

NETWORK INFERENCE FROM GROUPED DATA

by

Charles W. Weko III

A Dissertation

Submitted to the
Graduate Faculty

of

George Mason University

In Partial Fulfillment of



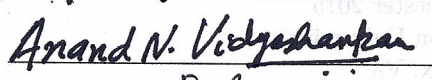
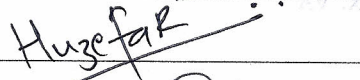
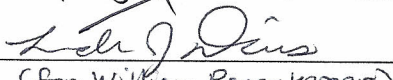

The Requirements for the Degree

of

Doctor of Philosophy

Statistical Science

Committee:

	Dr. Yunpeng Zhao, Dissertation Director
	Dr. Daniel B. Carr, Committee Member
	Dr. Anand N. Vidyashankar, Committee Member
	Dr. Huzefa Rangwala, Committee Member
 (for William Rosenberger)	Dr. William F. Rosenberger, Department Chair
	Dr. Kenneth S. Ball, Dean, Volgenau School of Engineering

Date: 29 APR 15

Spring Semester 2015
George Mason University
Fairfax, VA

Network Inference from Grouping Data

A dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy at George Mason University

By

Charles W. Weko III
Master of Science
Naval Postgraduate School, 2009
Bachelor of Science
Rose-Hulman Institute of Technology, 1995

Director: Dr. Yunpeng Zhao, Professor
Department of Statistics

Spring Semester 2015
George Mason University
Fairfax, VA

Copyright © 2015 by Charles W. Weko III
All Rights Reserved

Dedication

This dissertation is dedicated to my wife. I owe you one PHD.

Acknowledgments

I would like to thank the following people who made this work possible:

- Dr. David Alderson of the Naval Postgraduate School for providing early guidance and critical professional advice.
- Sam Mendelson and other GMU Mathematics Department Graduate Students for their patient assistance.

Table of Contents

	Page
List of Tables	viii
List of Figures	ix
Abstract	x
1 Introduction	1
1.1 Notation	1
1.2 Network Science	1
1.3 Network Inference	3
1.4 Data	4
1.4.1 Format of Grouped Data	4
1.4.2 Data Examples	8
1.5 Techniques of Network Inference	10
1.5.1 The Social Networks Perspective	10
1.5.2 Dyadic Behavior	11
1.5.3 Affiliations Networks	12
1.5.4 Inference of Transmission Paths from Co-Occurrences	16
1.6 Summary	17
2 Modeling Grouped Data with Star Models	18
2.1 Motivation	18
2.2 Star Models	18
2.2.1 Data Generation Process	20
2.3 Known Star Model	20
2.3.1 Maximum Likelihood Estimators	21
2.4 The Symmetric Star Model	22
2.4.1 Maximum Likelihood Estimate of the Symmetric Star Model	25
2.4.2 EM Algorithm	29
2.4.3 Symmetric Star Model EM Algorithm	30
2.5 Simulation Results	33
2.5.1 Generating the True Model Parameters	33
2.5.2 Simple Example of SSM	34

2.5.3	Convergence of SSM	34
2.5.4	Visualization	36
2.6	Data Analysis	38
2.6.1	Dream of the Red Chamber	38
2.6.2	Dolphins network	46
2.6.3	Self-sparsity	49
2.7	Conclusion	51
3	Inducing Sparsity with the Penalized Rho Star Model	53
3.1	Introduction	53
3.1.1	Motivation	53
3.2	Methodology	55
3.2.1	Linear Penalty	55
3.2.2	Formulation of Optimization Problem	55
3.2.3	Algorithm	56
3.3	Selecting Tuning Parameters	58
3.4	Simulation Studies	59
3.4.1	Toy Example	60
3.4.2	BIC for Increasing η with Fixed Number of Parameters	63
3.4.3	Estimating Parameters for Large Sparse Networks ($n = 50$)	63
3.5	Data Analysis	64
3.5.1	Penalized Rho Star Model for Dolphins	64
3.5.2	Penalized Rho Star Model for <i>Dream of the Red Chamber</i>	68
3.6	Conclusion	73
4	Future Work	74
4.1	Identifiability of Star Models	74
4.1.1	Definition of Multivariate Bernoulli Random Variables	74
4.1.2	Definition of a Finite Mixture Distribution	75
4.1.3	Identifiability of Finite Mixtures of Multivariate Bernoulli Random Variables	76
4.1.4	Identifiability	76
4.1.5	Necessary and Sufficient Condition for Identifiability	77
4.2	General Grouping Model	77
4.3	Enumeration of Connected Graphs	78
4.4	Additional Work	81
4.4.1	Singleton Free Datasets and Alternative Handling of Singletons	81
4.4.2	Effect of Removing Individuals	81

4.4.3	Behavior Transitions Across Time	82
4.4.4	Social Group Evolution	82
4.4.5	Non-Member Grouping Effects	82
4.4.6	Measurement Error	82
4.4.7	Population Bounding	83
4.5	Outreach	83
A	Notation	85
B	Properties of Grouped Data Under Star Models	87
C	Runtime of Star Models	89
	Bibliography	90

List of Tables

Table	Page
1.1 Grouped Dataset for Six Children and Three Birthday Parties	3
1.2 Relationship of Observations	5
2.1 Example True Adjacency Matrix	20
2.2 General results	35
2.3 Average and Standard Deviation of Mean Absolute Error as Observations Increase	36
2.4 Values of ρ for characters in <i>Dream of the Red Chamber</i>	42
2.5 Percentiles of Standard Deviation in \hat{A} estimated by SSM for <i>Dream of the Red Chamber</i>	43
2.6 Relationships of Lin Daiyu and Xue Baochai to other characters in <i>Dream of the Red Chamber</i>	46
2.7 Two Different estimates of ρ for Dolphins	49
3.1 Example of a Population With Sparse ρ 's	61
3.2 Frequency Table Low Observations Example	61
3.3 Bayesian Information Criterion as η Increases	62
3.4 Estimated Adjacency Matrix for $\eta = 1$	62
3.5 Estimated Adjacency Matrix for $\eta = 500$	63
3.6 Average and Standard Deviation of Mean Absolute Error as Observations Increase	64
3.7 $\hat{\rho}_1$ for Dolphin Data	66
3.8 $\hat{\rho}_2$ for Dolphin Data	67
3.9 $\hat{\rho}$ for Reduced <i>Dream of the Red Chamber</i> Data	71
3.10 $\hat{\rho}$ for Reduced <i>Dream of the Red Chamber</i> Data	72
4.1 Adjacency Matrix for N_{100}	79
4.2 Upper Triangular Matrix for N_{100}	80
A.1 Notation Table	86
C.1 Average Runtime in Seconds	89

List of Figures

Figure	Page
1.1 Two Different Ways Nodes $\{v_1 \dots v_6\}$ Can Be Connected	6
2.1 Example of a Star Graph	19
2.2 Comparison of Estimation Techniques	32
2.3 Distribution of Beta Function	34
2.4 Symmetric Star Model Estimates Improve as T Increases	38
2.5 Estimates for <i>Dream of the Red Chamber</i>	40
2.6 Estimates Brothers in <i>Dream of the Red Chamber</i>	41
2.7 Estimates for <i>Dream of the Red Chamber</i> Represented by Plot Area	44
2.8 Overlaid Estimates for <i>Dream of the Red Chamber</i> (<i>O-Black, HWI-Red, SSM-Green</i>)	45
2.9 Conventional Estimates for Dolphins	47
2.10 Adjacency Matrix Estimates for Dolphins	48
3.1 Number of Parameters as n_o Changes	60
3.2 Penalized Adjacency Matrix Estimates for Dolphins ($\eta = 20$)	68
3.3 Estimated Relationships for <i>Dream of the Red Chamber</i> ($\eta = 500$)	70

Abstract

NETWORK INFERENCE FROM GROUPING DATA

Charles W. Weko III, PhD

George Mason University, 2015

Dissertation Director: Dr. Yunpeng Zhao

In the past two decades, the interest in network analysis has expanded rapidly. Most network analysis methods start from observed network topology. However, network structure is not directly observed in many fields, especially in social sciences. Thus, a methodology for inferring implicit network structure is required to effectively apply network analysis. One area of research involves the inference of network structure from *grouped data*. Grouped data records the manner in which a population forms subsets or smaller groups.

In the existing social science literature, inference of network structure from grouped data is performed using descriptive statistics. Researchers have defined a collection of measures to quantify the strength of interaction among members of the population and use these measures to infer a network structure. Classic examples of these measures include the *co-occurrence matrix* and the *half weight index*.

This dissertation defines stochastic models called Star Models for modeling group formation. Each observed group is assumed to have a single leader who has brought the group together. We derive maximum likelihood estimators for the model parameters. The parameter estimation of Star Models fits naturally into the framework of the Expectation-Maximization algorithm. The resulting parameters have an intuitive interpretation as the

assertiveness of individual nodes and their popularity within the population.

We apply the new methods to simulated data to compare our results with the existing methods. Additionally, we apply these techniques to the famous 18th century Chinese novel, *Dream of the Red Chamber* to demonstrate the superior performance of the Star Model.

The number of parameters for Star Models is order $O(n^2)$, where n is the size of the network. This presents a challenge that the model requires a large number of observations to accurately estimate parameters even when the network size is moderate. In practice, the number of observations may be limited. To resolve this issue, we further propose a technique called the *Penalized Rho Star Model*, which is based on the assumption that only a few members of the population can generate groups even though the total population is large. Simulation studies and data analysis are performed to compare the Star Model to the Penalized Rho Star Model.

The structure of this dissertation begins with a literature review of existing techniques. Chapter 2 presents the Star Model and shows that assuming that relationships are symmetric leads to an identifiable version which we call the Symmetric Star Model. In Chapter 3, we present the Penalized Rho Star Model and show that this technique introduces sparsity into the parameters and improves estimation of large networks. We conclude with a discussion of a broad number of extensions of the existing work.

Chapter 1: Introduction

1.1. Notation

A network encodes fundamental information about how the elements of a system are connected to each other. Each network consists of a set of n discrete *verticies*, or *nodes*, $V = \{v_1, \dots, v_n\}$. In most applications, including those presented in this dissertation, the number of verticies is assumed to be known and fixed.

Verticies are related to each other by an $n \times n$ *adjacency matrix*, A . Each element of the adjacency matrix represents the relationship between an ordered pair of verticies. Specifically, A_{ij} represents the relationship between the pair of nodes, $\{v_i, v_j\}$. The elements (possibly weighted) of the adjacency matrix are often called *edges*, *links*, *arcs*, or *relationships*. The adjacency matrix may be symmetric, i.e., $A_{ij} = A_{ji}$, for all i, j , or asymmetric.

