone number to itself. We refer to these self-calls as voicemails since they can be thought of as calls to check the voicemail system for that telephone number. When a number calls itself it is as if it is connecting to its voicemail messaging system. These voicemail calls were counted and then ignored and played no part in the network. We did not consider the implications to the network of nodes having links to themselves, otherwise known as loops.

Approximately 40% of the calls in any given day were repeated calls, i.e., the second, third, fourth, or more calls from TN A to TN B. We refer to the first call
from TN A to TN B as a unique call. All subsequent calls from TN A to TN B are referred to as repeated calls. We counted the number of repeated calls and then ignored them too. Each link in the graph of the network represents at least one directed call between two telephone numbers. A call from TN A to TN B and another call from TN B to TN A were treated as two unique calls.

4.2.1 Visualizing data

Visualization of a network can be a useful start to exploratory data analysis. To visualize the networks we drew the graph of the networks using Graphviz software (North, 2004) on the Unix/Linux command line. Given a list of dyads, the Graphviz program draws either directed or undirected links between the nodes of the dyads. The nodes in the graph are positioned in a layout such that their distances in the graph correspond to the path length between nodes. These “optimal” layouts are achieved using multidimensional scaling on the path lengths (Buja et al., 2008). The closer one node is to another node in terms of path length, the closer together those nodes will be in the drawn graph.

Figure 4.2 shows the graphs for four non-overlapping weeks of the calling network. We drew these graphs by first aggregating all the dyads in the first week (May 5-11) to form a one-week network. The voicemails in the network were counted and then deleted since we ignored loops. The repeated calls were also counted and ignored in order to focus on the unique calls. The unique calls determined the dyads of the network, and these dyads were drawn by Graphviz to form the graph of the May 5-11 network. Note that for security and confidentiality reasons, the labels on the nodes in the graphs are blank. Ordinarily each node would be labeled with its telephone number. Also note that since the data set
Figure 4.2: Four successive weeks of streaming call dyads are aggregated into weekly networks. The graphs of these four networks show varying degrees of connectedness with the third week graph (lower left) being completely connected.
was subsetted to show the largest connected component, one of these one-week networks must be completely connected, and it happened to be the network of the third week (May 19-25).

Similar graphs were drawn for the one-day networks for each day in May. These graphs show varying degrees of connectedness, with the daily networks of May 22-23 being the most connected. These graphs are too numerous to display within the text of this chapter, but can be seen in Figure 4.15 in Appendix 4.6. In our experience, visualizing a sequence of ten or more graphs is too cumbersome and overwhelming to be a very effective analytical tool. However, viewing a selection of graphs can be helpful in getting a sense of the data.

4.2.2 Difficulties in the data

Our goal is to detect anomalous behavior in a continuous stream of dyadic data. Traditionally, these dyads are aggregated to form a sequence of static networks. Comparisons between the graphs of these static networks can be difficult due to the high number of effective dimensions for each graph. For example, the number of possible directed graphs with $n$ nodes is $2^{(n^2-n)}$ since every node could have a directed edge to and/or from each other node in the graph. A simple parametrization of this graph using the $p_1$ model of Holland and Leinhardt (1981) (see Equation 1.1) would require a productivity and attractiveness parameter for each node for a total of $2n$ parameters. Even more parameters would be needed if the $\binom{n}{2}$ dyads were not assumed to be independent. Alone, visualization techniques of sequences of static graphs of the network are insufficient and non-intuitive to handle this high-dimensional data structure, and the problem is compounded when the graphs are dynamic due to the constant appearance and disappearance of nodes.
and edges in the graphs.

Our strategy is to first aggregate the stream of dyads to form daily networks. We summarize each static daily network using global statistics of the network and summary statistics of the local behavior in the network. This effectively reduces the dimension of our networks. We then track those statistics over time in order to detect anomalies.

4.3 Detecting Anomalous Behavior

4.3.1 Summaries of the network

For each day in May (except May 4th), we aggregated the stream of dyads for the day into a one-day network. For each of these one-day networks we summarized the network by counting certain global attributes such as the total number of calls, the number of telephone numbers/nodes sending or receiving a telephone call to or from another telephone number in the data set, the number of voicemail calls sent and received by the same telephone number, the number of repeated calls from one number to another number, and the number of unique calls during that day. The number of unique calls was the total number of calls minus the voicemails and minus the repeated calls.

Figure 4.3 shows four of these summary statistics for the one-day networks for the month of May. The total number of nodes and the number of unique calls rose and fell with each other, seemingly in tandem. Both showed an increase from May 20 to May 23 beyond the normal daily fluctuation, with a peak between May 22nd and 23rd. The number of voicemails in the network was low and stayed relatively constant throughout the month. The number of repeated calls showed
more fluctuation, but did not exhibit any particular pattern. The increase in the number of nodes and unique calls did not correspond with an increase in the number of repeated calls.

Using non-statistical methods we investigated this increase in calling activity. An examination of local newspapers during the days surrounding the peak calling activity of May 22 and 23 revealed nothing of obvious importance. There were no local elections or major events (as reported in the local papers) that might provoke cascades of conversation over telephone lines. The nationally televised singing competition *American Idol*, which spurs voting by telephone that often
swamps the national calling networks (Becker, 2006), held its season finale on May 23rd. However, calls to a national, toll-free number was an unlikely cause for the increase in local telephone calls, and the time of the telephone voting (or potential subsequent phone conversations) did not correspond with the times of the vast majority of the phone calls (see Figure 4.12). The biggest event in the papers was a children’s parade organized by a local Chamber of Commerce. In short, nothing reported in the papers or appearing on TV indicated a departure from the residential telephone customers’ quotidian that would account for the sudden increase in the number of local telephone calls. A more likely cause will become evident later in this chapter.

Figure 4.4: A and B form a mutually connected dyad. The pair C and D are an asymmetric dyad. A and C (also A and D, B and C, and B and D) form an “empty” dyad since no link exists between them.

A different approach to describing or summarizing the daily networks is to use the number of mutually connected dyads and asymmetric dyads appearing in each one-day network. Two telephone numbers are mutually connected if each number initiates at least one call that is received by the other telephone number. Two telephone numbers connected by a telephone call in only one direction is an asymmetric dyad. For example, if TN A calls TN B, and TN B also calls TN A during the same day, A and B form a mutually connected dyad. If TN C calls
TN D, but TN D does not call TN C that day, then the pair form an asymmetric dyad (see Figure 4.4).

![Graph showing unique calls, asymmetric dyads, and mutual dyads](image)

**Figure 4.5**: Total number of unique calls, asymmetric dyads, and mutual dyads by day for the one-day networks.

Figure 4.5 shows the number of mutually connected dyads ($N_{md}$), the number of asymmetric dyads ($N_{ad}$), and the total number of unique calls ($N_{UC}$) for each one-day network during the month of May. It is worth noting that the number of unique calls is a function of the number of mutual and asymmetric dyads.

$$2 \cdot N_{md} + N_{ad} = N_{UC}. \quad (4.1)$$

The number of mutually connected dyads was relatively constant throughout the
month, whereas the number of asymmetric dyads showed an increase from May 20 to May 23 with a peak between May 22nd and 23rd. As can be seen in Figure 4.5, this increase in asymmetric dyads is largely responsible for the peak in the number of nodes and the number of unique calls.

4.3.2 Triad census

According to Couzin et al. (2005) global structure of networks can often be inferred by local structure, i.e., by the types of dyads and triads in the network. In section 4.3.1 we discussed two of the three types of dyads: mutually connected and asymmetric dyads. TN A and TN C in Figure 4.4 form the third category of dyad, namely, the empty or null dyad. No link—neither mutual nor asymmetric—exists between TN A and TN C, though one could consider TN A and TN C to be connected by a null link.

A triad is a subgraph of three nodes in a network, and as such consists of three nodes and up to six directed links between the nodes. Hence, there exist $2^6 = 64$ possible configurations of the triad. However, many of these configurations are isomorphic to each other if we disregard node labels. Figure 4.6 shows the sixteen isomorphism classes of triads labeled by their number of mutual dyads, asymmetric dyads, null dyads, and a letter to further distinguish the triad if necessary. Wasserman and Faust (1994) define and discuss the sixteen possible types of triads in a network. They provide examples of how the enumeration of triads can help researchers test for the presence of special structural patterns in the network such as transitivity, reciprocity, and differential popularity. The triad census may be regarded as a way of reducing the network to a smaller set of summary statistics, and can be used to test the tendency of such structural
patterns to hold across a network (Holland and Leinhardt, 1976).

\[
N_T = \binom{N_n}{3}.
\]

Moody (1998) developed algorithms to calculate the triad census using simple manipulations of the adjacency matrix of a network.

In the daily networks, the median number of nodes was about 700, which meant that the total number of triads was on the order of \(5 \times 10^7\). By far the most common triad was the empty triad, or the 003 triad in Figure 4.6. The next most common was the 012 triad which consists of an asymmetric dyad and another node unconnected to the dyad. The third most common triad was the
102 triad which consists of a mutual dyad and another node unconnected to the dyad. In the daily calling networks we studied, the vast majority of the triads were of these three, uninteresting varieties which are mainly determined by the number of nodes in the network of the day, not by any special properties of the network or the underlying behavior of the telephone customers. In fact, the three most common types of triads accounted for about 99.9995% of all the triads in the daily networks. In other words, only about 1 in every 500,000 triads in the daily networks were of the other, more interesting varieties.

Table 4.1 shows, as an example, the triad census for the May 11th one-day calling network (see Figure 4.1 for the graph of this network). The graph of this network had 672 nodes and thus \( \binom{672}{3} = 50,351,840 \) total triads. Of these total triads, nearly all of them (50,061,960) were the 003 empty triad, while the vast majority of the rest were the 012 “asymmetric-dyad” triad (213,404) and the 102 “mutual-dyad” triad (76,254). There were only 222 of the remaining thirteen types of triads.

In the May 11th network, of the remaining thirteen categories of triads, seven had counts of two or fewer. These seven rare triad types 030T, 030C, 120D, 102U, 120C, 210, and 300 are the only triads in which a link connects all three nodes of the triad (see Figure 4.6). For the May 11th network, the median number or count of these triads was one. This distribution of these seven triads was typical.
of all the daily calling networks, as the median number of these rare triads for each daily network was also one.

Plots of the counts of the seven rare triads, which only varied from zero to six in the daily networks, can be seen in Figure 4.16 in Appendix 4.6.

Our goal in enumerating each type of triad for each of the daily networks was to identify subsets of triad types that would be useful summaries of the global network structure. The first three types of triads, the empty triad and the two "dyad plus one" triads, were just too numerous to give insight into the behavior of the network. The addition or subtraction of just a few dyads to these networks would cause the counts of these triads to fluctuate wildly. For example, the addition of one asymmetric dyad unconnected to the rest of the \( n \) nodes in a network increases the 003 empty triad by \( 2 \times \binom{n}{2} \) and the number of 012 triads by \( n \). The reciprocation of just one telephone call between an asymmetric dyad unconnected to the other nodes would increase the number of 102 triads by \( (n - 2) \) and would decrease the number of 012 triads also by \( (n - 2) \).

On the other hand, the seven rare triads numbered too few in the daily networks to be useful indicators of network structure. The addition or subtraction of a few links which could complete existing triads would have little practical significance. A network with two mutually complete 300 triads is not much different from a network with, say, five of those completely connected triads.

Of the sixteen categories of triads, there were six types of triads—neither too numerous nor too rare—which showed enough practical variation in the one-day networks to serve our purposes well as indicators of the network structure. These six interesting triads—the 021D, 021U, 021C, 111D, 111U, and 201 varieties (see Figure 4.6 and Figure 4.7)—are the only six triads having exactly one empty or
null link. In the May 11th network, counts of these interesting triads ranged from 18 to 50, with a median of 38. The distribution of these six triads in the May 11th network was typical of the daily calling networks, as the median number of these interesting triads for each daily network was 37.

Figure 4.7: Four of the most interesting types of triads. From left to right: the 021D triad, 021C, 111D, and 201.

These triads are interesting, not only in that they showed a reasonable amount of daily variation in our data set, but also because they are meaningful indicators of transitivity, structural holes, and other properties in the network. Briefly, the 021C, 111U, and 201 triads are examples of intransitive triads which violate the transitive property of: if \( A \rightarrow B \) and \( B \rightarrow C \), then \( A \rightarrow C \). An additional link between \( A \) and \( C \) would “promote” the intransitive 021C triad to become the transitive 030T triad as shown in Figure 4.8. In many social networks which evolve over time, the trend is towards more transitive triads (Davis et al., 1971).

Figure 4.8: The addition of a link between \( A \) and \( C \) to the intransitive 021C triad creates the transitive 030T triad.

The 021D and the 201 triads are two examples of structural holes in the network.
which can provide an opportunity for one node to broker the flow of information between two nodes (Burt, 1992). In the 201 triad, B is mutually linked with both A and C which are not linked to each other. Therefore B can control their communication.

Figure 4.9: Daily counts of four of the most interesting, variable triads.