Appendix A contains a list of notations used throughout this dissertation.

1.2. Network Science

As documented in a 2006 National Research Council report (Alderson, 2008), the research field called *network science* is focused on an interdisciplinary view of complex network systems. Broadly speaking, the scientific questions that are interesting to researchers address *network structure*, *universal laws* governing structure, and *vulnerabilities* inherent in complex networks. The applications include the Internet and World Wide Web, friendship and communication networks in the social sciences, food webs and gene-regulatory networks in biology, network games in economics, and many others.

Network science generally takes *random graphs* as a foundation for research. Random

graph models specify a probability distribution over a collection of graphs. The precise analytical characterization of many of the summary structural measures of these models, such as clustering coefficients and degree distributions makes random graph models ideally suited to explaining observed network structures (Kolaczyk, 2009). Some well studied random graphs models are the Erdős-Renyi graph, the preferential attachment graph, and the configuration model (Newman, 2011).

The *Erdős-Renyi graph model*, or the *classical random graph model*, is probably the simplest model in network science. This model places an equal probability on all edges in a network (Erdos and Renyi, 1960). This can be done in two different ways. In the first technique, the $G(n, p)$ model, a graph is defined to have n nodes and the probability of each edge is defined to be p . Each edge is included in the network independently of every other edge; therefore, the number of edges in the graph will have a binomial distribution. In the second technique, the $G(n, m)$ model, the number of edges in the network is set by m and the set of edges in the network is selected uniformly from any of the possible networks with m edges.

Despite its simplicity, the *Erdős-Renyi graph model* is not able to model some important characteristics observed in real-world networks, especially the heavy-tailed distribution of degrees. The *degree* of a node is the number of edges originating from the node connected to the rest of the nodes. The *preferential attachment model* is proposed to model the observed power-law degree distributions. The preferential attachment model assumes that observed network structure is the result of a generation process where new nodes entering the network are more likely to form links to well connected nodes than nodes with few connections (Barabasi and Albert, 1999). The simplest form of the preferential attachment model begins with a pair of connected nodes. As a new node is added to the network, it forms a single link with one of the existing nodes. The node selected to form this link is selected with a probability that is exactly proportional to the degree of the node.

In the *configuration model*, the degree of each node in the network is predetermined. The generation process randomly links nodes based on their fixed degree. This type of model

is particularly useful for mimicking networks with arbitrary degree distributions (Newman, 2011).

Random graphs prove foundational when modeling known network structure by providing a probability distribution over all possible graphs.

1.3. Network Inference

An alternative research area is *network inference*. Network inference begins with some observed network behavior and then attempts to infer the network’s structure from that behavior. In contrast to the models for network formation discussed in the last section, models for network inference should provide a probability distribution over all possible observed behaviors.

Social Network Analysis: Methods and Applications (Wasserman and Faust, 1994) introduces the problem of inferring the relations among a collection of children based on their attendance at birthday parties. In this introductory dataset, the names of children represent the column headings, and the birthday parties represent the row headings. If a specific child attended a party, he or she is represented by the numeral 1, and 0 indicates that he or she did not.

Table 1.1: Grouped Dataset for Six Children and Three Birthday Parties

Party	Child					
	Allison	Drew	Eliot	Keith	Ross	Sarah
1	1	0	0	0	1	1
2	0	1	1	0	1	1
3	1	0	1	1	1	0

In this dissertation, a collection of individuals observed in the same sample is called a *group*, and a dataset of these observations is referred to as *grouped data*. In Wasserman

and Faust’s example, each party defines a group and the set of all parties is the grouped data. Two individuals are said to *co-occur* if they appear in the same group. For example, in Table 1.1 Ross and Sarah co-occur in Parties 1 and 2 but not in Party 3.

In order to make inferences about this population, a method of finding the most likely network structure that could have generated the observations is needed. Clearly, classical network models cannot be directly used to accomplish this task because random graphs model the generation of the network itself rather than the generation of the observed groups. Therefore, in this dissertation, we introduce models for the generation of grouped data and apply maximum likelihood estimation to infer the latent network structure.

1.4. Data

1.4.1 Format of Grouped Data

Grouped data consists of observations $\{V^{(t)}, t = 1, \dots, T\}$, which are subsets of a global population V , taken at different time points, $V^{(t)} \subset V$.

The observed subset $V^{(t)}$ can be coded by an n length row vector $G^{(t)}$ where

$$G_i^{(t)} = \begin{cases} 1 & \text{if } v_i \in V^{(t)} \\ 0 & \text{if } v_i \notin V^{(t)} \end{cases}$$

The number of members in the global population is n and the number of nodes in group $G^{(t)}$ is denoted by $n_t = \sum_i G_i^{(t)}$. The full set of observations is denoted by $G = \{G^{(1)}, \dots, G^{(T)}\}$.

Note that this data structure is only an abstraction of “groups”. The criteria that researchers use to define a “group” varies between research situations. For example, a researcher studying a corporation may define a group as the people attending the same meeting. Another researcher studying marine mammals may define a group based on the physical distance between individuals.

This dissertation focuses on the situation where only one group is observed at a single point in time. We call this grouped data *partially observed*. There are two situations which are excluded from this focus. First, *fully observed* populations are present when all members of V can be observed at the same time. This kind of data is often present in studies of controlled populations such as captive chimpanzee populations (Schel et al., 2013). Secondly, *simultaneously observed* populations are present when multiple groups can be observed at the same time without observing the full population. For example, suppose a researcher observes a population of children on a school playground. When the researcher focuses on a large play structure, some children may be hidden from view and at the same time children who are not interacting with each other may be in view. This type of data introduces aspects of measurement error that are addressed briefly in Chapter 4.

An important point concerning partially observed populations is that no inference can be made about the grouping behavior of the unobserved nodes in observation t . That is, when two nodes are simultaneously unobserved, we cannot tell whether they are together or apart in a different group. Table 1.2 summarizes that an inference about the co-occurrence of two nodes can only be made when one of the nodes is observed.

Table 1.2: Relationship of Observations

	Node v_i Observed	Node v_i Not Observed
Node v_j Observed	v_i and v_j together	v_i and v_j apart
Node v_j Not Observed	v_i and v_j apart	unknown

The observed group $G^{(t)}$ is assumed to represent a *component* of an adjacency matrix, $A^{(t)}$ generated by A . A component is a subset of the vertices of a network such that there exists at least one path from each member of that subset to each other member, and such that no other vertex in the network can be added to the subset while preserving this property.

$A^{(t)}$ is an unweighted adjacency matrix where $A_{ij}^{(t)} = 1$ with probability A_{ij} and $A_{ij}^{(t)} = 0$ otherwise. The central challenge of network inference is that there are multiple $A^{(t)}$ which could have produced the same $G^{(t)}$. For an edge, A_{ij} , where nodes v_i and v_j are in $G^{(t)}$, we say the edge is *active* if $A_{ij}^{(t)} = 1$ and *inactive* otherwise.

As an example of the way in which a single observed group may be the result of different behaviors, consider a population of ten nodes. In sample t , $G^{(t)} = \{1, 1, 1, 1, 1, 1, 0, 0, 0, 0\}$ is observed. That is, nodes 1-6 are observed to co-occur. Two possible grouping behaviors that could have caused this observation are shown in Figure 1.1. In Figure 1.1a, the group is formed by a star subgraph. In Figure 1.1b, the group is the result of a ring subgraph.

In Figure 1.1a, we can see that A_{15} is active and that A_{56} is inactive. In practice, whether a link is active is unknown and must be inferred from the data.

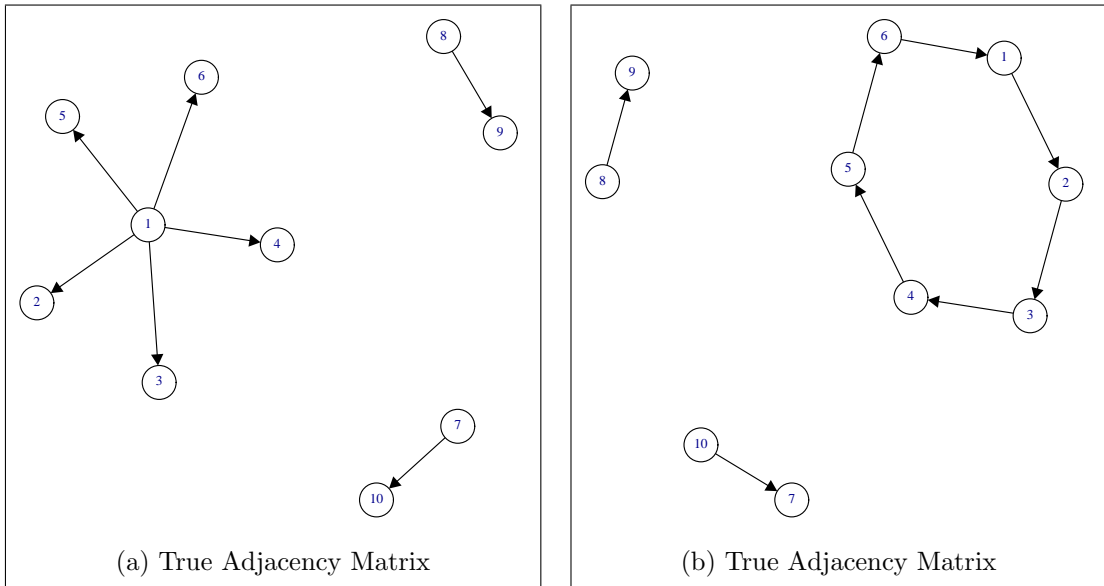


Figure 1.1: Two Different Ways Nodes $\{v_1 \dots v_6\}$ Can Be Connected

The terms active and inactive have important implications for inferring the network. Two nodes which co-occur do not necessarily have an active relationship between them;

it is only known that some set of active relationships have connected them. However, we can make a stronger statement about two nodes, $\{v_i, v_j\}$, when only one node is observed. These two nodes do not have any active set of relationships connecting them and $A_{ij}^{(t)} = 0$.

As a simple example of a set of partially observed grouping data, consider a population of four individuals ($n = 4$). Five samples of these individuals are observed ($T = 5$). This grouped dataset can be encoded in the 5×4 matrix shown below.

$$G = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Throughout this dissertation, groups are assumed to be independent and identically distributed within the dataset G . That is, G^t is not the result of a transformation on G^{t-1} , and we could reorder the elements of G without effecting our inference. This assumption is consistent with techniques that are used in social network analysis, which require sightings to be spaced at least one day apart to provide independence (Bejder et al., 1998).

Under the assumption of independence, a method of data compression is to build a frequency table, F , which counts the number of times that a unique group appears. This reduces the number of rows in the grouping data from T to at most 2^n . As an additional means of organizing the data, we sort the groups from least number of members to highest number of members.

Continuing with our example G above, the resulting frequency table appears below. Consider the third row of G where only node v_1 is observed. This group becomes the first row of F , and we place a 1 in the fifth column because this group is observed only once. Compare this to the groups $t = 1$ and $t = 4$. These groups have the same membership and appear in the frequency table on the second row with a 2 in the fifth column.

$$F = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 2 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

1.4.2 Data Examples

Two datasets are used in this dissertation to demonstrate inference of network structure.

Dream of the Red Chamber

As noted by Kolaczyk (2009), an important problem of parameter estimation, in the context of network inference, is that there is usually no way to verify the extent to which the estimate matches reality for a real world dataset. That is, there is no “ground truth” or “golden standard” to evaluate the performance of the estimated results against. Therefore, to test the performance of Star Models, it is useful to analyze data where there is some knowledge of the relationships between nodes.

To this end, we constructed a dataset of characters from the 18th century Chinese novel *Dream of the Red Chamber*, also known as *The Story of the Stone*. Since novels contain a qualitative social structure that is familiar to readers, the results of quantitative analysis can be compared to this standard. This novel was selected because the relationships between the characters are subtle and complex. In this way, the story presents a challenge to the task of estimating the social structure of the characters.

Traditional approaches to building datasets from novels require carefully reading the text and identifying dyadic interactions between characters based on criteria established by the researchers, e.g., character *A* has a conversation with character *B* (MacCarron and Kenna, 2013). Though this method may construct high quality datasets, identifying the dyadic interactions requires readers who are familiar with the novel’s language and who have time to build the datasets. Since *Dream of the Red Chamber* is written in classical

Chinese and the English translation runs over 2,600 pages, directly generating the dataset would be excessively time consuming.

To address the novel's size, the dataset is built using text-mining. We define a group as characters who co-occur in the same paragraph. Paragraphs containing no character names are ignored.

We analyze the relationships of 29 important characters. The character names are based on their original pinyin pronunciations and the David Hawkes translation (Hawkes, 1974). A publicly available Chinese version of the novel was used for text-mining.

This complete novel contains 120 chapters, but we focus on the first 80 chapters because it is commonly believed that the last 40 chapters are written by a different author and may not reflect the original themes of the novel. The resulting dataset has 1,389 observations of groups containing at least one of the 29 characters.

South Island Hector's Dolphin Data

This data set consists of observations of Hector's dolphins that were taken over 1996-1997 in the South Island of New Zealand's inshore waters. The full population is estimated to contain 50-70 individuals.

Hector's dolphins are most often observed in groups of two-to-eight individuals. These groups often fuse together and split up over periods of several hours. The researchers considered individuals associated if they were members of the same group or cluster of groups. Groups of dolphins were considered part of the same cluster of groups if groups merged in the time span when observations were being taken.

Observations were recorded by photographs. Photography is a noninvasive tool that is frequently used to study the social structure of cetaceans and other social animals. Groups are defined entirely based on photographic records. That is, individuals seen, but not photographed, are not included in the observed group (Bejder et al., 1998).

The dataset focuses on the grouping of 18 individuals. A total of 40 observations are taken of the population. Throughout this dissertation, this dataset is used as an example

of a situation where there are very few observations relative to the number of members in the population under study.

1.5. Techniques of Network Inference

1.5.1 The Social Networks Perspective

The majority of literature dealing with grouped data comes from social network research where it is common to collect data on subgroups of a global population. This approach is founded on a qualitative approach to social interaction known as the *social networks perspective*. This perspective was introduced in a book written in 1934 by J.L. Moreno (Moreno, 1934).

The social networks perspective begins by recognizing that individual actors (nodes) and their actions are interdependent, rather than autonomous and independent. Under this perspective, actors are connected by relational ties (edges) that serve as conduits for information or resources. Network models focusing on individuals view the network's structure as providing opportunities for, or constraints on individual action (Wasserman and Faust, 1994).

This perspective is nicely summed up in the American Psychological Association record review of Moreno's book. In the following passage, the editors point out that the strength of relationships (which is the parameter of interest in social network analysis) are essentially unobserved data. This suggests that the Expectation-Maximization algorithm will be useful in solving the network inference problem.

Wherever two or more people are functioning as a social group, that group not only consists of those individuals, but . . . the relations which maintain between them. It is these intangible, imponderable and invisible aspects of the situation which enable the mathematical sum of a certain number of individuals to function as a social group (PsycNet, 2012).

One question that the social networks perspective does not address is whether behavior actually demonstrates interdependencies between individuals or is merely the result of independent behavior. This issue suggests a “null model” for group formation which will be addressed in greater detail later in this dissertation.

1.5.2 Dyadic Behavior

In order to derive network structure from grouped data, the relations of interest must first be defined. One of the most straightforward ways to accomplish this is to define the relations in a manner that simplifies the derivation of the network structure. The easiest approach is to define *dyadic relationships* which exist only between a pair of individual nodes. Examples of these types of relationships include dating, mating, and financial transactions.

Dyadic relationships enforce a constraint on grouped data, $G^{(t)}$. Under a dyadic relationship, every group is a pair, $\sum_i G_i^{(t)} = 2 \quad \forall t$. This is a fairly artificial constraint which often has to be aggressively enforced.

Alternatively, if a dyadic relationship is defined such that it is directly observable by the researcher, network inference is reduced to a trivial task.

With this goal in mind, one of the most basic ways to infer the relationships between individuals is simply to conduct a survey in which each member of the population lists his favorite members of the population. This technique can be expanded into what is called the *Moreno-Davis Experiment* where respondents list their most favorite as well as least favorite members of the population (Freeman et al., 1989).

Unfortunately, this type of data collection is not always feasible. For example, the population under interest in the Moreno-Davis Experiment may not be willing or able to provide survey responses. This would clearly be the case when the population consists of inanimate objects like telecommunication routers, brain cells, or proteins. Further, animal populations are not capable of providing such detailed responses. Even human populations may not be able to describe their preferences either because they cannot answer such questions (e.g. young children) or because their responses are subject to bias (e.g. corporate

organizations).

One way to overcome the problem of survey taking is to use the *Bales Experiment*. In the Bales Experiment, the behavior of a population is discretely observed by researchers who classify dyadic behavior by a scheme of defined categories. This method has two distinct advantages. First, since the data is collected by the researchers under a set scheme of classification, the measurement of interactions is standardized for all members of the population. Secondly, since data collection occurs independent of the interaction of individuals, data can be collected over time to describe how the relationships unfold.

The Bales Experiment has two drawbacks. First, data collection is very time intensive, because it must be done manually. Therefore, the set of data collected is limited by the resources available for observation. Secondly, the characterization of observed behavior is subject to the interpretation of the researcher. While this limitation may be fine for observations on human beings, inferences about the interaction of non-human populations can be biased by the researcher's perceptions (Freeman et al., 1989).

1.5.3 Affiliations Networks

While dyadic behavior analysis simplifies the network inference problem by describing groups based on some pairwise relationship, *affiliation network data analysis* attempts to infer network structure by representing the interactions between a set of actors (usually greater than two) and a set of events (Wasserman and Faust, 1994).

Affiliation networks describe collections of actors rather than the ties between individuals. Affiliations are technically a combination of two sets of nodes, actors and events, which are often represented by bipartite graphs, but interest usually focuses on translating the information into a network that describes the connections between actors. This data is essentially equivalent to grouped data as described above.

The study of affiliation networks is based on the importance of individuals' membership in collective organizations. That is, multiple group affiliations form the basis for defining the social identity of individuals.

An aspect common to many of the views about affiliations is the idea that actors are first brought together by their common participation in social events. Joint participation then provides an opportunity for individuals to interact. Ultimately, interaction increases the probability that direct pairwise ties will develop between actors (Wasserman and Faust, 1994).

Note that this perspective does not suggest that co-occurrence implies a direct tie, but that the probability of a direct tie increases as a result of co-occurrence. This *joint participation perspective* is the intuitive basis for the maximum likelihood approach used in this dissertation.

One popular method for measuring affiliation networks is with a *co-membership matrix* or a *co-occurrence matrix*. The co-occurrence matrix can be represented either as a count of co-occurrence, $O^\#$, or as a frequency of co-occurrence, O . This method is described in detail by Wasserman and Faust; however, it enjoys such popularity that it is often applied without explanation or reference.

Each element of the co-occurrence matrix, $O^\#$, is calculated by summing the number of times that nodes v_i and v_j are observed together in the same group, $O_{ij}^\# = \sum_t G_i^{(t)} G_j^{(t)}$. Since co-occurrence is non-directional, the co-occurrence matrix is symmetric, $O_{ij}^\# = O_{ji}^\#$. To fully define the co-occurrence matrix, let $O_{ii}^\# = O_i^\#$, that is the number of times that v_i is observed in a group.

Once the co-occurrence matrix is calculated, a threshold, α , can be used to translate the co-occurrence matrix, $O^\#$, into an unweighted, undirected adjacency matrix, A . If $O_{ij}^\# > \alpha$, then $A_{ij} = 1$, otherwise $A_{ij} = 0$.

We can easily see that there is a problem with using the co-occurrence matrix to infer A . The problem is that the choice of the threshold α may be arbitrary and subjective.

This challenge was explored in detail by (Choudhury et al., 2010) when they observed that by generating a family of networks parameterized by different choices of the threshold, they produced networks with different structures. In particular, the authors used three

different thresholds that had been used by previous researchers. They found that the three resulting networks differed vastly in terms of density, connectivity, and clustering among other properties.