Figure 4.9 shows the counts of four of these interesting triads (021D, 021C, 111D, and 201) for each of the one-day networks in May.
4.4 Results

Though it is unclear from the time series plot in Figure 4.9 how the counts of the triads 021C, 111D, and 201 for the daily networks change with time, it is strikingly obvious that the behavior of the 021D triad is anomalous. The plot shows that something is very different about the May 22nd and 23rd one-day networks from the rest of the daily networks. The counts of 3558 and 3249 for the May 22nd and 23rd networks were well above the median value of 49 for the daily counts of the 021D triad.

4.4.1 Explanation of the 021D triad

![Figure 4.10](image-url): A five-node network consisting of four asymmetric dyads. This graph has two types of triads: six 021D triads and four 003 empty triads.

The 021D triad consists of one telephone number calling two other numbers, both of which never return the call nor call each other (see Figure 4.7). If, for example, one telephone number made four calls to distinct numbers which did not
return the call and did not call each other, and these four asymmetric dyads were aggregated to form a network, its graph would be as shown in Figure 4.10. Of the ten total triads in this example network, six would be of the 021D variety and four would be 003 empty triads. In this situation where one TN calls $n$ distinct numbers, the number of 021D triads is given by the equation:

$$N_{021D} = \binom{n}{2}.$$  

(4.2)

If the TN called one additional distinct number (for a total of $n+1$ called numbers), $N_{021D}$ would increase by $n$.

### 4.4.2 Identifying cause of anomalous 021D triad spike

Further investigation into the data revealed that nine telephone numbers were registered to the same nine-digit zip code. This indicated that they all belonged to the same residence. These nine suspicious TNs initiated a total of 219 calls on May 22nd and received a total of 0 calls. All of these 219 calls were unique and resulted in the creation of 3497 021D triads. On May 23rd the nine suspicious TNs received no calls but initiated 205 calls. Again, none of these calls were repeated calls, i.e., on that day none of these nine TNs called the same number twice. This unusual calling behavior resulted in the creation of 3147 021D triads in the May 23rd network.

Figure 4.11 shows the COI of degree two for TN “α”, one of the nine suspicious TNs. TN “α” initiated 41 calls, the most out-going calls of the nine suspicious TNs on May 22nd. Of these 41 calls, all were unique and all were received by numbers which did not return the call or call each other. Using Equation 4.2 we calculate that these 41 calls from TN α resulted in the creation of 820 021D triads.
Figure 4.11: The communities of interest of degree two for TN $\alpha$ for May 22nd. The graph shows all the numbers TN $\alpha$ called and all the calls associated with those 41 numbers.

Table 4.2: The interesting triads created by the nine suspicious telephone numbers.

<table>
<thead>
<tr>
<th>Triads</th>
<th>021D</th>
<th>021U</th>
<th>021C</th>
<th>111D</th>
<th>111U</th>
<th>201</th>
</tr>
</thead>
<tbody>
<tr>
<td>May 22nd</td>
<td>3497</td>
<td>33</td>
<td>19</td>
<td>17</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>May 23rd</td>
<td>3147</td>
<td>31</td>
<td>18</td>
<td>13</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

in the May 22nd network.

Table 4.2 shows the number of interesting triads created by these nine suspicious TNs in the May 22nd and 23rd networks. Only the 021D triad shows a dramatic increase, though many of the other triads were created by these nine TNs. None of the six rare, completely connected triad types were created during
these two days by the nine suspicious TNs. Since these TNs created 219 and 205 dyads for the two daily networks, they created on the order of $10^7$ 003 empty triads, $10^5$ 021 asymmetric-dyad triads, and $10^4$ 102 mutual-dyad triads.

4.4.3 Recognizing the anomalous and fraudulent behavior

We suspected that the nine TNs were engaging in some sort of fraudulent behavior due to their similar, unusual calling patterns and their registration at the same residence. The facts presented so far:

1. Initiated many calls
2. Rarely received calls
3. Rarely called the same telephone number twice
4. The telephone numbers called rarely made calls with each other.
5. All nine TNs were registered to the same nine-digit zip code (residence).

Further examination of the calling behavior of these nine suspicious TNs revealed one more interesting fact. The time of day in which the nine TNs initiated calls showed a different pattern than for the other 973 telephone numbers in the data set which initiated calls on May 22nd and 23rd. Figure 4.12 shows the histogram of the hourly calling pattern for all the calls initiated on May 22nd and 23rd. The overlay of the histograms for the calls from the telephone numbers other than the nine suspicious TNs with the calls initiated by the nine TNs shows the striking differences in the hourly calling patterns of the nine suspicious TNs. These TNs initiated calls on May 22nd and 23rd strictly from 9:00AM to 8:00PM. Furthermore, this pattern was consistent for every weekday throughout the month.
The nine suspicious TNs only initiated calls from 9:00AM to 8:00PM Mondays through Fridays, and from 10:00AM to 2:00PM on Saturdays. During the month of May, these 9 TNs initiated no calls on only five days: the four Sundays in the month and on Monday, May 29th which was the Memorial Day holiday.

Together, these facts made the identification of the behavior of the nine suspicious TNs clear. They were telemarketers. Nine telemarketers at the same location made numerous calls between the hours of 9:00AM and 8:00PM Monday to Friday and 10:00AM to 2:00PM on Saturdays to distinct telephone numbers which they had never called before, which rarely returned the calls, and which rarely called each other.
All the telephone numbers in the data set were registered to customers who had signed up for residential telephone plans. Since businesses contracting for telephone service were not eligible for these residential service plans and rates, the nine telemarketers were contractually obligated to use an alternative telephone service designed for businesses. The use of the residential telephone lines by the nine telemarketers was a violation of their service agreement and constituted fraud against the telecommunications company. Though other clues were present, only the massive generation of the 021D triads by the nine telemarketers in the May 22nd and 23rd networks alerted us to their fraudulent behavior.

4.4.4 Daily networks without nine telemarketers

The daily networks with the dyads generated by the nine telemarketers removed were not much different from the full networks, except for the spike in the 021D triads on May 22nd and 23rd. The six interesting triad types continued to show variability from day to day as shown for four of these types in Figure 4.13. The dynamic nature of these triads might be worthy of further study.

The global summary statistics of the daily calling networks were only marginally affected by the removal of the nine telemarketers from the network. Figure 4.14 shows the daily number of nodes, unique calls, repeated calls, and voicemails for the one-day networks without the nine telemarketers. The increase in nodes and unique calls from May 20 to May 23 is still evident, though not as pronounced as when the dyads created by the nine telemarketers were included (see Figure 4.3). The nine telemarketers did not create any additional voicemails, and their contribution to the number of repeated calls was negligible. Compared to the changes in the daily counts of interesting triads, these global summary statis-
Figure 4.13: Daily count of the four most interesting, variable triads without the 9 telemarketers included in the network.

...tics were relatively robust to the inclusion or exclusion of those nine telephone numbers.

4.5 Conclusions and Future Research

Given a stream of dyadic data, we were able to identify meaningful anomalies and changes in the structure of the network by tracking the counts of the sixteen possible triad types. This led to the identification of nine fraudulent TNs whose behavior was a violation of their service agreement.
We showed that tracking local behavior of a network over time gave us a useful summary of the larger scale network activity. Changes in the absolute and/or relative frequencies of triads in a network can be used to infer changes in the underlying structure of the network and can be correlated to real-life scenarios such as the activity of telemarketers. Since the triad census is relatively cheap to compute, we believe that this method of summarizing a dynamic network has potential for widespread application.

Future research directions might include building a model to predict the ways in which the triad counts change as dyads are added to the network. Such a model
might use a birth/death process for each type of triad in the dynamic network, allowing not only for the testing of theories on the past behavior of network, but also for predictions of future network behavior. Another direction might be to use models for the propagation of diseases to study how information might propagate through the network via telephone calls.

4.6 Appendix: Daily Graphs and Triad Plots

Figure 4.15 shows the graphs of each one-day network from May 8th to May 31st. The dyads for each day were aggregated, and then only unique calls were used to form the graphs. The graphs from the first week of May were omitted due to space constraints and for being noninformative.

The seven rare types of triads in the daily networks only varied from zero to six (see Figure 4.16). Removing the nine telemarketers from the data set did not affect the counts of any of these triads for any of the one-day networks.
Figure 4.15: Daily graphs for May 8-31.
Figure 4.16: Daily counts of the seven rare triads: 030T, 030C, 120D, 102U, 120C, 210, and 300. These were unaffected by the nine telemarketers.
Chapter 5

The DataSpheres Method for Change Detection in Network Data Streams

5.1 Introduction

In the previous chapters we studied methods for analyzing and understanding streams of dyads in dynamic networks, and how to detect changes or anomalous behavior in these streaming data sets. In this chapter we continue that theme by developing and applying methods for detecting managerially significant changes in the underlying multivariate distribution of massive, rapidly accumulating data sets. We conclude this chapter by applying a new method for change detection in real data streams that is simple for data managers to implement and interpret.

Data streams have become ubiquitous in many industrial and scientific applications. IP networks, financial institutions, and environmental sensors are just a few sources that frequently generate complex data at short time intervals, and it is now common for statisticians to encounter and analyze these massive streaming data sets. The rapid rate of accumulation, the large size and dimensionality, and the often dynamic, unknown underlying distributions of these data streams pose
unique challenges for scientists and researchers who wish to analyze them. Furthermore, the size of the data sets often dictate that the data cannot be stored—even temporarily—and can therefore be accessed only once. The new data must be processed immediately before being erased forever. Any method for analyzing such data streams must account for these special challenges.

Change detection for multivariate data streams is similar to the problem in industrial statistics of quality control for processes which generate multivariate outputs. Detecting when a process exhibits special variation beyond the historical patterns (Shewhart, 1931) is at the heart of statistical process control and has been led to multivariate control charts in, for example, Lowry *et al.* (1992) and Liu (1995).

The Hotelling $T^2$ statistic is a generalization for multivariate data of the univariate Student’s $t$ statistic (Hotelling, 1931). Its application is widespread in multivariate process control (Stoumbos *et al.*, 2000), but its use should be limited to cases when the data are actually multivariate normal—or nearly so—which is rarely a reality in industrial data sets. Section 5.4.1 below further discusses and tests a change detection algorithm for multivariate data using the Hotelling $T^2$ statistic.

Since streaming data change over time we may want to compute and update an exponentially weighted moving average (EWMA) $Y_t$ of the current state of the data using a linear combination of the old average $Y_{t-1}$ and the new streaming data $X_t$:

$$Y_t = \theta X_t + (1 - \theta) Y_{t-1}, \quad 0 < \theta \leq 1 .$$  \hspace{1cm} (5.1)

This EWMA $Y_t$ puts a weight of $(1 - \theta)$ on the previous time step’s smoothed average and could be used to track slow evolution in the center of the data stream.
In the context of intrusion detection of computer systems using anomaly detection, Ye et al. (2003) use an EWMA of proportions of event types to test against the norm profile. They use a $\chi^2$ test to detect statistically significant differences in the current EWMA proportions of event types in a data stream compared to the typical profile proportions of event types.

To improve database query processing, Berchtold et al. (1998) developed an indexing method for high-dimensional data spaces which divided the space into hyperpyramids and then cut the hyperpyramids into parallel slices. This idea of partitioning data spaces into hyperpyramids based on the directions of the maximum deviation of the data forms the basis for the method (Johnson and Dasu, 1998) used to discover which subpopulations within a data set are different across two data sets. Johnson and Dasu (1999) extend their method of DataSpheres to recursively partition the data set into hyperspheres and hyperpyramids. The method of DataSpheres, as applied to data streams, will be discussed in Section 5.3.1.

5.2 Solution Requirements

In a typical scenario, a data stream allows those monitoring it only a single glimpse of the data at a given time point. A large IP network data stream results in several gigabytes of data during just one five-minute slice, and therefore only summaries of this data can be stored for further analysis. Most of the research to date on how to manage large streams of data has focused on univariate analysis (see Babcock et al., 2002), whereas most data streams today are multivariate.
5.2.1 Special challenges

Scale and speed

When new data from a stream arrives, it is typically stored in the buffer, or temporary storage area. After one time interval new data arrives and replaces the old data in the buffer, resulting in loss of the old data (Cranor et al., 2003). For this reason, the data must be quickly summarized before it is lost. Therefore, only computationally inexpensive summaries of the data—such as counts, first, second and third moments, quantiles (assuming the cut-off points have already been defined), and sums or products of attributes—can be considered when the data streams are very large and accumulate very rapidly.

High dimension and correlation

Multivariate histograms can be used to summarize the data. The frequency of data points are recorded for each fixed user-defined region within the range of data. The smaller the regions, the more bins of data, and the closer the histogram summary approximates the data set.

In multivariate data sets, the number of bins in a standard histogram explodes as the number of dimensions increases. If each attribute in a $p$-dimensional data set were partitioned into $q$ classes, the total number of bins in the histogram would be $q^p$. For this reason, the method of approximating a multivariate distribution with a conventional histogram can not be scaled up to high dimensions, although some methods, such as $kdq$-trees (see Dasu et al., 2006), have been developed to control the number of bins by imposing minimum thresholds on the number of data points in each bin and on the volumes of the bins.
Another issue with high dimensional data is the often unknown and complex correlations between the various attributes in a data set. Attributes along different dimensions can be highly correlated, rising and falling in lagged tandem so that treating each dimension as independent will not suffice.