Proportional Affiliation

The co-occurrence matrix can be modified to produce a weighted, symmetric matrix, O , where the value of O_{ij} is the probability that nodes v_i and v_j will be observed to co-occur. To do this, simply divide the elements of $O^\#$ by the total number of observations, $O_{ij} = \frac{O_{ij}^\#}{T} = \frac{\sum_t G_i^{(t)} G_j^{(t)}}{T}$. For the purposes of visualization, we will define $O_{ii} = 1$. This notation can be generalized to any subset of V . For example, O_{ijkl} would represent the co-occurrence of nodes v_i, v_j, v_k, v_l , $O_{ijkl} = \frac{\sum_t G_i^{(t)} G_j^{(t)} G_k^{(t)} G_l^{(t)}}{T}$. This level of complexity is not used in this dissertation; however, the simple concept of O_i is used to represent the probability that node v_i is observed. To avoid confusion, the term *co-occurrence matrix* will refer to the frequency version of the matrix.

One might initially expect that the elements of O will estimate the probability that a relationship between two nodes is active. However, careful consideration of how co-occurrence occurs will show that O_{ij} is not equivalent to the probability that i and j are connected. $O_{ij}^\#$ will include every observation where the edge between nodes v_i and v_j is active, it will also include instances where the nodes of interest are interacting through a common node, or set of nodes, at the moment of observation. This technique implicitly assumes that nodes v_i and v_j are not interacting when they are unobserved.

These problems with the co-occurrence matrix are partially due to an unstated assumption about observed groups. That is, each observed group is assumed to be the result of a *clique*. This means that the nodes in the observed group are fully connected and an edge exists between every pair of nodes in the group. While this may be plausible for very small groups, it is highly unlikely in large groups.

Coincidence Index / Half Weight Index

L. R. Dice introduced a measure of association called the *coincidence index* to describe the co-occurrence of plant species within a habitat (Dice, 1945). This measure had the form:

$$\text{coincidence index}_{ij} = \frac{2 \sum_t G_i^{(t)} G_j^{(t)}}{\sum_t G_i^{(t)} + \sum_t G_j^{(t)}} \quad (1.1)$$

The *half weight index* (H) was introduced later (Cairns and Schwager, 1987). H for two individuals, v_i and v_j , is calculated as follows:

$$H_{ij} = \frac{\sum_t G_i^{(t)} G_j^{(t)}}{\sum_t G_i^{(t)} G_j^{(t)} + 0.5[\sum_t G_i^{(t)}(1 - G_j^{(t)}) + \sum_t (1 - G_i^{(t)})G_j^{(t)}]} \quad (1.2)$$

Since Equation 1.1 predates, is equivalent to, and is easier to interpret than H , this form is used for calculations. However, we will continue to use the term half weight index because H is more common in current literature.

H ranges in value from 0 to 1 and is symmetric. Values that are close to zero indicate that individuals do not co-occur under observation while values that are close to 1 indicate that individuals almost always co-occur when observed. H estimates the likelihood that nodes v_i and v_j will co-occur given that one of them is observed, i.e. $\mathbb{P}(G_i G_j = 1 | G_i = 1 \text{ or } G_j = 1)$. As a result, H is simply another attempt to describe the probability that an edge or relationship is active in a group.

There is a minor technical problem with this measure which needs to be addressed. When a given pair of nodes is never observed, H is undefined (i.e. $\frac{0}{0}$). In the classical applications of H , this is not a problem because if a node is not observed it is not included as an element of V . However, this is an issue for our work because we will want to estimate the variance of H by bootstrapping and this will occasionally result in a pair of nodes being completely unobserved. To deal with this, we define this situation as $H_{ij} = 0$.

Polish Manufacturing Company E-mail Estimator

Michalski et al. (2014) set out to compare the social network and the corporate hierarchy of a mid-sized manufacturing company. The stated purpose of this effort was to demonstrate a technique which could be used by corporate management to address the question of whether the employees were properly aligned within the organization.

In order to construct the social network of the company, the authors chose to use e-mail records. These records provided subgroups of the organization along with the sender of the e-mail; therefore, it could be reduced to dyadic behavior.

The authors proposed the following measure without proof or reference.

$$A_{ij} = \frac{\sum e_{ij}}{\sum e_i}, \quad (1.3)$$

where e_{ij} is an email sent from node v_i to node v_j , and e_i is an email sent from node v_i (Michalski et al., 2014).

As we will see, this measure is the maximum likelihood estimator for the Known Star Model proposed in Chapter 3.

1.5.4 Inference of Transmission Paths from Co-Occurrences

An alternative to the social network perspective or joint participation is the so-called *internally sensed network tomography* problem. In this problem, the network that is inferred is a complex communication system. Nodes represent routers and switches while edges represent the connections between these system components. In this scenario, transmissions are carried over the telecommunication network along a path between a source and terminal node (Rabbat et al., 2006).

In some cases, it is impossible to directly observe the order in which routers/switches handle the transmission. However, sensors are able to identify which routers/switches were active at roughly the same time. The internally sensed network tomography problem aims to recover network structure from un-ordered lists of network elements along transmission

paths.

Rabbat and his colleagues modeled the process of group formation as a random walk on the network subjected to an unknown permutation to account for the lack of order information. Treating permutations as missing data, they derived an *expectation-maximization* (EM) algorithm for estimating the random walk parameters.

While the model and the EM algorithm significantly simplified the problem, the reconstruction process grew exponentially in the length of each transmission path. For observations with many nodes, path length became long and the E-step became computationally intractable. For an observed group of n_t nodes, there were $n_t!$ different paths which could have connected the nodes. To address this challenge, the authors employed a polynomial-time *Monte Carlo EM* (MCEM) algorithm based on importance sampling to estimate the parameters of the model.

1.6. Summary

Not many methods for network inference from group data have been proposed in the current literature. One of the most basic simplifications is to define relationships in such a way that only pairs are considered. When analyzing groups with more than two nodes, heuristic methods using descriptive statistics such as the co-occurrence matrix and half-weight index are proposed but lack rigorous justification. Rabbat et al. (2006) proposed a model-based approach based on random walks. Since the number of possible paths on n nodes is factorial, a Monte Carlo EM algorithm has been applied to control the computational cost. In the next chapter, a model of group formation is introduced that is linear in n . This model allows for complexity in group formation without requiring the use of Monte Carlo solutions to estimate the parameters.

Chapter 2: Modeling Grouped Data with Star Models

2.1. Motivation

In Chapter 1, we saw that there is a long history of calculating descriptive statistics from grouped data to characterize the relationships within a population. These descriptive statistics are often mistaken to estimate the probability of links between nodes. However, there are no statistical models justifying the usage of these estimators. In other words, given the measures of a set of grouped data, there is no mechanism to simulate a new set of data with statistics that are similar to the original dataset.

The objective of this chapter is to develop statistical models of grouping behavior and find estimators for those parameters of the models. The approach will start with simple models related to the E-mail Estimator, introduced in Section 1.5.3, and then the constraints on the models will be relaxed to allow for increased sophistication.

2.2. Star Models

This chapter proposes a family of stochastic models for group formation called *Star Models*. This name captures the central assumption that each observed group is connected by a *star graph*. A star graph on n nodes is a connected graph with $(n - 1)$ edges and a single central node of degree $(n - 1)$.

Figure 2.1 represents what it means when every observed group is connected by a star graph. This example graphically depicts one of the six ways in which nodes 1 through 6 might have been connected. Note that if star graphs are the only way that observed groups can be connected, there are only n_t graphs which could have produced such an observation.

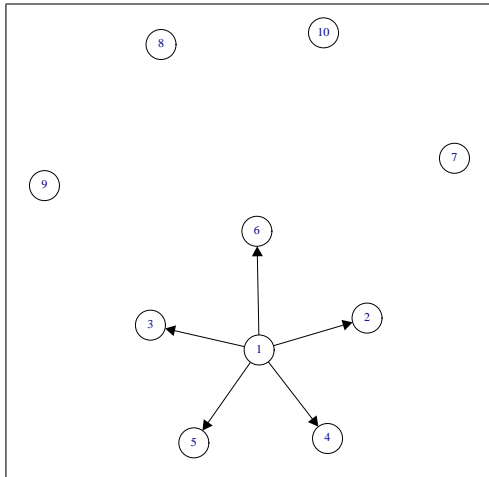


Figure 2.1: Example of a Star Graph

In each sample, an n length vector $S^{(t)}$ indicates the member of the group which is the center of the star graph.

$$S_i^{(t)} = \begin{cases} 1 & \text{if } v_i \text{ is the center of sample } t \\ 0 & \text{if } v_i \text{ is not the center of sample } t \end{cases}$$

$S_i^{(t)}$ is under the constraint that $\sum_i S_i^{(t)} = 1$.

$S^{(t)}$ is not necessarily observed. This point will be addressed in the following sections.

Star Models make the assumption that there is a latent network structure defined by the weighted adjacency matrix A which fully describes the preferences of members of the population. The adjacency matrix defines a relation where A_{ij} is the probability that the link between nodes v_i and v_j is active in sample t . We take $A_{ii} = 1$ for all diagonal values by convention, which captures the fact that v_i is always in the group that node v_i created.

2.2.1 Data Generation Process

For Star Models, the probability that a node chooses to form a star is given by $\rho = \{\rho_1, \dots, \rho_n\}$ where $\sum_i \rho_i = 1$. For each sample, a node v_k is selected with probability ρ_k to be the central node of a group. Node v_k then considers every other node in the population for inclusion in the group independently, with probability A_{kj} .

Qualitatively, this generation process means that the parameter ρ represents the “leadership” or “assertiveness” of the individual and A represents the “popularity” of an individual.

Table 2.1: Example True Adjacency Matrix

i	ρ_i	j				
		1	2	3	4	5
1	0.2	1.00	0.01	0.01	0.01	0.01
2	0.5	0.01	1.00	0.01	0.90	0.90
3	0.1	0.01	0.01	1.00	0.01	0.90
4	0.1	0.01	0.90	0.01	1.00	0.90
5	0.1	0.01	0.90	0.90	0.90	1.00

As an example, consider the adjacency matrix represented in Table 2.1. For sample $t = 1$, suppose that v_2 is selected as the central node. Based on the probability that v_2 will select the other nodes in the population, it would be very likely to observe a group containing nodes 2, 4, and 5. That is, $G^{(1)} = \{0, 1, 0, 1, 1\}$ and $S^{(1)} = \{0, 1, 0, 0, 0\}$.

From this generating process, it is possible to estimate commonly observed properties of grouped data (e.g. the average size of a group). The formulas for these properties are presented in Appendix B.

2.3. Known Star Model

The initial motivation for Star Models comes from the structure of e-mail datasets. In these datasets, each group is the result of a population member sending a message to a subset

of the population. Since e-mail logs generally indicate the sender, it is possible to identify the unique star graph that produced an observed group. Such a model is called the *Known Star Model* (KSM). Under the KSM, $S^{(t)}$ is observed and the number of parameters to be estimated is:

$$d = (n - 1)^2. \quad (2.1)$$

Additionally, note that for the KSM, A is not necessarily symmetric.

The following section formally proves that the measure used by Michalski et al. (Equation 1.3) is the maximum likelihood estimator for A_{ij} when the center of the star forming the observed group is known.

2.3.1 Maximum Likelihood Estimators

By definition, there is only one graph that could have produced any observed group. Therefore, the probability of $G^{(t)}$ and $S^{(t)}$, given A , is provided by Equation 2.2

$$\mathbb{P}(G^{(t)}, S^{(t)}|A) = \prod_{i=1}^n \prod_{j=1}^n [A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{(1-G_j^{(t)})}]^{S_i^{(t)}}. \quad (2.2)$$

The likelihood function for the full observation is given by Equation 2.3

$$\mathbb{P}(G, S|A) = \prod_{t=1}^T \prod_{i=1}^n \prod_{j=1}^n [A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{(1-G_j^{(t)})}]^{S_i^{(t)}}. \quad (2.3)$$

Taking the log of the likelihood function produces

$$\mathcal{L} = \log \mathbb{P}(G, S|A) = \sum_{t=1}^T \sum_{i=1}^n \sum_{j=1}^n S_i^{(t)} [G_j^{(t)} \log A_{ij} + (1 - G_j^{(t)}) \log(1 - A_{ij})]. \quad (2.4)$$

The derivative with respect to A_{xy} is

$$\frac{\partial \mathcal{L}}{\partial A_{xy}} = \sum_{t=1}^T S_x^{(t)} \left[\frac{G_y^{(t)}}{A_{xy}} - \frac{1 - G_y^{(t)}}{1 - A_{xy}} \right]. \quad (2.5)$$

Setting this equation equal to zero and solving for A_{xy} yields

$$\frac{1}{A_{xy}} \sum_{\{t:G_y^{(t)}=1\}} S_x^{(t)} = \frac{1}{1 - A_{xy}} \sum_{\{t:G_y^{(t)}=0\}} S_x^{(t)} \quad (2.6)$$

and

$$\hat{A}_{xy} = \frac{\sum_t G_y^{(t)} S_x^{(t)}}{\sum_t S_x^{(t)}}. \quad (2.7)$$

This is exactly the form of the estimator that was used by Michalski et al. (2014).

It is also worth noting that it is not necessary to estimate ρ because $S^{(t)}$ is directly observable for this model. However, estimates of ρ_i are needed in the more general model introduced in the following section.

2.4. The Symmetric Star Model

For the general Star Model, the probability that a group is observed is

$$\mathbb{P}(G^{(t)}|A, \rho) = \sum_i \rho_i G_i^{(t)} \prod_j A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{(1-G_j^{(t)})} \quad (2.8)$$

and the log likelihood function is:

$$\mathcal{L}(G|A, \rho) = \sum_t \log \left[\sum_{i=1}^n \rho_i G_i^{(t)} \prod_j A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{1-G_j^{(t)}} \right]. \quad (2.9)$$

Since most existing techniques for inferring network structure produce a symmetric adjacency matrix, a reasonable constraint to consider is symmetry, i.e., $A_{ij} = A_{ji}$ for all i and j . The following theorem shows that symmetry of A is a sufficient condition for identifiability of a Star Model.

Theorem 1. Let A and A^* be symmetric adjacency matrices. If $\mathbb{P}(G = g|A, \rho) = \mathbb{P}(G = g|A^*, \rho^*)$ for all g , then $\{A, \rho\} = \{A^*, \rho^*\}$.

Proof. Let g^k and g^l denote the singleton groups which consist only of nodes v_k and v_l , respectively. Further, let g^{kl} denote the group representing the pair of v_k and v_l .

From (2.8) the probability of the singletons is given by the following equations:

$$\mathbb{P}(G = g^k|A, \rho) = \rho_k(1 - A_{kl}) \prod_{j \neq \{k, l\}} (1 - A_{kj}) \quad (2.10)$$

$$\mathbb{P}(G = g^l|A, \rho) = \rho_l(1 - A_{kl}) \prod_{j \neq \{k, l\}} (1 - A_{lj}) \quad (2.11)$$

In (2.11) we take advantage of the symmetry of A to replace A_{lk} with A_{kl} .

Now, we consider the probability of g^{kl} .

$$\begin{aligned} \mathbb{P}(g^{kl}|A, \rho) &= \rho_k A_{kl} \prod_{j \neq \{k, l\}} (1 - A_{kj}) + \rho_l A_{kl} \prod_{j \neq \{k, l\}} (1 - A_{lj}) \\ &= A_{kl} \left[\rho_k \prod_{j \neq \{k, l\}} (1 - A_{kj}) + \rho_l \prod_{j \neq \{k, l\}} (1 - A_{lj}) \right] \\ &= A_{kl} \left[\frac{\mathbb{P}(G = g^k|A, \rho)}{(1 - A_{kl})} + \frac{\mathbb{P}(G = g^l|A, \rho)}{(1 - A_{kl})} \right] \\ &= \frac{A_{kl}}{(1 - A_{kl})} \left[\mathbb{P}(G = g^k|A, \rho) + \mathbb{P}(G = g^l|A, \rho) \right], \end{aligned} \quad (2.12)$$

which implies that:

$$A_{kl} = \frac{\mathbb{P}(G = g^{kl}|A, \rho)}{\mathbb{P}(G = g^k|A, \rho) + \mathbb{P}(G = g^l|A, \rho) + \mathbb{P}(G = g^{kl}|A, \rho)}. \quad (2.13)$$

Therefore, $A_{kl} = A_{kl}^*$ for all k and l .

To complete the proof, consider an arbitrary node v_k which appears as a singleton represented by g^k :

$$\mathbb{P}(G = g^k|A, \rho) = \rho_k \prod_{j \neq k} (1 - A_{kj}). \quad (2.14)$$

If $A_{kl} = A_{kl}^*$ for all k and l and $\mathbb{P}(G = g|A, \rho) = \mathbb{P}(G = g|A^*, \rho^*)$ for all g , then:

$$\rho_k \prod_{j \neq k} (1 - A_{kj}) = \rho_k^* \prod_{j \neq k} (1 - A_{kj}^*) \quad (2.15)$$

and it is easy to see that $\rho_k = \rho_k^*$ for all k . □

We will refer to the Star Model with the symmetry condition as the *Symmetric Star Model* (SSM) for the rest of this dissertation. Note that for the SSM, the number of parameters being estimated is:

$$d = \binom{n}{2} + (n - 1). \quad (2.16)$$

Remarks: Before proceeding we would like to make two short remarks about (2.13).

Firstly, (2.13) suggests a method of moments estimator for A_{kl} based on the frequencies of doubletons and singletons. However, this estimator requires that the probability of doubletons and singletons be estimated accurately, so this technique would be very inefficient in practice because small groups appear very infrequently in many datasets. Therefore, we will continue to consider MLE which presumably uses all available information.

Secondly, note that the form of (2.13) is similar to the form of the half weight index shown in (1.2). Roughly speaking, the SSM only uses the probabilities of singletons and doubletons, while the half weight index uses frequencies of non-co-occurrence and co-occurrence.

2.4.1 Maximum Likelihood Estimate of the Symmetric Star Model

The maximum likelihood estimator of SSM does not have a closed-form solution. In this section, we derive equations for the conditions that the MLE $\{\hat{A}, \hat{\rho}\}$ must satisfy. Then we will show that solving these equations iteratively is equivalent to an EM algorithm. The details of the EM algorithm will be given in the next section.

Solving the MLE of SSM is an optimization problem with the equality constraints $\sum_i \rho_i = 1$, and $A_{ij} = A_{ji}$ for all i and j . We denote the log likelihood function as $\mathcal{L}(G|A, \rho)$. This gives us the following Lagrange function:

$$\Lambda(G|A, \rho) = \mathcal{L}(G|A, \rho) - \lambda_o[(\sum_i \rho_i) - 1] - \sum_{i < j} \lambda_{ij}(A_{ij} - A_{ji}). \quad (2.17)$$

$$\frac{\partial \Lambda(G|A, \rho)}{\partial A_{xy}} = \frac{\partial \mathcal{L}(G|A, \rho)}{\partial A_{xy}} - \lambda_{xy} = 0 \text{ if } x < y, \quad (2.18)$$

$$\frac{\partial \Lambda(G|A, \rho)}{\partial A_{yx}} = \frac{\partial \mathcal{L}(G|A, \rho)}{\partial A_{yx}} + \lambda_{xy} = 0 \text{ if } x > y. \quad (2.19)$$

Therefore,

$$\frac{\partial \mathcal{L}(G|A, \rho)}{\partial A_{xy}} = -\frac{\partial \mathcal{L}(G|A, \rho)}{\partial A_{yx}}. \quad (2.20)$$

We now focus on deriving the derivative of the log likelihood function of the general Star Model given in Equation (2.9):

$$\frac{\partial}{\partial A_{xy}} \mathcal{L}(G|A, \rho) = \sum_t \frac{\rho_x G_x^{(t)} \left[\prod_{j \neq y} A_{xj}^{G_j^{(t)}} (1 - A_{xj})^{1-G_j^{(t)}} \right] \frac{\partial}{\partial A_{xy}} \left(A_{xy}^{G_y^{(t)}} (1 - A_{xy})^{1-G_y^{(t)}} \right)}{\sum_{i=1}^n \rho_i G_i^{(t)} \prod_j A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{1-G_j^{(t)}}}. \quad (2.21)$$

Note that the derivative on the right hand side of (2.21) is equal to 1 if node v_y is in observation $G^{(t)}$ and -1 if v_y is not in the observation. We represent this by the function

$$\gamma(G_y^{(t)}) = \begin{cases} 1 & \text{if } G_y^{(t)} = 1, \\ -1 & \text{if } G_y^{(t)} = 0. \end{cases}$$

Therefore,

$$\frac{\partial}{\partial A_{xy}} \mathcal{L}(G|A, \rho) = \sum_t \frac{\rho_x G_x^{(t)} \left[\prod_{j \neq y} A_{xj}^{G_j^{(t)}} (1 - A_{xj})^{1-G_j^{(t)}} \right] \gamma(G_y^{(t)})}{\sum_{i=1}^n \rho_i G_i^{(t)} \prod_j A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{1-G_j^{(t)}}}. \quad (2.22)$$