**Heterogeneous, unknown underlying distributions**

Data streams are often noisy. An attribute can vary widely but have no practical significance to the analysis. Methods that rely on multivariate normal assumptions are inadequate. In fact, making parametric assumptions about the real-life processes generating complex, dynamic data streams is problematic.

Also, due to the dynamic nature of the unknown, underlying distributions generating the data streams and the dubious quality of the data, one outlier data point should not inordinately affect the summaries of the data stream. The test statistics used to determine whether or not a change in the data stream has occurred should be robust to outliers. To this end we focus our efforts on non-parametric solutions robust to outliers.

### 5.2.2 Discussion of previous solution attempts

Partitioning data into pre-defined classes is known as *stratification*. Many database management systems use stratification for efficient querying and retrieval of data (Ullman, 1988). These classes are known before any data is seen or stored. An example of this strategy is the “data cube” Gray et al. (1997) used to aggregate data efficiently in order to construct histograms, facilitate data mining, and respond to database queries.

Other partitions are data driven. One example is to partition the data based
on a grid of cut points in each dimension. The cut points are often chosen from quantiles of each attribute. In order to avoid the explosion in the number of partitions when the dimension gets large, \textit{kdq}-trees (Dasu \textit{et al.}, 2006) stop partitioning the data set when the number of points within a bin reaches a user-defined threshold, or when the cut points defining the bins get too close to one another. This eliminates the automatic creation of empty bins which plague many stratifications. Another partitioning scheme driven by the data is one induced by clustering methods (Fraley and Raftery, 1998). However, clustering is computationally expensive.

\section*{5.3 DataSpheres Solution Method}

The AT&T data streams that we consider are very large and highly multivariate. They accumulate rapidly and often are so large that memory limitations dictate only one-time access to the raw data. Each new data point that arrives must be quickly classified. We achieve this through the method of DataSpheres (Johnson and Dasu, 1998) in which each observation is summarized by two labels: depth, and direction of greatest deviation. Bins are then defined in terms of these labels, so that when each new datum arrives, the count of one of the bins is increased by one. Thus the stream of multivariate data is summarized by a set of bin counts.

Each point of the data stream lies in a $p$-dimensional space. When aggregated over some time interval the entire stream makes a cloud of points in that $p$-dimensional space. For any aggregation of the data stream we define a center of the cloud of points. This center is either the mean or one of the various spatial medians (see Small, 1990). Then for a scientifically appropriate metric we calculate the distance from each point to the center. We partition the $p$-
dimensional space into shells based on quantiles of the distances from the center. The points nearest to the center are in the innermost shell, or hypersphere, of the data cloud while the points farthest from the center are in the outermost shell or hypersphere. Data points get labeled by the particular hypersphere in which they reside. In this way, the $p$-dimensional data are partially summarized by a 1-dimensional value.

The second label assigned to the points in the data stream is the direction of greatest variation. Of the $p$ components of a data point, expressed in standard units, one of the components will be largest in absolute value. We then label the datum with the signed component of greatest variation (in standard units). Thus if the sixth component had the largest absolute z-score, the label would be $+6$ or $-6$, depending on the sign of the z-score. This flags the dimension that is most unusual, and whether the point is positive or negative in that most unusual component. The values of the other $p-1$ components of the point are then ignored and the observation is classified into the bin that corresponds to its depth label and the label showing its signed component of greatest variation. This means that each $p$-variate observation is summarized by two numbers. The number of bins is $2pq$, where $q$ is the number of shells.

5.3.1 DataSpheres algorithm

All the data points within a certain initial time period (chosen by the system administrators to represent a period when the stream is thought to contain no anomalous behavior) are used to define the center of the data and its covariance matrix. These are then used to calculate depth, in terms of the Mahalanobis distance between the observations and the center, and the direction of greatest
variation for those initial points and subsequent points in the data stream.

With the reference center and covariance matrix defined, points in the data stream can be binned according to their depth and direction labels. Below is the DataSpheres method for binning points in a data stream. This method is powerful when compared against other tests by Monte Carlo simulation in Section 5.4.2. In Section 5.5 this method is applied to real data from an AT&T network.

**Step 1: Restriction of data stream to a finite time interval and designation of a reference period for comparison and standardization**

Consider a collection of points in a data stream, \( D \), ordered in time:

\[
D = \{ x_1, x_2, \ldots, x_{t_1}, x_{t_1+1}, \ldots, x_{t_2}, \ldots, x_{t_k} \}.
\]

We segment the data stream into \( k \) sections of data \( D_1, \ldots, D_k \), such that \( D_i = \{ x_{t(i-1)+1}, \ldots, x_{t_i} \} \) for \( 1 \leq i \leq k \), and \( t_0 \equiv 0 \). The endpoints of the data sections \( t_1, \ldots, t_k \) often occur at natural breaks in the data stream such as the end of the hour, day, or week.

We designate \( D^* \) to be the section of the data stream which is the standard for comparing the other \( k - 1 \) sections of the data stream. \( D^* = \{ x_{t^*-1+1}, \ldots, x_{t^*} \} \).

We will test to see if the distribution of points into bins in section \( D_i \) is the same as the distribution for the points in \( D^* \).

**Step 2: Define center of DataSphere**

We define the center of the data, \( \bar{x}^* \), to be the mean of the data points in the reference section of the data stream \( D^* \).

Alternatives to the mean include trimmed means or generalized multivariate medians as measures of center. Small (1990) gives an overview of multivariate
spatial medians. In Section 5.5 we use the trimmed mean as an alternative robust measure of center for our analyses on real data.

**Step 3: Calculate distance from center for each point**

For each point \( x \) in \( D \), we calculate the Mahalanobis distance, \( M_x \), from \( x \) to the reference center \( \bar{x}^* \) using the sample covariance matrix \( S^* \) of \( D^* \).

\[
M_x = \sqrt{(x - \bar{x}^*)'(S^*)^{-1}(x - \bar{x}^*),
\]

\[
S^* = (D^*)'D^*/(n^* - 1).
\]

The value \( n^* \) is the number of points in reference section \( D^* \).

Any mapping of the data points which allows them to be ordered—typically a mapping to the positive real numbers—would work for our purposes. Alternative measures of distance will be discussed in Section 5.3.3.

**Step 4: Assign a depth layer for each point**

Each point \( x \in D \) is assigned a numerical depth layer \( d_x \in \{1, \ldots, q\} \) based on the \( q \) quantiles of the Mahalanobis distances of the points in \( D^* \) from the center \( \bar{x}^* \). The boundaries of the depth layers are determined by first ordering each \( x \in D^* \) from the center outward according to increasing Mahalanobis distances, \( x^*_{(1)}, x^*_{(2)}, \ldots, x^*_{(n^*)} \), where \( M_{x^*_{(1)}} \leq M_{x^*_{(2)}} \leq \ldots \leq M_{x^*_{(n^*)}} \). The first depth layer \( d = 1 \) is comprised of all points with Mahalanobis distance less than \( M_{x^*_{(\frac{n^*}{q})}} \). Thus the boundaries of the depth layers are \( M_{x^*_{(\frac{n^*}{q})}}, M_{x^*_{(\frac{2n^*}{q})}}, \ldots, M_{x^*_{(\frac{(q-1)n^*}{q})}} \). Any \( x \) such that \( M_x > M_{x^*_{(\frac{(q-1)n^*}{q})}} \) is assigned to the last depth layer \( d_x = q \).
Step 5: Determine direction of greatest variation

Assume that all points \( x \in D_i \) have been standardized with respect to the center \( \bar{x}^* \) and covariance matrix \( S^* \) of \( D^* \).

In addition to being assigned a depth layer, each point \( x = (x_1, \ldots, x_p) \) is labeled with its signed component of greatest variation \( v_x \), where \( v_x \in \{\pm 1, \ldots, \pm p\} \).

Let \( j \in \{1, \ldots, p\} \) be the index of the component which is the maximum of \( \{|x_1|, \ldots, |x_p|\} \) so that \( \max\{|x_1|, \ldots, |x_p|\} = |x_j| \). Then

\[
v_x = \text{sign}(x_j) \times j.
\]

For a \( p \)-dimensional data point there are \( 2 \times p \) possible positive or negative directions of greatest variation.

Step 6: Place points into bins according to depth layer and direction

Each point \( x \in D_i \) is classified as having one of \( 2p \) positive or negative directions of greatest variation and one of \( q \) depth layers. Thus each point can be placed into one of \( 2pq \) bins according to its direction \( v_x \) and its depth layer \( d_x \). We summarize the section of the data stream \( D_i \) by a \( 2q \times p \) matrix of counts of points in each of the \( 2pq \) bins. We denote this matrix \( B_i \). The \( p \) columns of \( B_i \) correspond to the \( p \) dimensions of the data. The \( 2q \) rows of \( B_i \) correspond to the depth layers and whether the signed component of greatest variation is positive or negative. The first \( q \) rows of \( B_i \) will contain the counts of points that have a greatest variation in the positive direction. The last \( q \) rows of \( B_i \) will consist of counts of points with greatest variation in the negative direction.

1If a point \( x' \) has not been thusly standardized, do so by subtracting \( \bar{x}^* \) and dividing each component by the corresponding element of the diagonal of \( S^* \) (i.e., \( x_i = (x'_i - \bar{x}^*_i)/\sqrt{S^*_ii} \) for \( i = 1, \ldots, p \)).
The matrix $B_i$ can be written in terms of its columns: $B_i = (b_{i1}, \ldots, b_{ip})$,
where

$$b_{ij} = \begin{pmatrix} \sum_{x \in D_i} I(v_x = +j) I(d_x = q) \\ \sum_{x \in D_i} I(v_x = +j) I(d_x = q - 1) \\ \vdots \\ \sum_{x \in D_i} I(v_x = +j) I(d_x = 1) \\ \sum_{x \in D_i} I(v_x = -j) I(d_x = 1) \\ \vdots \\ \sum_{x \in D_i} I(v_x = -j) I(d_x = q) \end{pmatrix} \text{ for } j \in (1, \ldots, p).$$

An example of such a matrix for real data can be seen in Section 5.5 in Figure 5.7.
We call $B_i$ the DataSphere of the data stream $D_i$. The DataSphere summarizes the $p$-dimensional data stream into $2pq$ counts of points. Likewise the summary for the gold standard data stream $D^*$ is denoted $B^*$.

### 5.3.2 Discussion

This method of DataSpheres for summarizing high-dimensional data sets so that the two data sets can be easily compared satisfies all of our requirements for supporting an algorithm for change detection in data streams, as long as the gold-standard comparison stream, $D^*$, is known a priori along with its center, covariance matrix, and the quantiles of the distance measure. Our method allows for analysis and summarization of an incoming stream with only one pass through the data. It scales to high dimensions, accounts for correlation between attributes, makes no parametric assumptions about the underlying distributions, and is robust to outliers.

Before further discussing the merits of the algorithm in detail, we present Figure 5.1 which graphically depicts our partitioning scheme for a 2-dimensional data
stream. In this example we partition the space into sixteen bins corresponding to four depth layers and four directions of greatest variation (counting positive and negative directions separately). Data point A is classified in depth layer 2 and direction \((-x)\). Point B is also in depth layer 2 but has direction \((-y)\). Point C has direction \((+y)\) and depth layer 4.

The DataSpheres partitioning is simple to compute, requiring only inexpensive calculations for an incoming data stream. In its most basic implementation, described above, a point’s distance from the center is calculated and its depth layer stored, and then the point is standardized and its direction of greatest variation is calculated and stored. Once the depth and direction are stored, the data point
can be ignored and/or erased. On the other hand, the determination of the center, covariance matrix, and quantiles of the distance measure of the comparison stream \( D^* \) requires two passes through the data (Johnson and Dasu, 1998). The mean and covariance of \( D^* \) can be easily updated as more points arrive in the stream, but determining the quantiles of the distance measure requires the complete order statistics of \( D^* \).

The DataSpheres partition of a data stream scales linearly with the number of dimensions. The number of bins in the partition is only \( 2^{pq} \) where \( p \) is the dimension of the data set. An ordinary grid stratification would require \( q^p \) bins, making the DataSpheres method preferable over a grid partition when \( p \) is higher than two.

![Diagram](image.png)

**Figure 5.2:** Left: Euclidean distance forms circular contours of depth and orders the points A, B, then C. Right: Mahalanobis distance forms elliptical contours of depth and orders the points C, B, then A.

When the Mahalanobis distance is used to define the distance of a point to the center (Equation 5.3), and hence its depth, correlations between the dimensions are taken into account. For example, in two dimensions, if \( x \) and \( y \) are posi-
tively correlated, the contours of equal Mahalanobis depth are ellipses. This is in contrast to the circular contours of Euclidean distance. In Figure 5.2, the center-outward depth of points A, B, and C is reversed depending on which distance is used.