The denominator of (2.22) is simply the probability of $G^{(t)}$ (see (2.8)). In addition, the term in the numerator can be made equal to $\mathbb{P}(G^{(t)}, S_x = 1)$ by multiplying $A_{xy}^{G_y^{(t)}} (1 - A_{xy})^{(1-G_y^{(t)})}$. This gives:

$$\frac{\partial}{\partial A_{xy}} \mathcal{L}(G|A, \rho) = \sum_t \frac{\gamma(G_y^{(t)}) \mathbb{P}(G^{(t)}, S_x^{(t)} = 1)}{A_{xy}^{G_y^{(t)}} (1 - A_{xy})^{(1-G_y^{(t)})} \mathbb{P}(G^{(t)}|A)}. \quad (2.23)$$

This equation can be further simplified by noticing that $\frac{\mathbb{P}(G^{(t)}, S_x^{(t)}=1)}{\mathbb{P}(G^{(t)}|A)}$ is equivalent to $\mathbb{P}(S_x^{(t)} = 1|G^{(t)})$:

$$\frac{\partial}{\partial A_{xy}} \mathcal{L}(G|A, \rho) = \sum_t \frac{\gamma(G_y^{(t)}) \mathbb{P}(S_x^{(t)} = 1|G^{(t)})}{A_{xy}^{G_y^{(t)}} (1 - A_{xy})^{(1-G_y^{(t)})}}. \quad (2.24)$$

Plugging (2.24) into (2.20), we get:

$$\sum_t \frac{\gamma(G_y^{(t)}) \mathbb{P}(S_x^{(t)} = 1|G^{(t)})}{A_{xy}^{G_y^{(t)}} (1 - A_{xy})^{(1-G_y^{(t)})}} = - \sum_t \frac{\gamma(G_x^{(t)}) \mathbb{P}(S_y^{(t)} = 1|G^{(t)})}{A_{yx}^{G_x^{(t)}} (1 - A_{yx})^{(1-G_x^{(t)})}}. \quad (2.25)$$

By applying symmetry and breaking the summations, this becomes:

$$\begin{aligned} \sum_{t:G_y^{(t)}=1} \frac{\mathbb{P}(S_x^{(t)} = 1|G^{(t)})}{A_{xy}} - \sum_{t:G_y^{(t)}=0} \frac{\mathbb{P}(S_x^{(t)} = 1|G^{(t)})}{1 - A_{xy}} = \\ - \sum_{t:G_x^{(t)}=1} \frac{\mathbb{P}(S_y^{(t)} = 1|G^{(t)})}{A_{xy}} + \sum_{t:G_x^{(t)}=0} \frac{\mathbb{P}(S_y^{(t)} = 1|G^{(t)})}{1 - A_{xy}}. \end{aligned} \quad (2.26)$$

With some simple algebra, it is easy to see that:

$$\hat{A}_{xy} = \frac{\sum_t G_y^{(t)} \mathbb{P}(S_x = 1|G^{(t)}) + G_x^{(t)} \mathbb{P}(S_y = 1|G^{(t)})}{\sum_t [\mathbb{P}(S_x = 1|G^{(t)}) + \mathbb{P}(S_y = 1|G^{(t)})]}. \quad (2.27)$$

It is worth restating that (2.27) is not a closed form solution for \hat{A}_{xy} . This is because the right hand side of the equation contains \hat{A}_{xy} .

We now proceed to derive the condition for $\hat{\rho}$. By taking the derivative of (2.17) with respect to ρ_x , we get the following:

$$\frac{\partial}{\partial \rho_x} \Lambda(G|A, \rho) = \sum_t \frac{G_x^{(t)} \prod_j A_{xj}^{G_j^{(t)}} (1 - A_{xj})^{1-G_j^{(t)}}}{\mathbb{P}(G^{(t)}|A)} - \lambda_o \quad (2.28)$$

$$= \sum_t \frac{\mathbb{P}(G^{(t)}|S_x^{(t)} = 1)}{\rho_x \mathbb{P}(G^{(t)}|A)} - \lambda_o \quad (2.29)$$

$$= \frac{1}{\rho_x} \sum_t \mathbb{P}(S_x^{(t)} = 1|G^{(t)}) - \lambda_o. \quad (2.30)$$

Solving this equation for zero, we obtain:

$$\sum_t \mathbb{P}(S_x^{(t)} = 1|G^{(t)}) = \rho_x \lambda_o. \quad (2.31)$$

Summing over all nodes, we get:

$$\sum_i \sum_t \mathbb{P}(S_i^{(t)} = 1|G^{(t)}) = \sum_i \rho_i \lambda_o \quad (2.32)$$

$$\sum_t \sum_i \mathbb{P}(S_i^{(t)} = 1|G^{(t)}) = \lambda_o \sum_i \rho_i \quad (2.33)$$

$$T = \lambda_o. \quad (2.34)$$

Which gives us,

$$\hat{\rho}_x = \frac{\sum_{t=1}^T \mathbb{P}(S_x^{(t)} = 1|G^{(t)})}{T}. \quad (2.35)$$

2.4.2 EM Algorithm

The last section ended with estimating equations where the probability $\mathbb{P}(S_x^{(t)} = 1|G^{(t)})$ on the right side of (2.27) and (2.35) depends on $\{\hat{A}, \hat{\rho}\}$. This implies a fairly intuitive algorithm iteratively updating $\{\hat{A}, \hat{\rho}\}$ and $\mathbb{P}(S_x^{(t)} = 1|G^{(t)})$, which can be fitted into the general framework of EM algorithm.

The central concept of the EM algorithm is to formulate a complete data model then solve the model as if some data is observed and other data is missing. In this case, the Known Star Model serves as the complete data model; G is the observed data, and S is the missing data. Each iteration of the EM algorithm consists of an expectation step followed by a maximization step. In the most general form of the EM algorithm, the MLE of the observed data cannot be directly calculated because of the missing data. To overcome this, the EM algorithm maximizes the expected value of the log likelihood function of the complete data model. This log likelihood function is given in (2.4).

Based on this, the first step of the EM algorithm involves calculating the following function where $A^{(k)}$ is the current iteration of the adjacency matrix.

E-Step

$$Q(A|A^{(k)}) = E \left[\sum_{t=1}^T \sum_{i=1}^n \sum_{j=1}^n S_i^{(t)} [G_j^{(t)} \log A_{ij}^{(k)} + (1 - G_j^{(t)}) \log(1 - A_{ij}^{(k)})] \middle| G \right] \quad (2.36)$$

Since (2.36) is linear in the unobserved data, $S_i^{(t)}$, the E-Step (on the $(k+1)^{th}$ iteration) simply requires calculating the current conditional expectation of $S_i^{(t)}$ given the observed data, $G^{(t)}$. See (McLachlan and Krishnan, 2008) for a detailed explanation.

$$\mathbb{P}(S_x^{(t)} = 1 | G^{(t)}, A) = E[S_x^{(t)} | G^{(t)}, A] = \frac{\rho_x G_x^{(t)} \prod_j A_{xj}^{G_j^{(t)}} (1 - A_{xj})^{1-G_j^{(t)}}}{\sum_{i=1}^n \rho_i G_i^{(t)} \prod_j A_{ij}^{G_j^{(t)}} (1 - A_{ij})^{1-G_j^{(t)}}} \quad (2.37)$$

M-Step

In the general form of the EM algorithm, the M-Step consists of selecting the value of A which maximizes (2.36). However, since we are able to derive the MLE condition for the SSM, the M-Step on the $(k + 1)^{th}$ iteration simply requires replacing $\mathbb{P}(S_x^{(t)} = 1 | G^{(t)})$ in (2.27) and (2.35) with $E[S_x^{(t)} | G^{(t)}]$.

Theoretical Properties

The Expectation Maximization algorithm has several standard properties which Star Models inherit. These include numerical stability with each EM iteration increasing the likelihood, easy implementation, and a low cost per iteration which can offset the larger number of iterations needed for the EM algorithm compared to competing procedures. Since Star Models belong to classical mixture models, the MLEs also achieve consistency and asymptotic normality (McLachlan and Krishnan, 2008).

2.4.3 Symmetric Star Model EM Algorithm

Algorithm 1 illustrates the details of the algorithm for the SSM.

There are two techniques included in the algorithm to avoid using a bad starting point. First, we use ten different starting points. Second, we limit the number of iterations taken through the algorithm. This is based on the observation that when this algorithm has a bad starting point, it will take a very long time to converge; and the point that it converges to is not a maximum.

Also, notice that when $\rho_i = \rho_j = 0$, A_{ij} is undefined under (2.27). In this case, the value of A_{ij} does not affect the likelihood function for the observed data; therefore, we define $A_{ij} = 0$.

```

Data: G
Result:  $\hat{A}, \hat{\rho}$ 

Initialize:
 $\mathcal{L}(G|\hat{A}) = -\infty$ 

for rep=1 to 10 do
  Initialize:
   $\hat{A}_{ij}^{(0)} = \text{unif}(0, 1) \quad \forall \{i, j\}$ 
   $X_i = \text{unif}(0, 1) \quad \forall i$ 
   $\hat{\rho}_i^{(0)} = \frac{X_i}{\sum_k X_k}$ 
   $\Delta\mathcal{L}(G|A^{(0)}) = 10^4$ 

  counter=1
  while  $|\frac{\Delta\mathcal{L}(G|A^{(m+1)})}{\mathcal{L}(G|A^{(m)})}| > 10^{-4}$  and counter < 100 do
    E-Step
    Update  $\mathbb{P}(S_k^{(t)} = 1|G^{(t)})$  by Equation 2.37
    M-Step
    Update  $A^{(m+1)}$  by Equation 2.27
    Update  $\rho^{(m+1)}$  by Equation 2.35
     $\Delta\mathcal{L}(G|A^{(m+1)}) = \mathcal{L}(G|A^{(m+1)}) - \mathcal{L}(G|A^{(m)})$ 

    counter=counter+1
  end

  if  $\mathcal{L}(G|A^{(m+1)}) > \mathcal{L}(G|\hat{A})$  then
    if  $\hat{A}_{ij} \leq 10^{-4}$  then
       $\hat{A}_{ij} = 0$ 
    else
       $\hat{A}_{ij} = A_{ij}^{(m+1)}$ 
    end
  end
end

```

Algorithm 1: Symmetric Star Model EM Algorithm

To illustrate the difference in the performance of the traditional techniques and the Star Models, consider the toy example presented in Figure 2.2a where the strength of the relationships is represented by the width of the links. In this example, there is a pair of nodes, v_1 and v_2 , that never voluntarily approach each other. Despite the aversion between

these two nodes, they share a number of relationships with common nodes. This situation may be thought of as two coworkers who do not particularly like one another, but who are often required to cooperate because of their relationships to intermediaries. In Figure 2.2b, we can see that the co-occurrence matrix mistakenly assigns a relatively strong relationship to nodes v_1 and v_2 . In Figure 2.2c, the half-weight index arrives at a very similar conclusion as the co-occurrence matrix. In both Figures 2.2b and 2.2c, the non-existent relationship between nodes v_1 and v_2 is actually estimated to be stronger than all other relationships. By contrast, the Star Model results in Figure 2.2d clearly capture the social structure of the population.

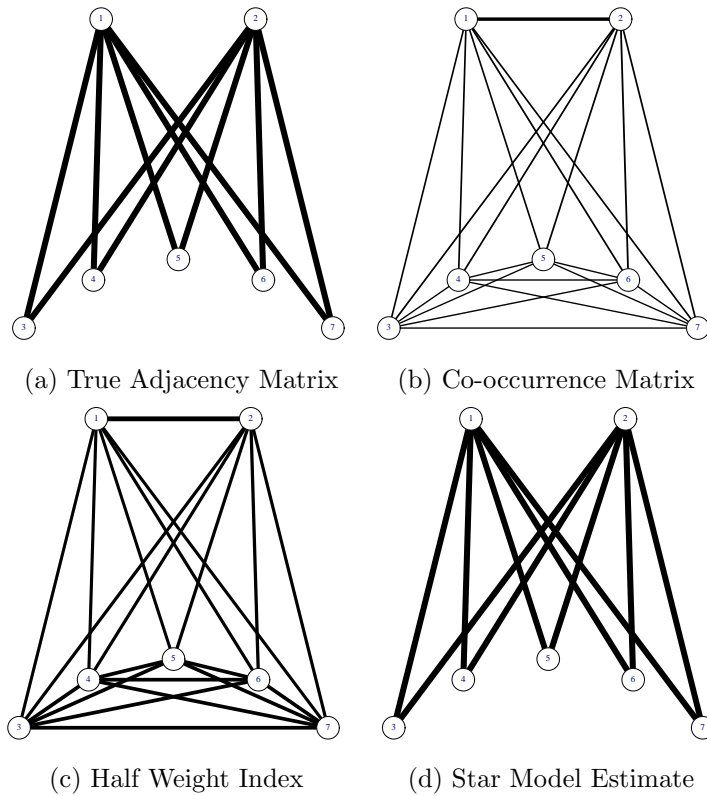


Figure 2.2: Comparison of Estimation Techniques

2.5. Simulation Results

2.5.1 Generating the True Model Parameters

When performing data analysis, the “true” parameters are unobservable. However, we can evaluate the performance of different methods by comparing them with simulated parameters.

To generate ρ , we select n i.i.d. random numbers uniformly, X_i , and divide each random number by the sum of all X_i 's. That is, $\rho_i = \frac{X_i}{\sum_j X_j}$.

We use a two step process to generate A . First, we create an unweighted, undirected Erdos-Renyi random graph on n nodes where an edge is present with probability p . Then each present edge is assigned a relationship strength with a beta distribution, i.e., for $i < j$,

$$A_{ij} = \begin{cases} \text{Beta}(\alpha, \beta) & \text{w.p. } p \\ 0 & \text{otherwise} \end{cases}$$

we let $A_{ji} = A_{ij}$ because we assume A is symmetric.

In the following examples, a common setup is used so that all examples are similar. The probability that a relationship exists is 0.30. When relationships do not exist, the probability of two nodes interacting is 0.001. And when a relationship does exist, the probability of two nodes interacting is given by the beta distribution $\text{Beta}(\alpha = 10, \beta = 2)$. This distribution has the form shown in Figure 2.3 and a mean of $\frac{10}{10+2} = 0.8333$. This produces a latent network structure where relationships are clearly distinguished from random encounters.

Although the following examples require the estimation of $\hat{\rho}$, the results focus on \hat{A} to explore convergence and compare techniques.

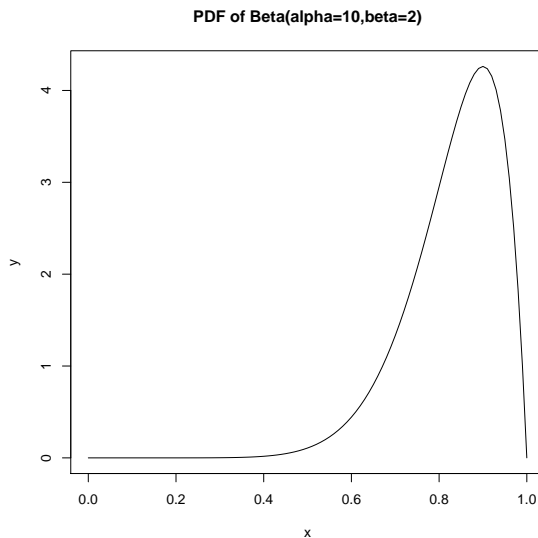


Figure 2.3: Distribution of Beta Function

2.5.2 Simple Example of SSM

The initial example presents a small population of only 5 nodes to limit the estimation to 14 parameters. The purpose of this example is to show that convergence does, in fact, occur. The true adjacency matrix for this population is shown on left-hand side of Table 2.2. To ensure convergence, 10,000 observations are generated.

Comparing the estimated adjacency matrix on the right hand side of Table 2.2 with the true adjacency matrix on the left, we can make two important observations. First, the estimated values of \hat{A}_{ij} are generally close to their true values. Secondly, there is no instance of a link with no relationship being mistaken for a link with a relationship.

2.5.3 Convergence of SSM

In this section, we briefly explore how the estimator behaves as the number of observations increases for different network sizes.

To measure the difference between A and \hat{A} , we define the *mean absolute error* (MAE)

Table 2.2: General results

i	True Adjacency Matrix					Estimated Adjacency Matrix				
	j					j				
	1	2	3	4	5	1	2	3	4	5
1	1.0000	0.0010	0.0010	0.0010	0.0010	1.0000	0.0010	0.0004	0.0001	0.0015
2	0.0010	1.0000	0.0010	0.7452	0.8324	0.0010	1.0000	0.0144	0.7465	0.8302
3	0.0010	0.0010	1.0000	0.0010	0.5885	0.0004	0.0144	1.0000	0.0043	0.5536
4	0.0010	0.7452	0.0010	1.0000	0.8594	0.0001	0.7465	0.0043	1.0000	0.8573
5	0.0010	0.8324	0.5885	0.8594	1.0000	0.0015	0.8302	0.5536	0.8573	1.0000

over the matrix as:

$$MAE = \frac{1}{\binom{n}{2}} \sum_{i < j} |\hat{A}_{ij} - A_{ij}|. \tag{2.38}$$

There are other metrics from computer science literature to evaluate the effectiveness of the estimated A matrix. Some of these metrics focus on false positives, false negatives, or functions of these values (Hastie et al., 2009). Typically these methods, in a network setting, work under some assumption of sparsity. Since we make no assumptions regarding sparsity in this thesis, we have not compared (2.38) with these other metrics. It is worthwhile to compare our metric with these alternative metrics under a sparsity assumption and we leave this for future work.

In Table 2.3, we consider four different scenarios with $n = \{10, 20, 30, 50\}$. The latent network structure for each scenario was generated with $p = 0.3$, $\alpha = 10$, and $\beta = 2$. For each scenario, we run 100 replicates and report the mean and standard deviation of the MAE over these replicates.

Not surprisingly, as the number of observations increases, the average MAE tends to decline. For the first two scenarios of $n = 10$ and $n = 20$, the standard deviation of the MAE declines monotonically as the number of observations increases. For the second two

scenarios of $n = 30$ and $n = 50$, although the standard deviation of the MAE ultimately decreases as the number of observations increases, for low values of T , the standard deviation seems to rise before it decreases.

Table 2.3: Average and Standard Deviation of Mean Absolute Error as Observations Increase

	$n = 10$		$n = 20$		$n = 30$		$n = 50$	
Obs	Avg MAE	StDev MAE	Avg MAE	StDev MAE	Avg MAE	StDev MAE	Avg MAE	StDev MAE
200	0.021	0.004	0.031	0.023	0.184	0.047	0.256	0.024
500	0.013	0.002	0.012	0.002	0.110	0.053	0.237	0.027
1000	0.009	0.002	0.009	0.001	0.067	0.056	0.207	0.033
2000	0.006	0.001	0.006	0.001	0.020	0.024	0.183	0.043
4000	0.005	0.001	0.004	0.001	0.010	0.017	0.145	0.053
10000	0.003	0.001	0.003	0.000	0.005	0.011	0.081	0.061
20000	0.002	0.000	0.002	0.000	0.003	0.002	0.060	0.058
50000	0.001	0.000	0.001	0.000	0.002	0.002	0.020	0.033

2.5.4 Visualization

The mean absolute error of the matrix is only a measure of overall estimator performance and cannot give details of each adjacency matrix element. To explore how each element of the estimate changes when the number of observations increases, we introduce a visualization of the relationships between members in the population.

In this visualization, the adjacency matrix is represented as an $n \times n$ grid where the $i^{th} \times j^{th}$ cell represents the relationship A_{ij} . The strength of a relationship is represented by the cell's color. That is, nodes with weak relationships have light cells while nodes with strong relationship have dark cells. Cells representing relationships of intermediate strength are shaded along the gray scale.

Figure 2.4 presents estimates of an adjacency matrix with $n = 20$. Figure 2.4a shows the

true matrix and Figures 2.4b-2.4d give estimates as the number of observations increases from $T = 100$ to $T = 10,000$. As suggested by Table 2.3, when there are relatively few observations for a population, the estimated structure contains important errors. In this case, there are 209 parameters which need to be estimated. With only $T = 100$ (Figure 2.4b), we can see that some of the true adjacency matrix's basic structure is identified, but there is considerable error in some relationships, consider $A_{1,20}$ as an example. In the true adjacency matrix, this relationship does not exist, but in the first estimate the relationship is estimated to be strong. However, as the number of observations increases to $T = 1,000$ (Figure 2.4c) the set of non-zero relationships in the true adjacency matrix is captured, and the only errors exist in degrees of strength. As an example of this refinement, consider $A_{2,20}$ which is under estimated in Figure 2.4c but estimated more accurately in Figure 2.4d.