Any outliers in the comparison data set $D^*$ would affect the mean and covariance of $D^*$. If the mean were used to define the center, and if the Mahalanobis distance were used as the distance measure, those outliers might affect the efficacy of the DataSpheres method. However, we assume that $D^*$ is generated by a process that is in control, and therefore contains no gross outliers. If an outlier occurred in the test section $D_i$, it would have no more influence over the DataSphere $B_i$ than any other point. To make our method further insensitive to outliers, a trimmed mean or median could be used as the measure of center of $D^*$, as well as a covariance matrix formed with outliers in $D^*$ excluded. These minor variants to the method are briefly discussed in Section 5.3.3.

5.3.3 Variants

Measures of depth and center

Our measure of the depth of data points within a cloud of data is a center-outward, low-to-high ordering of the points by increasing Mahalanobis distance from the center. Points very near the center with low Mahalanobis distance have low depth; points far away from the center have high depth. This notion of depth deviates in a trivial fashion from the conception of Mahalanobis depth proposed by Liu and Singh (1993) who order the data points $\{x_i\}$ high-to-low according to the depth function $[1 + M_{x_i}]^{-1}$, where $M_{x_i}$ is the Mahalanobis distance from $x_i$ to the center. This alternative version of depth satisfies the four desirable properties of
depth functions as defined by Zuo and Serfling (2000). These four properties are:

P1. Affine invariance

P2. Maximality at center

P3. Monotonicity relative to deepest point

P4. Vanishing at infinity.

There are numerous alternative depth functions, many of them more desirable for our purposes that the Mahalanobis depth, but also much more difficult to compute. Two examples are the halfspace depth of Tukey (see Donoho and Gasko, 1992) which defines depth of a point $\mathbf{x} \in \mathbb{R}^p$ by the minimum number of points contained in a closed halfspace through $\mathbf{x}$, and simplicial depth (Liu, 1990) which is the probability that the point $\mathbf{x}$ is contained inside a simplex formed by $p + 1$ random data points. Additional depth functions are reviewed by Zuo and Serfling (2000).

Alternative measures of center include trimmed means, the $L_1$ median (see Vardi and Zhang, 2000), and any other center defined by the point attaining the maximal value of a depth function. The median defined by the Mahalanobis depth is $\bar{\mathbf{x}}$. Likewise, the Tukey halfspace median is the point for which halfspaces through that point contain half of the data. The simplicial median is the point contained in the most simplices with $p + 1$ vertices. We have found that the trimmed mean is a nice compromise between robustness and computational difficulty. The simplicial and halfspace medians are more robust to outliers than the mean, but are very difficult to compute for large data sets (see Struyf and Roussseeuw, 2000).
Distance measures

We use Mahalanobis distance to define the depth of a data point in the Data-Spheres method. As discussed above, alternative depth functions could be used—including any other 1-dimensional distance—to order the data points from the center outward. An alternative distance that could be used is one tuned to the scientific and managerial questions at hand. For example, if the data stream contained categorical variables, a distance could be defined by placing extra weight on deviations within specific attributes. Certain data points could be penalized for having deviations above a threshold if the managers of the data thought that these points were more outlying than other points.

5.3.4 Testing DataSpheres

Our goal is to detect managerially significant changes in data streams in order to assist AT&T manage their data, and we do so for real data in Section 5.5. Our method of DataSpheres produces a matrix of counts of data points, $B_i$, for each section of the data stream $D_i$. These matrices can be considered 2-dimensional histograms and presented to data managers visually in the form of heatmaps. Each cell of the histogram has the number of points of each depth layer and direction of greatest variation. In many cases the differences between data streams are obvious and immediately apparent after inspecting the histograms of $B_i$.

We want a formal hypothesis test for comparing two DataSpheres. We also might wish to conduct hypothesis tests to automatically detect any significant differences between DataSpheres without any human oversight. Two simple, standard tests will be described below for the case in which a data stream is to be compared to the gold-standard data stream $D^*$, and for the case in which data
streams from two separate time intervals, $D_i$ and $D_j$, are to be compared. We also describe two Bayesian tests for comparing bins with multinomial counts of data points.

**The $\chi^2$ goodness-of-fit**

When the gold-standard data stream $D^*$ defines what it means for the data stream to be in control, its DataSpheres summary $B^*$ defines how the DataSpheres for other time intervals should be distributed. To compare data stream $D_i$ with $D^*$ we compare $B_i$ to $B^*$. Specifically, we assume that the counts of data points in each bin of $B_i$ come from a multinomial distribution with probabilities determined by the counts in $B^*$. We use a $\chi^2$ goodness-of-fit test to test the null hypothesis that $B_i$ comes from the same distribution as $B^*$. Assuming that the number of points in $D_i$ is the same as in $D^*$:

$$
\chi^2 = \sum_{k=1}^{p} \sum_{j=1}^{2q} \left( \frac{b_{jk} - b_{jk}^*}{b_{jk}^*} \right)^2.
$$

(5.3)

The degrees of freedom for the $\chi^2$ goodness-of-fit statistic is $q(2p - 1)$ since each of the $q$ depth layers in $B^*$ has an equal number of data points. Given the total number of data points $n$ in $D^*$, each of the $2q$ rows in $B^*$ is constrained to equal $\frac{n}{q}$.

**The $\chi^2$ test of homogeneity**

The second testing scenario compares two sections of the data stream $D_i$ and $D_j$ by comparing their DataSpheres $B_i$ and $B_j$. The depth layers in $B_i$ and $B_j$ are not constrained to have equal numbers of points. Therefore our null hypothesis
is that $\text{vec}(B_i) = b^i$ and $\text{vec}(B_j) = b^j$ are two realizations of the same, unknown multinomial distribution. To test this we use a $\chi^2$ test of homogeneity on the two $2pq$-vectors of bin counts $b^i$ and $b^j$. Assuming the number of points in $D_i$ and $D_j$ is the same:

$$\chi^2 = \sum_{k=1}^{2pq} \left( \frac{(b^i_k - E_k)^2}{E_k} + \frac{(b^j_k - E_k)^2}{E_k} \right),$$

(5.4)

$$E_k = \frac{b^i_k + b^j_k}{2}.$$

The expected count $E_k$ for each of the $2pq$ bins is the average count for that bin if the number of data points in $D_i$ and $D_j$ is equal, i.e. $n_i = n_j$. Otherwise, $E_{ik} = \frac{n_i (b^i_k + b^j_k)}{n_i + n_j}$, and the expected counts $E_{jk}$ for $b^j$ are similarly re-weighted. This test statistic is approximately $\chi^2$ distributed with $(2pq - 1)$ degrees of freedom (Rice, 1995).

**Bayesian multinomial tests**

For a given data stream and reference period, the DataSpheres method produces the counts of data points in each of $2pq$ bins. We assume these counts are multinomial and wish to test whether the counts from different bins come from the same multinomial distribution or a different distribution for each DataSphere. Specifically, we test the hypothesis that two sample data streams, $X$ and $Y$, come from the same distribution by testing whether their DataSpheres vectors of counts, $b^x$ and $b^y$, come from the same multinomial distribution. The alternative is that the counts of points in the bins come from independent multinomial distributions.

The first hypothesis, $H_1$, is that $b^x$ and $b^y$ come from the same multinomial distribution. The second hypothesis, $H_2$, is that $b^x$ and $b^y$ come from independent
multinomial distributions. We consider the posterior odds of each hypothesis given the data, which is the prior odds times the Bayes Factor:

$$\frac{P(H_1 \mid \text{Data})}{P(H_2 \mid \text{Data})} = \frac{P(H_1)}{P(H_2)} \times \frac{P(\text{Data} \mid H_1)}{P(\text{Data} \mid H_2)}. \quad (5.5)$$

Assuming we are indifferent to the hypotheses a priori so that \(\frac{P(H_1)}{P(H_2)} = 1\), the Bayes Factor will indicate which of the two hypotheses is most supported by the data. To calculate the Bayes Factor we compute the joint distribution of the data and the parameters given the model, then integrate over the space of parameters to find the marginal likelihood for the data. We do this for each hypothesis and evaluate the ratio of marginal likelihoods.

For this hypothesis test, the Bayes Factor is given by

$$\text{BF } (H_1; H_2) = \frac{P(\text{Data} \mid H_1)}{P(\text{Data} \mid H_2)} = \frac{\int p(b^x, b^y \mid \theta_1, H_1) p(\theta_1 \mid H_1) d\theta_1}{\int p(b^x, b^y \mid \theta_2, H_2) p(\theta_2 \mid H_2) d\theta_2},$$

where

\[
\begin{align*}
  b^x &= (x_1, x_2, \ldots, x_k), \\
  b^y &= (y_1, y_2, \ldots, y_k), \\
  \theta_1 &= (\theta_1, \theta_2, \ldots, \theta_k), \\
  \theta_2 &= (\theta_x, \theta_y) = (\theta_{x_1}, \ldots, \theta_{x_k}, \theta_{y_1}, \ldots, \theta_{y_k}).
\end{align*}
\]

The likelihood for the hypothesis, \(H_1\), of the common multinomial parameter
\( \theta_1 \) is the product of two multinomials with shared \( \theta_1 \):

\[
p(b^x, b^y \mid \theta_1, H_1) = p(b^x \mid \theta_1, H_1) \cdot p(b^y \mid \theta_1, H_1)
= \left( \sum_{i=1}^{k} x_i \right) \theta_1^{x_1} \cdots \theta_k^{x_k} \left( \sum_{i=1}^{k} y_i \right) \theta_1^{y_1} \cdots \theta_k^{y_k}
= \frac{(\sum x_i)!}{\prod_{i=1}^{k} (x_i)!} \frac{(\sum y_i)!}{\prod_{i=1}^{k} (y_i)!} \prod_{i=1}^{k} \theta_i^{(x_i + y_i)}
= \frac{\Gamma(\sum x_i + 1)}{\prod_{i=1}^{k} \Gamma(x_i + 1)} \frac{\Gamma(\sum y_i + 1)}{\prod_{i=1}^{k} \Gamma(y_i + 1)} \prod_{i=1}^{k} \theta_i^{(x_i + y_i)}
= c_1 \prod_{i=1}^{k} \theta_i^{(x_i + y_i)}.
\]

The likelihood for the hypothesis of independent multinomials, \( H_2 \), is the product of the two different multinomial likelihoods:

\[
p(b^x, b^y \mid \theta_2, H_2) = \left( \sum_{i=1}^{k} x_i \right) \theta_1^{x_1} \cdots \theta_k^{x_k} \left( \sum_{i=1}^{k} y_i \right) \theta_1^{y_1} \cdots \theta_k^{y_k}
= \frac{\Gamma(\sum x_i + 1)}{\prod_{i=1}^{k} \Gamma(x_i + 1)} \frac{\Gamma(\sum y_i + 1)}{\prod_{i=1}^{k} \Gamma(y_i + 1)} \prod_{i=1}^{k} \theta_i^{x_i} \theta_i^{y_i}
= c_1 \prod_{i=1}^{k} \theta_i^{x_i} \theta_i^{y_i}.
\]

We choose a Dirichlet prior, which is conjugate for multinomial data, for \( \theta_1 \):

\[
\theta_1 \sim \text{Dir}(\alpha), \quad \alpha = (\alpha_1, \ldots, \alpha_k), \quad \text{where } \alpha_i > 0 \quad \text{for } i = 1, \ldots, k.
\]

\[
p(\theta_1 \mid H_1) = \frac{\Gamma(\sum \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{(\alpha_i - 1)}.
\]

For \( \theta_2 \) we choose independent Dirichlet priors for \( \theta_x \) and \( \theta_y \) since we hypothesize
that our data come from separate multinomials:

\[
P(\theta_2 \mid H_2) = p(\theta_x, \theta_y \mid H_2) = p(\theta_x \mid H_2)p(\theta_y \mid H_2),
\]

\[
\theta_x \sim \text{Dir}(\alpha_x), \alpha_x = (\alpha_{x1}, \ldots, \alpha_{xk}), \text{ where } \alpha_{xi} > 0 \text{ for } i = 1, \ldots, k,
\]

\[
\theta_y \sim \text{Dir}(\alpha_y), \alpha_y = (\alpha_{y1}, \ldots, \alpha_{yk}), \text{ where } \alpha_{yi} > 0 \text{ for } i = 1, \ldots, k,
\]

\[
p(\theta_x \mid H_2) = \frac{\Gamma(\sum \alpha_{xi})}{\prod_{i=1}^{k} \Gamma(\alpha_{xi})} \prod_{i=1}^{k} \theta_{xi}^{(\alpha_{xi}-1)},
\]

\[
p(\theta_y \mid H_2) = \frac{\Gamma(\sum \alpha_{yi})}{\prod_{i=1}^{k} \Gamma(\alpha_{yi})} \prod_{i=1}^{k} \theta_{yi}^{(\alpha_{yi}-1)},
\]

\[
P(\theta_2 \mid H_2) = \frac{\Gamma(\sum \alpha_{xi})\Gamma(\sum \alpha_{yi})}{\prod_{i=1}^{k} \Gamma(\alpha_{xi})\Gamma(\alpha_{yi})} \prod_{i=1}^{k} \theta_{xi}^{(\alpha_{xi}-1)}\theta_{yi}^{(\alpha_{yi}-1)}.
\]

The marginal probabilities of the data under both hypotheses are straightforward to compute since the Dirichlet priors are conjugate with the multinomial likeli-
hoods:

\[
P(\text{Data} \mid H_1) = \int_{\Theta_1} p(b^x, b^y \mid \theta_1, H_1) p(\theta_1 \mid H_1) \, d\theta_1
\]

\[
= \int_{\Theta_1} c_1 \prod_{i=1}^{k} \theta_i^{(x_i+y_i)} \frac{\Gamma(\sum \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{(\alpha_i-1)} \, d\theta_1
\]

\[
= c_1 \frac{\Gamma(\sum \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \int_{\Theta_1} \prod_{i=1}^{k} \theta_i^{(x_i+y_i+\alpha_i-1)} \, d\theta_1
\]

\[
= c_1 \frac{\Gamma(\sum \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \Gamma(x_i+y_i+\alpha_i) \prod_{i=1}^{k} \theta_i^{\alpha_i} \Gamma(\sum x_i + \sum y_i + \sum \alpha_i).
\]

\[
P(\text{Data} \mid H_2) = \int_{\Theta_2} p(b^x, b^y \mid \theta_2, H_2) p(\theta_2 \mid H_2) \, d\theta_2
\]

\[
= \int_{\Theta_x} \int_{\Theta_y} p(b^x \mid \theta_x, H_2) p(b^y \mid \theta_y, H_2) p(\theta_x \mid H_2) p(\theta_y \mid H_2) \, d\theta_x \, d\theta_y
\]

\[
= \int_{\Theta_x} p(b^x \mid \theta_x, H_2) p(\theta_x \mid H_2) \, d\theta_x \int_{\Theta_y} p(b^y \mid \theta_y, H_2) p(\theta_y \mid H_2) \, d\theta_y
\]

\[
= \int_{\Theta_x} \frac{\Gamma(\sum x_i + 1)}{\prod_{i=1}^{k} \Gamma(x_i+1)} \frac{\Gamma(\sum \alpha_{x_i})}{\prod_{i=1}^{k} \Gamma(\alpha_{x_i})} \prod_{i=1}^{k} \theta_i^{(x_i+\alpha_{x_i}-1)} \, d\theta_x \times
\]

\[
\int_{\Theta_y} \frac{\Gamma(\sum y_i + 1)}{\prod_{i=1}^{k} \Gamma(y_i+1)} \frac{\Gamma(\sum \alpha_{y_i})}{\prod_{i=1}^{k} \Gamma(\alpha_{y_i})} \prod_{i=1}^{k} \theta_i^{(y_i+\alpha_{y_i}-1)} \, d\theta_y
\]

\[
= \frac{\Gamma(\sum x_i + 1)}{\prod_{i=1}^{k} \Gamma(x_i+1)} \frac{\Gamma(\sum \alpha_{x_i})}{\prod_{i=1}^{k} \Gamma(\alpha_{x_i})} \prod_{i=1}^{k} \Gamma(x_i+\alpha_{x_i}) \times
\]

\[
\frac{\Gamma(\sum y_i + 1)}{\prod_{i=1}^{k} \Gamma(y_i+1)} \frac{\Gamma(\sum \alpha_{y_i})}{\prod_{i=1}^{k} \Gamma(\alpha_{y_i})} \prod_{i=1}^{k} \Gamma(y_i+\alpha_{y_i})
\]

\[
= c_1 \frac{\Gamma(\sum \alpha_{x_i}) \Gamma(\sum \alpha_{y_i})}{\prod_{i=1}^{k} \Gamma(\alpha_{x_i}) \Gamma(\alpha_{y_i})} \prod_{i=1}^{k} \Gamma(x_i+\alpha_{x_i}) \Gamma(y_i+\alpha_{y_i}) \prod_{i=1}^{k} \theta_i^{\alpha_i} \Gamma(\sum x_i + \sum \alpha_{x_i}) \Gamma(\sum y_i + \sum \alpha_{y_i}).
\]

127
Therefore the Bayes Factor \((H_1; H_2)\) is the ratio of these two marginal data likelihoods:

\[
BF(H_1; H_2) = \frac{c_1 \Gamma(\sum \alpha x_i) \Gamma(\sum \alpha y_i) \prod_{i=1}^{k} \Gamma(x_i + \alpha x_i) \Gamma(y_i + \alpha y_i)}{\prod_{i=1}^{k} \Gamma(\alpha i) \Gamma(\sum x_i + \sum y_i + \sum \alpha i)}.
\] (5.6)

This Bayes Factor is our measure for which hypothesis is most supported by the data. If the Bayes Factor is less than one we reject \(H_1\) and accept the hypothesis indicating a difference between the DataSpheres. When the Bayes Factor is greater than one we accept \(H_2\), the hypothesis of no significant change between the data streams.

However, the Bayes Factor can be sensitive to the choice of hyper-parameters \(\alpha\), \(\alpha_x\), and \(\alpha_y\) (see Kass and Raftery, 1995; Good, 1965). In the power simulations below in Section 5.4.2, we use the median Intrinsic Bayes Factor (Berger and Pericchi, 1996) for our hypothesis test. In short, we create a more informative prior by using a random sample of the data. We first calculate the posterior distribution of the multinomial parameters from a small sample from the data streams \(X\) and \(Y\) and non-informative priors on \(\theta_1\) and \(\theta_2\). We use this posterior distribution as the new prior for the rest of the data (not including the sample) to calculate the marginal likelihoods of the data under the two hypotheses. We repeat this process multiple times using different samples of the data each time in order to produce a sequence of Bayes Factors. We choose the median of this sequence as the Intrinsic Bayes Factor and base our Bayesian hypothesis test on its value.
5.4 Evaluation of the DataSpheres Method

The uniformly most powerful test for detecting a shift in means between two samples is Hotelling’s two-sample $T^2$ test when the data are multivariate normal (Mardia et al., 1979). Just as the Student’s $t$ statistic can be used to compare two samples of univariate, normally-distributed data, Hotelling’s $T^2$ statistic can be used to test if the difference in means between two multivariate normal samples is more or less than expected. We compare our method of DataSpheres to this test for simulated multivariate normal data to see how much power our test loses. We also compare DataSpheres with the Hotelling’s $T^2$ test for heavy-tailed data.

5.4.1 Power and the asymptotic relative efficiency of DataSpheres

Hotelling’s $T^2$ test

If $X = (x_1, \ldots, x_{n_x})'$ with $x_i \sim N_p(\mu, \Sigma)$ and $Y = (y_1, \ldots, y_{n_y})'$ with $y_i \sim N_p(\mu, \Sigma)$ are two matrices ($n_x \times p$ and $n_y \times p$, respectively) of independent samples from a $p$-variate normal distribution, and $S = \frac{X'X + Y'Y}{n_x + n_y - 2}$ is the pooled sample covariance matrix, then:

$$T^2 = \frac{n_x n_y}{n_x + n_y} (\bar{x} - \bar{y})' S^{-1} (\bar{x} - \bar{y}).$$  \hspace{1cm} (5.7)

This Hotelling $T^2$ statistic is proportional to the Mahalanobis distance between the two means $\bar{x}$ and $\bar{y}$ and has an $F$ distribution (Mardia et al., 1979) under the null hypothesis of equal means:

$$\frac{n_x + n_y - p - 1}{(n_x + n_y - 2)p} T^2 \sim F(p, n_x + n_y - p - 1).$$  \hspace{1cm} (5.8)

129
If an alternative hypothesis is that the means are different, i.e., \( \mathbf{x}_i \sim N_p(0, \Sigma) \) and \( \mathbf{y}_i \sim N_p(\mu_1, \Sigma) \), then the difference in sample means will be proportional to a non-central \( \chi^2 \) distribution with \( p \) degrees of freedom and a non-centrality parameter of \( \frac{1}{2} \mu_1' \Sigma^{-1} \mu_1 \):

\[
(\bar{\mathbf{x}} - \bar{\mathbf{y}})'\Sigma^{-1}(\bar{\mathbf{x}} - \bar{\mathbf{y}}) \sim \frac{n_x + n_y}{n_x n_y} \chi^2 \left( p, \frac{1}{2} \mu_1' \Sigma^{-1} \mu_1 \right).
\] (5.9)

In our example \( (\bar{\mathbf{x}} - \bar{\mathbf{y}}) \sim N_p(\mu_1, \frac{n_x + n_y}{n_x n_y} \Sigma) \), and if \( \mathbf{a} \sim N_p(\mathbf{\xi}, \Omega) \), then \( \mathbf{a}'\Omega^{-1}\mathbf{a} \sim \chi^2 \left( p, \frac{1}{2} \mathbf{\xi}'\Omega^{-1}\mathbf{\xi} \right) \), where \( \frac{1}{2} \mathbf{\xi}'\Omega^{-1}\mathbf{\xi} \) is the non-centrality parameter of the \( \chi^2 \) distribution with \( p \) degrees of freedom.

Hence, for a given significance level, the power of Hotelling’s \( T^2 \) test is the integral of the non-central \( \chi^2 \left( p, \frac{1}{2} \mu_1' \Sigma^{-1} \mu_1 \right) \) distribution from the minimal \( F \) rejection value to infinity. This power is a function only of the sample sizes \( n_x \) and \( n_y \), the dimension \( p \), and the Mahalanobis distance of \( \mu_1 \) from \( 0 \).

**Power of DataSpheres \( \chi^2 \) test**

For testing the similarity of two DataSpheres \( \mathbf{b}^x \) and \( \mathbf{b}^y \) from samples of data \( \mathbf{X} \) and \( \mathbf{Y} \), the approximate distribution of our test statistic (Equation 5.5) under the null hypothesis that the multinomial parameters \( \mathbf{\theta}_x \) and \( \mathbf{\theta}_y \) are equal is \( \chi^2 \) with \( 2pq - 1 \) degrees of freedom. If the alternative is true and \( \mathbf{\theta}_x \neq \mathbf{\theta}_y \) with \( \mathbf{\theta}_x \) and \( \mathbf{\theta}_y \) known, the distribution of our test statistic becomes approximately non-central \( \chi^2 \) with \( 2pq - 1 \) degrees of freedom and a non-centrality parameter that is a function of the difference between \( \mathbf{\theta}_x \) and \( \mathbf{\theta}_y \).

The expected multinomial probabilities \( \mathbf{\theta}_x \) and \( \mathbf{\theta}_y \) for the counts in the DataSpheres bins can be determined through integration if the distribution of the data
samples $X$ and $Y$ are known. These distributions are rarely known in practice, but for argument’s sake, let us assume that they are multivariate normal and that the data have been centered and standardized according to Step 5 of the DataSpheres method, using the data stream $X$ as the reference period. Then:

$$X \sim N_p(0, \Sigma), \text{ and}$$

$$Y \sim N_p(\mu_1, \Omega).$$

The variance matrix $\Sigma$ has ones along the diagonal, but $\Omega$ need not.

The probability is $\frac{1}{2pq}$ that a point $x_i$ from the data stream $X$ is placed in bin $b_{(d,v)}^x$, where $d \in \{1, \ldots, q\}$ is the depth layer and $v \in \{-1, \ldots, 1\}$ indicates the direction of greatest variation for point $x_i$. Each bin is equally probable, and thus $\theta_x = \frac{1}{2pq} 1$. This can be demonstrated by considering:

1. The depth layers are determined by the $q$ quantiles of Mahalanobis distances of the data stream $X$ and thus have equal counts.

2. The direction of greatest variation is solely determined by the marginal distributions of $x_{i1}, \ldots, x_{ip}$ which are the standard normal distributions. This implies equal expected counts for each positive or negative direction.

3. The depth and direction of greatest variation are independent since the multivariate normal distribution is spherically symmetric about the origin.

Determining the bin probabilities $\theta_y$ for $b^y$ is made easier by the fact that the boundaries of the bins are fixed by $b^x$. The boundaries of the depth layers are determined by the $q$ quantiles of the Mahalanobis distances of $X$, and can be represented as $M_{x_1(\frac{1}{2p})}, \ldots, M_{x_q(\frac{1}{2p})}$, using the notation of Step 4 in Section 5.3.1. The
last, \( q \)-th depth layer boundary \( M_{x,\frac{q-1}{q}} \equiv \infty \) since any point \( y \) with Mahalanobis distance greater than \( M_{x,\frac{q-1}{q}} \) is assigned to the \( q \)-th depth layer. The actual values of these depth layer boundaries are of course determined by the data, but the expected values of these boundaries are the quantiles of the \( \chi^2_p \) distribution since \( x'\Sigma^{-1}x \) has a \( \chi^2_p \) distribution with \( p \) degrees of freedom.

In general, the depth and direction of greatest variation for point \( y \) are not independent. Therefore the probability that a point \( y \) is in bin \( b_{(d,v)}^y \) is the integral of the multivariate normal density \( \phi(Y; \mu_1, \Omega) \) over the region of each fixed bin. For example, consider the bivariate case with \( p = 2 \), direction \( v = +x \), and depth equal to \( d \) where \( d \in \{1, \ldots, q\} \) for a data stream \( Y \sim N_2(\mu_1, \Omega) \). Then

\[
P(y \in b_{(d,+x)}^y) = \theta_{(d,+x)}^y \]

where:

\[
\theta_{(d,+x)}^y = \int_{-x}^{x} \int_{M_{x,\frac{(d-1)n}{q}}}^{M_{x,\frac{dn}{q}}} \phi(\mu_1, \Omega) \, dy \, dx .
\] (5.10)

Note that if \( d = 1 \), the lower limit of integration along the \( x \) direction is defined to be 0, i.e., \( M_{x,\frac{q}{q}} \equiv 0 \). This integral (5.10) can be generalized to higher dimensions by integrating \( \phi(\mu_1, \Omega) \, dz \) over the values \((-x, +x)\) for each additional dimension \( z \), just as \( \phi(\mu_1, \Omega) \, dy \) was integrated from \((-x, +x)\) for the \( y \) dimension in our example:

\[
\theta_{(d,+x)}^y = \int_{-x}^{x} \cdots \int_{-x}^{x} \phi(\mu_1, \Omega) \, dz \, \cdots \, dy \, dx .
\] (5.11)

**Asymptotic relative efficiency of bivariate sign test**

Bennett (1962) developed a non-parametric sign test for paired bivariate normal data \( X \) and \( Y \) where \( X \sim N_2(0, \Sigma) \) and \( Y \sim N_2(\mu_1, \Sigma) \). In subsequent work he
showed that the asymptotic relative efficiency of the bivariate sign test compared to the Hotelling \( T^2 \) test was between 0.637 and 0.726 depending on \( \rho \), the correlation coefficient of \( \Sigma \), i.e., \( \text{Corr}(y_{i1}, y_{i2}) = \rho \). For \( \rho = 0 \), the ARE is the minimum 0.637, and when \( \rho \) is approximately 0.7, the ARE attains its maximum (Bennett, 1968).

With several assumptions, the DataSpheres \( \chi^2 \) test of homogeneity can be made equivalent to the bivariate sign test through rotation of the data, and thus shares its ARE with the Hotelling \( T^2 \) test when the data are bivariate normal.

### 5.4.2 Multivariate normal simulation

We achieve a more thorough investigation of the power achieved by the DataSpheres method compared to Hotelling’s \( T^2 \) test through Monte Carlo simulations under a variety of conditions. Our simulated data are samples from two multivariate normal distributions with different means but the same, known covariance matrix \( \Sigma \). We assume the first sample \( X \) has mean \( 0 \), the second sample \( Y \) has mean \( \mu_1 \), and the pooled sample covariance matrix is \( S \):

\[
X = (x_1, x_2, \ldots, x_{n_x})', \quad \text{where } x_i \sim N_p(0, \Sigma) \quad \text{for } i = 1 \ldots n_x,
\]

\[
Y = (y_1, y_2, \ldots, y_{n_y})', \quad \text{where } y_j \sim N_p(\mu_1, \Sigma) \quad \text{for } j = 1 \ldots n_y,
\]

\[
S = \frac{XX' + YY'}{n_x + n_y - 2}.
\]

The two-sample Hotelling \( T^2 \) test statistic (Equation 5.7) is calculated from the two samples \( X \) and \( Y \). We reject the null hypothesis that the two samples have the same mean if the \( T^2 \) test statistic is too large and falls within the rejection region determined by Equation 5.8 at the significance level of \( \alpha = 0.01 \).
For our DataSpheres method we treat sample $X$ as the gold-standard and partition $X$ and $Y$ into bins according to the steps in Section 5.3.1. The resulting counts of points, the DataSpheres $B_X$ and $B_Y$, are then subjected to the $\chi^2$ test of homogeneity (Section 5.3.4) to determine whether they come from the same multinomial distribution. If the $\chi^2$ test statistic with degrees of freedom $(2pq - 1)$ is large and falls within the rejection region as determined by $\alpha = 0.01$, we reject the null hypothesis that $X$ and $Y$ come from the same distribution. Additionally, we perform a Bayesian hypothesis test using the Intrinsic Bayes Factor to decide if the multinomial counts in the bins come from the same or different underlying distributions.

The major factor determining the power of the tests was how large the true difference in means between samples $X$ and $Y$ was, or, rather, how large $\mu_1$ was. In this simulation study we tested several cases of different weightings of $\mu_1$. For example, we tested $\mu_1 = (\|\mu_1\|, 0, \ldots, 0)$ and $\mu_1 = (\|\mu_1\|/\sqrt{p}, \|\mu_1\|/\sqrt{p}, \ldots, \|\mu_1\|/\sqrt{p})$, as well as a scenario when $\mu_1$ had its weight distributed over half the directions and also a random distribution over all the directions. In each case we varied $\mu_1$ so that its norm, $\|\mu_1\|$, ranged between 0 and 1.

The results were insensitive to how $\mu_1$ was weighted, so we set it to be evenly distributed through the dimensions which we fixed at $p = 6$, i.e., $\mu_1 = (\|\mu_1\|/\sqrt{6}, \|\mu_1\|/\sqrt{6}, \ldots, \|\mu_1\|/\sqrt{6})$. We fixed the number of depth layers to be $q = 5$, and we drew 2000 samples for each data set, i.e., $n_x = n_y = 2000$.

We also varied the covariance matrix $\Sigma$ in our simulations to reflect “negative” correlations between the six dimensions, no correlations, and positive correlations. For example, the first correlation in our simulations was $\rho = -0.5$ which meant that the first column $x^{(1)}$ of $X$ had a correlation of $-0.5$ with the second column.
In general, the covariance matrices for our simulations were generated such that the correlation between column vectors \( \mathbf{x}^{(i)} \) and \( \mathbf{x}^{(j)} \) was \( \rho^{|i-j|} \). That is: 
\[ \Sigma_{ij} = [\rho^{|i-j|}] \text{ for } i, j = 1, \ldots, 6, \text{ and } \rho = -0.5, 0, \text{ and } 0.5. \]

**Figure 5.3:** For various shifts in the mean \( \|\mu_1\| \), the power of the Hotelling \( T^2 \) test and the DataSpheres method is plotted. The Hotelling test is more powerful for each of the three covariance matrices.

Figure 5.3 shows the results of our simulations for various values of \( \|\mu_1\| \) from 0 to 1, and \( \rho = -0.5, 0, \text{ and } 0.5. \) The power of the Hotelling \( T^2 \) test and the
DataSpheres method using both the $\chi^2$ and Bayesian test with the median Intrinsic Bayes Factor increased as the difference in the means of $X$ and $Y$ increased. The Hotelling $T^2$ test was more powerful than the DataSpheres method for each value of $\|\mu_1\|$ and $\rho$. All three tests appeared most powerful for $\rho = -0.5$ and least powerful for $\rho = 0.5$. However, this difference in power within the tests was due to the fact that, for normal data, the Mahalanobis distance of the difference in means ($\mu'_1 \Sigma^{-1} \mu_1$) is the quantity of interest. That value—and not $\|\mu_1\|$—determines the $T^2$ statistic and the probabilities that data will fall into certain depth layers.

The power of the Bayesian multinomial hypothesis test was lower than both Hotelling’s $T^2$ test and the $\chi^2$ test of homogeneity. Due to the construction of our hypotheses, we computed a Bayes Factor for $H_1$ versus $H_2$ (see Section 5.3.4) which compared the first hypothesis of a shared multinomial distribution with $2pq = 60$ parameters against the second hypothesis of independent multinomials which had $2 \times 2pq = 120$ parameters. Before accepting the more complicated model, the Bayesian test required a very large discrepancy between data sets since it had a built-in “Occam’s Razor” effect that favored simpler hypotheses. The jump from 60 parameters in our simulation to 120 parameters placed a high threshold for the data to overcome before the Bayesian test favored the more complicated model.

We also ran Monte Carlo simulations of the power of the three tests for different values of the other quantities in our study. For example, when the sample size was increased or decreased from $n_x = n_y = 2000$ the power of all three tests correspondingly increased or decreased as expected. Increasing the number of depth layers from $d = 5$ decreased the power for the $\chi^2$ test. Increasing or decreasing the dimension from $p = 6$ similarly decreased or increased the power of all three
tests. Changing the significance level from $\alpha = 0.01$ affected the power of the $T^2$ and the $\chi^2$ tests in the obvious way.

Even in the case of normal data when the Hotelling $T^2$ test is uniformly most powerful, the DataSpheres method performed rather well. When detecting a shift in means it did not seem to lose much power—a very encouraging finding. We suspect that the method of DataSpheres would perform even better if shifts in the covariance structure were also tested.

### 5.4.3 Heavy-tailed simulation

For this simulation study we used the same methods as above on data drawn from $t$ distributions with 1, 2, and 4 degrees of freedom in order to compare the power of Hotelling’s $T^2$ test to the DataSpheres method for heavy-tailed data. An attractive feature of the DataSpheres method is that it is robust against outliers when a trimmed mean and covariance matrix from the reference data stream are used for determining depth and direction (see Section 5.3.3). For this simulation study we used a 5% trimmed mean and covariance matrix just as we did in our analysis of the real AT&T data below (see Section 5.5.3 for more details on the trimming procedure). We fixed the covariance matrices to correspond to a correlation of $\rho = 0.5$.

Figure 5.4 shows the results of our study for various shifts in the norm of $\mu_1$ between 0 and 1, and for varying degrees of heaviness in the tails of the simulated data. When the degrees of freedom was 1 for the $t_1$ distributed data, the Hotelling $T^2$ test was utterly ineffective. That test was powerless to detect differences between the sample means of Cauchy data. As the heaviness in the tails of the data decreased, the Hotelling $T^2$ test performed better and was more powerful
Figure 5.4: For various values of $\|\mu_1\|$ along the x-axis, the power of the Hotelling test and DataSpheres method is shown on the y-axis. The DataSpheres method is more powerful than the Hotelling $T^2$ test when the data have very heavy tails. The DataSpheres method outperforms the Hotelling test and DataSpheres method for $t$-distributed data with degrees of freedom 4 and higher.

The heaviness in the tails affected the DataSpheres $\chi^2$ test very little. It performed about as well as when the data were normal. The Bayesian hypothesis test using the median Intrinsic Bayes Factor was again less powerful than the $\chi^2$ test.
test, but performed better than the Hotelling $T^2$ test for Cauchy data and about as well for $t_2$ distributed data.

Variations in the sample size, the number of dimensions, the covariance matrices, the significance level, and the weighting of the mean vector either changed the power very little or affected the power in a manner consistent with our expectations.

## 5.5 Application of DataSpheres to Network Server Data

In this section we demonstrate how our DataSpheres method can be used in practice by managers of the data to detect significant changes within a large data stream. Our data come from the communications company AT&T. The data derive from one server in a specialized network of servers which handles web-hosting and e-commerce for a popular sports website. For five weeks in 2006 this server was polled for data every five minutes. At each five-minute interval 27 variables were recorded, including the number of in-packets, out-packets, CPU usage, number of procedures, ping latency, and the number of threads. Table 5.1 shows all of the variables and their explanations or unit of measurement if available.

Some of the 27 variables showed little variation from interval to interval, whereas other variables varied in a correlated manner from one five-minute interval to the next. Our goal was to detect anomalous traffic patterns and to find any patterns in the data which could predict future anomalous behavior.

Managers at AT&T intend to monitor this data stream continually. Because of the high traffic volume, it is problematic to do analyses that require extensive
Table 5.1: The 27 variables in the Server data set and their definitions.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Name</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>Application Log</td>
<td>number of events</td>
</tr>
<tr>
<td>$x_2$</td>
<td>CPU Idle %</td>
<td></td>
</tr>
<tr>
<td>$x_3$</td>
<td>CPU System %</td>
<td>CPU used by system</td>
</tr>
<tr>
<td>$x_4$</td>
<td>CPU Used %</td>
<td></td>
</tr>
<tr>
<td>$x_5$</td>
<td>CPU User %</td>
<td>CPU used by users</td>
</tr>
<tr>
<td>$x_6$</td>
<td>Free Memory</td>
<td>MB</td>
</tr>
<tr>
<td>$x_7$</td>
<td>Free Swap</td>
<td>GB</td>
</tr>
<tr>
<td>$x_8$</td>
<td>Free SysDisk Space</td>
<td>GB</td>
</tr>
<tr>
<td>$x_9$</td>
<td>In Errors</td>
<td>number</td>
</tr>
<tr>
<td>$x_{10}$</td>
<td>In Packets</td>
<td>number</td>
</tr>
<tr>
<td>$x_{11}$</td>
<td>Load Average</td>
<td></td>
</tr>
<tr>
<td>$x_{12}$</td>
<td>Number of Procs</td>
<td>number</td>
</tr>
<tr>
<td>$x_{13}$</td>
<td>Number of Threads</td>
<td>number</td>
</tr>
<tr>
<td>$x_{14}$</td>
<td>Out Errors</td>
<td>number</td>
</tr>
<tr>
<td>$x_{15}$</td>
<td>Out Packets</td>
<td>number</td>
</tr>
<tr>
<td>$x_{16}$</td>
<td>Page In</td>
<td>number</td>
</tr>
<tr>
<td>$x_{17}$</td>
<td>Page Out</td>
<td>number</td>
</tr>
<tr>
<td>$x_{18}$</td>
<td>Ping Packet Loss %</td>
<td></td>
</tr>
<tr>
<td>$x_{19}$</td>
<td>Ping latency</td>
<td>millisecs</td>
</tr>
<tr>
<td>$x_{20}$</td>
<td>Security Log</td>
<td>number of events</td>
</tr>
<tr>
<td>$x_{21}$</td>
<td>System Log</td>
<td>number of events</td>
</tr>
<tr>
<td>$x_{22}$</td>
<td>Total Memory</td>
<td>MB</td>
</tr>
<tr>
<td>$x_{23}$</td>
<td>Total Swap</td>
<td>GB</td>
</tr>
<tr>
<td>$x_{24}$</td>
<td>Total SysDisk Space</td>
<td>GB</td>
</tr>
<tr>
<td>$x_{25}$</td>
<td>Used Memory %</td>
<td></td>
</tr>
<tr>
<td>$x_{26}$</td>
<td>Used Swap %</td>
<td></td>
</tr>
<tr>
<td>$x_{27}$</td>
<td>Used SysDisk Space %</td>
<td></td>
</tr>
</tbody>
</table>

A good solution requires a compromise between traditional multivariate methods—such as Hotelling’s $T^2$ test which requires computationally expensive calculations of covariance matrices and their inverses—and alternative methods which summarize the data, and by doing so, sacrifice potentially useful detail.
We propose that our method of DataSpheres can provide these managers with a computationally inexpensive method that makes no assumptions on the distributions of the data streams and can detect managerially-significant changes in the data from one time period to the next. We use this server data as a test case and demonstration of the efficacy of our method.

5.5.1 Variable selection and elimination

Identification of non-informative variables

In order to detect changes in the 27-dimensional data stream we first eliminated the non-informative variables from further analysis. Several variables did not change at all over the five-week period, or changed very little. Other variables changed in a predictable and meaningless way. Still other variables were highly correlated—or perfectly correlated—with each other. We combined or dropped some of these correlated variables.

The variables In Errors, Out Errors, Total Memory, Total Swap, and Total SysDisk Space were constant throughout the five-week period of study. The variable Load Average was almost always zero and therefore did not be used in our analysis. The variables Application Log, System Log, and Security Log changed only very rarely. These three Log variables might be worthy of further study because of their infrequency of change, but were excluded from our study.

The two variables Free SysDisk Space and Used SysDisk Space were complements of each other and varied in a predictable and non-informative manner. The Used SysDisk Space increased linearly with time until it reached a certain threshold when it abruptly reset to a low value. Also complements of each other, of the two variables Free Swap and Used Swap, only Used Swap was selected for
our analysis.

The variables Free Memory and Used Memory were highly correlated with Number of Procs, so these two memory variables were dropped in favor of the Number of Procs variable. Also highly correlated, the In Packets and Out Packets were added together and a new variable, Log Packets, was created by taking the log of this sum. This new variable was a measure of the traffic coming into and out of the network server.

Of the four related CPU variables, CPU Idle %, CPU System %, CPU Used %, and CPU User %, we used only one of them, the CPU Used % variable. Of the two related Page In and Page Out variables we used neither in our analysis.

Six chosen variables

After eliminating or combining the 27 variables—based on characteristics of the data and expert input from AT&T scientists—we were left with six variables which were judged to be important and which showed an interesting amount of variation throughout the study period. The six variables we selected—CPU Used %, Number of Procs, Number of Threads, Ping Latency, Used Swap, and Log Packets—provided an interesting test case for our method of DataSpheres. These variables exhibited some positive and negative pairwise correlations, though no correlation was extreme. Together we believe these were important variables for process measurement in ways that were descriptive of the data set as a whole.

5.5.2 Univariate EDA

Before implementing the DataSpheres algorithm on our test data streams, and as a guide for the variable selection above, we did some exploratory data analysis
of our six chosen variables. Figures 5.5 and 5.6 show boxplots of the data for each day of our five-week study period. Occasionally some of the data in the five-minute intervals were missing, though most days had the complete 288 data points. Relatively few of the data points differed greatly from the day’s mode. Many of the daily boxplots were lines at the median/mode value with a few appreciable outliers.

Most noticeable in the daily boxplots were the many jumps and shifts in Free Memory and Number of Threads (see Figure 5.6). Pairwise scatter plots showed no obvious patterns or correlations between all of the six variables.

For the six variables we also constructed boxplots for each hour of the day. No periodicities could be seen in these hourly boxplots. However, a plot of the standard units of the six variables for the entire 5 week study period showed occasional, regular spikes for the CPU Used % variable. Approximately every 25 hours for two weeks during the study period this variable had one 5-minute interval that showed an unusual increase.

5.5.3 Application of DataSpheres method

We segmented the data stream, $D$, into five one-week sections of data $D_1, \ldots, D_5$. The first week of data, $D_1$, was designated as the section of the data stream to serve as the standard for comparing the other four sections of the data stream.

A more robust mean and covariance matrix were computed from $D_1$ using trimmed data. We first calculated the mean and covariance matrix using the complete data and computed the Mahalanobis distance of each data point from this mean. The trimmed mean and covariance matrix was recomputed using the 95% most central data points. The 5% of the data farthest from the complete
Figure 5.5: Daily boxplots of three variables CPU Used %, Number of Procs, and Number of Threads.

data mean were excluded from these new calculations.

We denote this trimmed mean center of the data stream $\bar{x}^*$, and the trimmed covariance matrix $S^*$. For each point $x$ in $D$ we calculated the Mahalanobis
distance from $\bar{x}^*$ using Equation 5.3. Based on these distances, each point $x$ in $D$ was assigned a center-outwards depth layer from 1 to 5 with the outermost points classified in depth layer 5.
For each point we determined the direction of greatest variation in units standardized by \( \bar{x}^* \) and \( S^* \). For example, a standardized point \( x' = (0.3, 1.26, -2.42, -0.03, .6, .15) \) was labeled with the \((-\text{Threads})\) direction since the third variable, \textit{Number of Threads}, was most extreme of the six variables for point \( x' \) and was negative.

After all the points were labeled by their depth and direction of greatest variation, the points were tallied and the DataSphere bins \( B_1, \ldots, B_5 \) were created. In this example we analyzed the data stream all at once, but an attractive feature of our method is that points can be processed as soon as they arrive as long as \( \bar{x}^* \) and \( S^* \) have already been determined. It is computationally inexpensive to determine the depth and direction of a point, and the DataSphere bins can be easily updated as each new point arrives from the stream.

Figure 5.7 shows the DataSphere \( B_1 \) (represented as a heatmap) for Week 1. Darker colors represent higher counts of data points within each bin. Each depth layer contained an equal number of points. However, not every direction was equally represented, and depth and direction were not independent. Therefore the counts in each bin were not uniform as would have the case had the data been normally distributed. The variable \textit{Number of Threads} had the most heterogeneous distribution of the six dimensions for Week 1 and had the most number of points classified in its \((-\text{Threads})\) and \((+\text{Threads})\) directions.
Figure 5.7: The data points from Week 1 were binned according to the Data-Spheres method. Darker colors represent higher counts in each bin.

Figure 5.8 shows the DataSpheres of Weeks 2 through 5. Because Week 1 was the standard, each depth layer in the bins $B_2, \ldots, B_5$ did not necessarily contain 20% of the data points. In fact, the majority of the data points lay in the depth layer farthest from the center $\bar{x}^*$. This was especially true for Weeks 4 and 5 which had nearly all their points in the outermost depth layer 5. Weeks 2 and 3 had a few points in depth layers 1, 2, and 3, but most of their points were classified to be in the outer depth layers 4 and 5. As the data stream moved further from
the reference period, the points drifted towards a new center and thus occupied higher and higher depth layers when compared to the reference data stream.

Managers of the data can detect shifts in the centers of the data streams visually from these DataSpheres heatmaps. But they can also easily detect possible causes of these shifts by analyzing the directions of greatest variation. In

**Figure 5.8:** DataSpheres for Week 2, Week 3, Week 4, and Week 5. Darker colors represent higher counts in each bin.
Week 4, the variable Log Packets had the highest counts of points classified in the (−Packets) and (+Packets) directions of greatest variation, indicating that many of the points from Week 4 had either highly positive Log Packets or highly negative Log Packets. The variable CPU Used % also was heavily populated in both the positive and negative directions for Week 4.

In contrast, the DataSphere for Week 5 shows that the majority of points were classified into a single bin, the (+Packets) bin of depth layer 5. The points shifted from being highly variable in both Packets directions and several other directions, including (−CPU) and (+CPU), to being predominantly most extreme in only the (+Packets) direction. The data manager can detect from these heatmaps that the most significant feature of the data stream in Week 5—and its changes compared to previous weeks—was the increase in Log Packets.

5.6 Conclusion

The analysis of high-dimensional, rapidly accumulating data streams requires computationally inexpensive methods that scale well to high dimensions, make little or no distributional assumptions about the data, are robust against outliers, and can process data in real-time as they arrive. We believe that our method of Data-Spheres satisfies all of these requirements. Furthermore, we have demonstrated through Monte Carlo simulations that our method retains much of the power of the Hotelling $T^2$ test for normally distributed data, and can be more powerful when applied to heavy-tailed data. Our application to the data stream from an AT&T network server demonstrated how our method could be implemented by managers of the data to detect and diagnose important changes in the data stream.
Chapter 6

Future Research

6.1 Extensions and Modifications

We developed statistical methodology for the treatment of dynamic, network data and applied this methodology to observations of wild African elephants, telephone calls in a calling network, and a data stream from an AT&T network server. This work has suggested possible extensions and modifications of this methodology which could be applied to other areas.

6.1.1 Elephant families

In Chapter 2 we were able to answer the scientific questions of how the elephant social structure changes from the wet season to the dry season and what affect genetics and kinship have on this social structure. However, because we focused only on the AA Family our conclusions are limited in scope, though it seems reasonable to surmise that our findings have broader relevance since the family is not atypical of those found in Amboseli National Park. Fortunately, our methods can be applied without modification to any family of elephants for which their
kinships and genetics are known. One potential future research direction is to apply our model to more families to see if our findings hold for a wider sample of elephants.

Ideally, to answer broader questions about elephant social structure, we would model all elephant families together in the same model. Using the notation of Equation 2.1, and further indexing the pair of elephants \((i, j)\) by their family \(f\), one variation on the model that keeps some parameters family-specific and allows others to be shared through the population is:

\[
\text{logit}(p_{ij}^f) = \beta_0 + \xi_f + \alpha_i^f + \alpha_j^f + \beta_k^f k_{ij}^f + \gamma_{ij}^f + (z_i^f)' z_j^f .
\]  

(6.1)

In this specification, each family has its own random effect \(\xi_f\) governing the density of ties within the family. The sociabilities \(\alpha\) remain family-specific, as would the unstructured pairwise error term \(\gamma_{ij}^f\), and the social space effect \((z_i^f)' z_j^f\). Because interactions between specific elephants of different families are rarely recorded, each family requires its own social space. The idea behind sharing between families the baseline probability parameter \(\beta_0\) and the kinship parameters \(\beta_k\) is that inference on the seasonal effect and the kinship effects would then apply to the general population. After accounting for familial, individual, and pairwise idiosyncrasies with other terms in the model, the kinship parameters \(\beta_k\) would show the quantitative effect of kinships and genetics on the probability of two elephants in the same family socializing together.

Another modification to the model suggested by our work is to include the wet season and dry season data in the same model. Which parameters to keep season-specific would be dictated by the scientific questions posed. It would seem natural to hold the social space effect separate in order to visually depict seasonal changes, but to make the sociability random effects the same for both seasons.
Unclear to the author is whether to regard the kinship effects as the same or different for each season.

### 6.1.2 Calling networks

If one were interested in how covariate information—either on the individual telephone numbers or on the dyads in the network—related to the presence/absence and direction of links between telephone numbers, one could apply a similar model to the calling network data of Chapter 4 as was applied in Chapter 2 to the elephants. The daily calls could be aggregated for the entire month to form binomial relationship data between telephone numbers, and the model of Chapter 2 could be immediately applied, provided interesting covariate information were available. For example, the model could be used to show how the induction of a new telephone number into a promotional “circle” of friends affected calling behavior.

### 6.1.3 DataSpheres

Our method of DataSpheres in Chapter 5 could be applied to many other streaming data sets of interest to AT&T and other companies or scientific researchers. One data set to which we have already applied this method was the collection of the 32 routers of the internet backbone owned and managed by AT&T. The DataSpheres method worked well in detecting changes from week to week on these routers and in identifying abnormal behavior on a subset of them. Because they literally form the structure of the internet in the United States, detecting changes and abnormalities on these routers before they become major problems and cause disruptions of internet traffic is a worthwhile endeavor.

Suggested by the method for treating dynamic graphs of calling networks in
Cortes et al. (2003), we could use an exponentially weighted moving average (EWMA) of the DataSpheres to define the current state of a data stream. As a point arrives from the stream, our method classifies it and adds it to the appropriate bin of the DataSphere. Once the end of the section of the data stream is reached (the end of the week, for example), we stop adding to the DataSphere and create a new one for the next time period. We could, however, take an EWMA of the labels of each point in the data stream. More recent data would contribute more weight to the bin counts than would older data, and comparing this weighted DataSphere to the reference would tell us whether the “current state” had changed, and in which ways.

Another modification to our method could be to define the center and covariance of the reference data stream by the EWMA center and covariance matrix. A situation where this might be preferable to giving equal weight to each point in a short reference data stream (one week) is when the reference data stream is quite large, perhaps containing a year’s worth of data. Instead of just using a shorter, more recent data stream as the reference, the EWMA of a long stream would allow for some information from the past to contribute while still giving higher weight to the more recent data. This might result in obtaining a truer center and covariance for the data stream.

6.2 Agent-based Modeling of Dynamic Social Networks

In Epstein and Axtell (1996) the authors endow objects in a computer simulation, called agents, with certain properties and designate rules governing their behavior and movement in the “sugarscape.” The basic goal of each agent is to
find sugar to enable movement, search for a mate, and reproduce. The agents
do this—while interacting with each other—according to the rules of the simula-
tion model. Remarkably, by following just a few simple rules, the agents “create”
an artificial society, complete with resource gathering, immigration, trade, sexual
reproduction, cultural transmission, warfare, and disease propagation. This Sug-
arscape model of agent-based simulations demonstrates how specific actions on a
micro-level can produce global phenomena.

In unfinished work sketched below, we create dynamic social networks between
agents. The agents form ties with each other which can change over time, thus
making their social network dynamic. The rules governing the behavior of the
agents are constructed to mimic the terms in the $p_1$ model of Holland and Lein-
hardt (1981) and the terms in the distance model of Hoff, Raftery, and Handcock
(2002).

6.2.1 Dynamic social networks

The general idea behind our agent-based modeling of dynamic social networks is
that over time, agents become more like those with whom they share ties. The
more an agent shares personal characteristics with another agent, the more likely
the pair is to create a tie, and once linked, the more personal characteristics they
begin to share.

We set our agents in a grade-school classroom and henceforth consider them
to be students. Ties between students are **Friendships**, which are directed re-
lationships and are created according to a probability rule. The probability $p_{ijt}$
of a **Friendship** between student $i$ and student $j$ at time $t$ is a function of fixed
parameters (rules) $\eta_{ij}$ and the distance $|z_{it} - z_{jt}|$ between $i$ and $j$ in social space.
at time $t$.

We consider a 2-dimensional social space and, for convenience, label the axes *Sports* and *Fashion*. Each student starts off with a position in social space, i.e., has a value measuring his or her degree of “sportiness” and “sense of fashion.” Students who enjoy sports but do not care much about clothes will have positions in social space close to each other, and will thus be more likely to become friends with each other than with students farther away in social space.

The social network becomes dynamic by introducing a rule that students’ positions in social space move a fraction of the distance towards the average of their friends’ positions in social space. We encode into the agent-based model the idea that if a student’s friends all play sports, that student will become more likely to play sports in the future. Students modify their social space characteristics to be more similar to their friends’ characteristics.

### 6.2.2 Rules incorporating higher-order effects

We add rules to the simulations governing students’ behavior with respect to *Friendship* ties and movement in social space. The flexibility of agent-based modeling allows for many types of behavior seen in real social networks—such as differential popularity, reciprocity, and transitivity—to be easily incorporated into the social networks created by our simulations.

**Density parameter**

Recall from Chapter 1 that the Erdős-Rényi model (1959) governs the number of ties in a network through the density parameter $\theta$. In our first agent simulation we define the probability rule governing the creation of a *Friendship* tie between
student $i$ and $j$ at time $t$ to be:

$$\text{logit}(p_{ijt}) = \theta - |z_{it} - z_{jt}|. \quad (6.2)$$

As it turns out, rules governing the size of social space and the students’ initial positions in social space are also important for the outcomes of this basic simulation. If students are initially placed too far apart in social space, and if $\theta$ is not large, then no students will create Friendship ties and therefore no movement will occur in social space. For example, if two students are 5 units apart in social space, the probability of a tie from one to the other is only 0.0067 for $\theta = 0$.

The rules so far:

1. Initial social space placement;
2. Density parameter $\theta$;
3. Probability rule (as in Equation 6.2);
4. Fraction of movement towards friends in social space.

A complete list of rules used in our agent simulations will follow at the end of this section.

**Differential popularity**

An attractive feature of the $p_1$ model is the incorporation of the first order effect of differential popularity in the form of the productivity parameters $\alpha$ and attractiveness parameters $\beta$. We add a rule governing this first-order effect and call it “Charisma”. We define the charisma of a student $i$ for sending ties to be $\alpha_i$ and
for receiving ties to be $\beta_i$. Students with high values of charisma are more likely to form *Friendship* ties, as seen from this probability rule:

$$\text{logit}(p_{ijt}) = \theta + \alpha_i + \beta_j - |z_{it} - z_{jt}|,$$

where

$$\sum_i \alpha_i = \sum_i \beta_i = 0.$$

(6.3)

**Reciprocity**

A simple way to govern the degree of reciprocity in the *Friendship* social networks is to create an additional rule for the agent simulations. Assuming that the tie from $i$ to $j$ at time $t$—given by the probability $p_{ijt}$—has already been created (or not), we can add the presence (or absence) of this tie into the probability rule for $p_{jit}$. Let $y_{ijt} = 1$ indicate the presence of this tie, then:

$$\text{logit}(p_{jit}) = \theta + \alpha_j + \beta_i + \rho y_{ijt} - |z_{jt} - z_{it}|.$$

(6.4)

The parameter $\rho$ controls the force of reciprocation.

**Transitivity**

Because of the effect of social space, transitivity is already modeled in our simulations (see Section 1.2 and Figure 1.1). However, we could explicitly encode transitivity into our agent-based model through the creation of another rule. If student $i$ is friends with $j$, and $j$ is friends with $k$, the transitive tie is from $i$ to $k$. The new probability rule increasing or decreasing the chances of this transitive tie through the value of parameter $\tau$ is given by:

$$\text{logit}(p_{ikt}) = \theta + \alpha_i + \beta_k + \tau y_{ijt} y_{jkt} - |z_{it} - z_{kt}|.$$

(6.5)
Covariates

It should be apparent that our agent-based model can incorporate a wide variety of higher-order effects through the creation of simple rules governing the simulations. We can also quite easily add covariate information to the probability rules. For example, in Section 6.2.3 below we add gender covariates to the model. If students $i$ and $j$ are of the same gender, then the gender indicator $g_{ij} = 1$. We create a new rule to allow for gender to affect the probability of Friendship ties:

$$\logit(p_{ijt}) = \theta + \alpha_i + \beta_j + \gamma g_{ij} - |z_{it} - z_{jt}|. \quad (6.6)$$

Our aim in this section was to highlight some of the possibilities for creating rules in agent-based models of dynamic social networks. A list of the rules we discussed follows. In the next section we briefly discuss using agents to model a social network with bivariate ties.

Rules:

1. Initial social space placement;
2. Density parameter $\theta$;
3. Charisma popularity parameters $\alpha$ and $\beta$;
4. Reciprocity parameter $\rho$;
5. Transitivity parameter $\tau$;
6. Gender or other dyadic covariate parameter $\gamma$;
7. Probability rule (how to assimilate above rules);
8. Fraction of movement towards friends in social space.
6.2.3 Bivariate extension

We extend our social networks to the bivariate case, modeling two correlated relationships between individuals: Friendship and Respect.

In our first hypothetical, bivariate situation we have one classroom of grade-school children. The class has 10 male (in green numbers) and 10 female (lavender) students. Each directed pair of students can separately have a Friendship and Respect relationship. Friends do not have to respect each other, and those who respect each other need not be friends. However, we typically assume a positive correlation between these two relationships when creating the rules of the agent-based model. Figure 6.1 shows the graph of the network at a particular time point for one of the simulations with a small, positive correlation between Friendship and Respect. It is also evident from this graph that student 2 has been endowed with high charisma since his in- and out-degrees are very high compared to the other students.
Figure 6.2: Graph of student friendship (red) and respect (blue) relationships with importance placed on the gender covariate. Green numbers indicate male students while lavender numbers indicate females.

In Figure 6.2, two students of the same gender have a higher probability of Friendship, but a lower probability of a Respect relationship. In this simulation, there was no correlation between the relationships, and charisma was distributed uniformly.

6.2.4 Discussion

The goal of this research is to apply agent-based modeling techniques to real data to answer questions about the dynamic interactions of individuals. We wish to consider the individuals as agents, and then—given the data—infer the rules governing their interactions. Deciding between rules and fitting real data to agent-based models presents an exciting future research direction.
Appendix A

Additional Social Space and Pickiness Plots

Figures A.1-A.4 show the social space plots for the AA elephant family during the wet season and boxplots of the size of the social space pairwise effect for mother/daughter, sisters, and unrelated pairs. In addition, the figures show boxplots of the pickiness of the elephants which can be viewed as a measure of how well the model fits the elephants. As covariates were added to the model, the positions in social space became less clustered, and the size of the inner products decreased. The addition of informative covariates lessened the need for social space to explain the idiosyncrasies in the social network relational data.

Figures A.5-A.8 show boxplots of the individual pickiness of the elephants in the AA family in the wet and dry seasons. No elephant was especially picky. The wet season did showed higher pickiness than the dry season for the Null and DNA-only models. However, when the Mother/Daughter and Sisters indicators were added to the model, pickiness decreased—indicating much better fits to the data in the ModaSis and Full kinship models.
Figure A.1: Left: $\hat{Z}$ positions in social space in the wet season for the Null model 3.7. Four clusters of mother/daughter pairs are apparent. Right: boxplots show the size of the posterior inner products for various types of pairs of elephants in family AA.

Figure A.2: Left: $\hat{Z}$ positions in social space in the wet season for the DNA-only model 3.8. Four clusters of mother/daughter pairs are still apparent. Right: boxplots show that including DNA relatedness did not greatly affect the size of the posterior inner products for the various types of elephant pairs.
Figure A.3: Left: $\hat{Z}$ positions in social space in the wet season for the ModaSis model 3.9. The four clusters of mother/daughter pairs are much more spread out. Right: boxplots show that including the Mother/Daughter and Sisters indicators caused the sizes of the posterior inner products for the various types of elephant pairs to be similar.

Figure A.4: Left: $\hat{Z}$ positions in social space in the wet season for the Full kinship model 3.10. The four clusters of mother/daughter pairs are much more spread out. Right: boxplots show that including all the genetic and kinship covariates caused the sizes of the posterior inner products for the various types of elephant pairs to be similar.
Figure A.5: Pickiness of the elephants in the wet and dry seasons in the Null model.
Figure A.6: Pickiness of the elephants in the wet and dry seasons in the DNA-only model.
Figure A.7: Pickiness of the elephants in the wet and dry seasons in the ModaSis model.
Figure A.8: Pickiness of the elephants in the wet and dry seasons in the Full
kinship model.
Bibliography


information-theoretic approach to detecting changes in multi-dimensional data streams. (Interface '06), Pasadena, CA.


Biography

Eric Vance was born in Modesto, California on October 12, 1975. He received an A.B. degree from the University of California, Berkeley in May of 1997 with a triple major in Mathematics, Economics, and Statistics. After traveling around the world three times backpacking through over 65 countries, he joined the Institute of Statistics and Decision Sciences (ISDS) as a Ph.D. student in August of 2002 at Duke University in Durham, North Carolina.

Now a student in the renamed Department of Statistical Science, Eric has published two papers so far. The first is “Moment estimation in delay tomography with spatial dependence,” with Ian Dinwoodie, found in Performance Evaluation, Volume 64, August 2007. The second is “Age, musth and paternity success in wild male African elephants, Loxodonta africana,” with Julie A. Hollister-Smith, Joyce H. Poole, Elizabeth A. Archie, Nicholas J. Georgiadis, Cynthia J. Moss, and Susan C. Alberts. This appeared in Animal Behaviour, Volume 74, August 2007. A third paper titled “Social Networks in African Elephants,” with Elizabeth A. Archie and Cynthia J. Moss is in review.

Upon matriculating at Duke in 2002, Eric Vance received the James B. Duke Fellowship. While at Duke he has been an active member of GPSC, the Basketball Campout Committee, and the Student Health Insurance Advisory Committee; and has served as a student representative to the Board of Trustees Undergraduate Affairs Committee.

In August of 2008, Eric will join the faculty of the Department of Statistics at Virginia Tech in Blacksburg, Virginia as the Director of the newly-created Laboratory for Interdisciplinary Statistical Analysis.