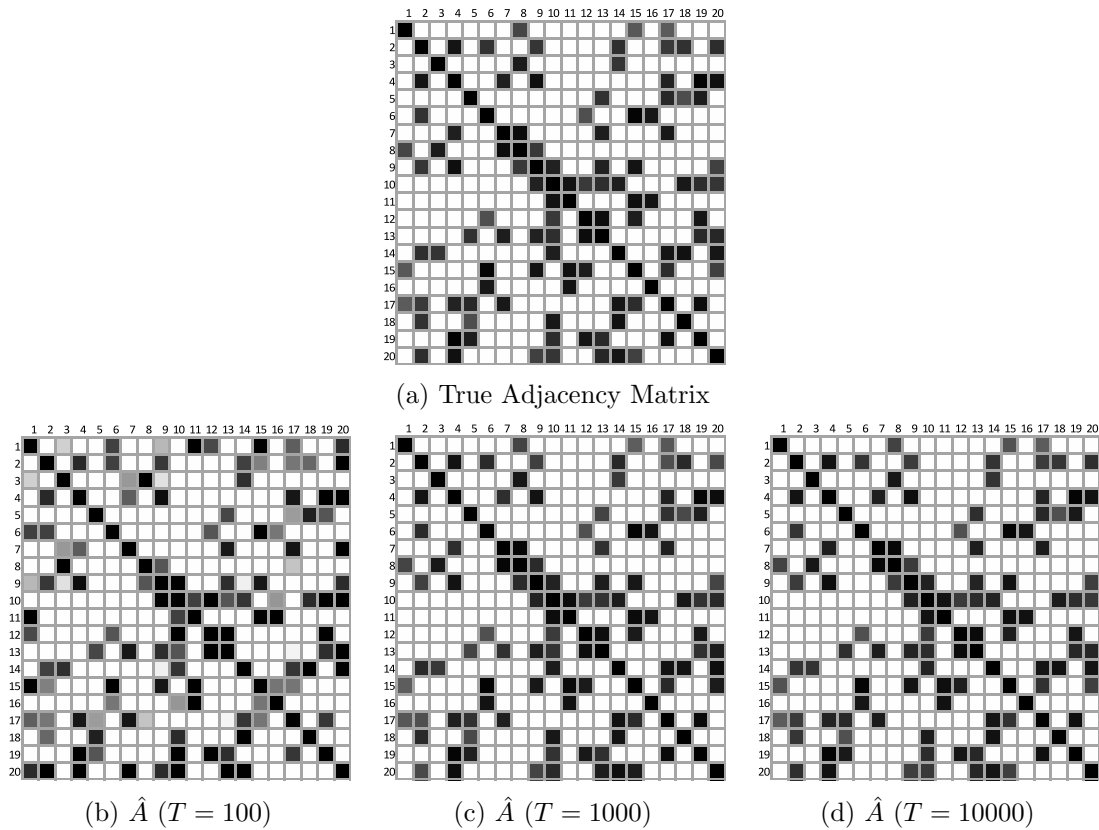


Figure 2.4: Symmetric Star Model Estimates Improve as T Increases

2.6. Data Analysis

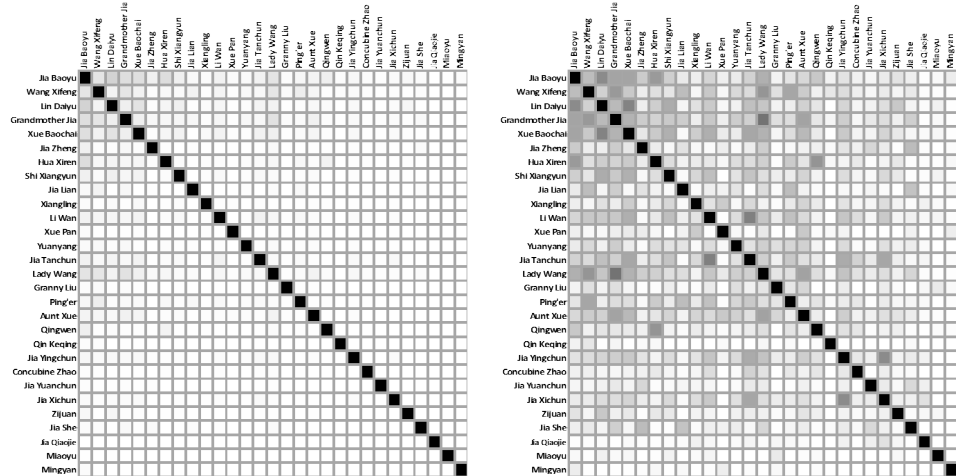
2.6.1 Dream of the Red Chamber

Overview

Dream of the Red Chamber was first published in print form in 1791 and is one of China's four great classical novels. The novel provides a detailed record of the wealthy and aristocratic Jia clan who live in two large, adjacent family compounds in the capital. The novel charts some thirty main characters and over four hundred minor ones as the Jias' fall from the height of their prestige.

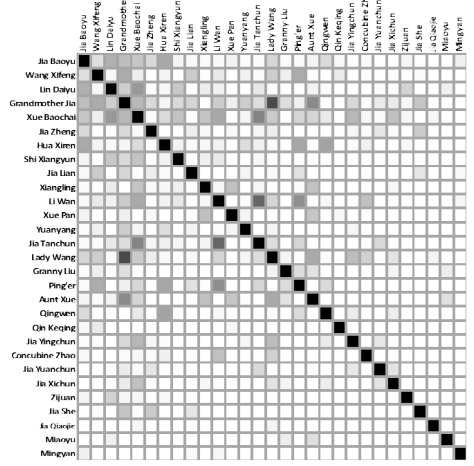
The main character of the novel is the adolescent male heir of the family, Jia Baoyu. He has a special bond with his sickly cousin Lin Daiyu, who shares his love of music and poetry. Baoyu, however, is ultimately tricked into marrying another cousin, Xue Baochai, whose grace and intelligence exemplifies an ideal woman, but with whom he lacks an emotional connection. The romantic rivalry and friendship among the three characters against the backdrop of the family's declining fortunes forms the main story in the novel.

Figure 2.5 presents the results of the three techniques for estimating social structure discussed throughout this dissertation. The characters have been ranked by the estimated value of ρ (see Table 2.4). In these estimates, the same performance is observed that has been seen in simulation studies. The co-occurrence matrix estimates all relationships as being very weak and it is difficult to differentiate the presence of a relationship from an absence of relationship. The half-weight index presents a much denser set of relationships, but there is evidence of relationships which have been imputed transitively.



(a) Co-Occurrence Matrix

(b) Half Weight Index



(c) SSM Adjacency Matrix

Figure 2.5: Estimates for *Dream of the Red Chamber*

For a specific example, consider Jia Yingchun and Jia Xichun in Figure 2.6. In the co-occurrence matrix, there seems to be no significant relationship between these characters and the three main characters. In the half-weight index, these characters share relationships with all the three main characters; but in the SSM adjacency matrix, these characters only share a relationship with the third main character. In general, SSM returns a much sparser network with clearer structure.

The EM algorithm of SSM provides very stable solutions. By selecting multiple starting points, we find that the adjacency matrix (Figure 2.5c) is repeatedly returned as the most likely parameter of the observed data.

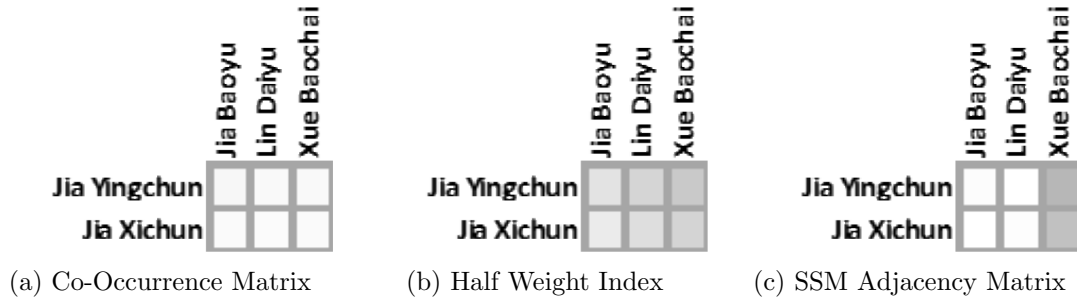


Figure 2.6: Estimates Brothers in *Dream of the Red Chamber*

In Table 2.4, we present the ρ values of the characters sorted from highest to lowest along with the degree centrality and eigenvector centrality of \hat{A} .

Characters with a high value of ρ are relatively more likely to initiate a group. The first six names on the list capture characters who are central to the story. Jia Baoyu is the main character. Wang Xifeng is the female family member who runs the household and wields political and economic power over the family. Lin Daiyu is the love interest of Jia Baoyu. Grandmother Jia is the highest living authority in the house and the oldest and most respected member of the entire clan. Xue Baochai is the “ideal” Chinese maiden and the predestined wife of Jia Baoyu. Finally, Jia Zheng is the father of the main character.

Degree centrality is a measure of the probability that an individual node will be selected to be a member of a group. Nodes with high degree centrality are more likely to be chosen to be members of a group than nodes with low degree centrality. Notice that degree centrality does not correspond directly to values of ρ . In particular, there are several characters with high degree centrality in Table 2.4 who have a very low value of ρ (e.g. Ping’er).

Eigenvector centrality is a measure of the influence of a node within a network. It

Table 2.4: Values of ρ for characters in *Dream of the Red Chamber*

Character	ρ	Degree Centrality	Eigenvector Centrality
Jia Baoyu	0.3061	3.201	0.553
Wang Xifeng	0.1573	2.723	0.479
Lin Daiyu	0.0997	2.789	0.502
Grandmother Jia	0.0838	5.135	1.000
Xue Baochai	0.064	4.789	0.937
Jia Zheng	0.0409	1.948	0.208
Hua Xiren	0.0375	2.799	0.403
Shi Xiangyun	0.0329	1.905	0.309
Jia Lian	0.029	1.820	0.202
Xiangling	0.0175	2.075	0.259
Li Wan	0.0173	3.441	0.718
Xue Pan	0.0151	1.690	0.139
Yuanyang	0.0148	1.645	0.196
Jia Tanchun	0.0144	2.987	0.657
Lady Wang	0.0135	3.764	0.815
Granny Liu	0.0118	1.651	0.178
Ping'er	0.0072	2.951	0.501
Aunt Xue	0.0068	2.793	0.517
Qingwen	0.0068	1.959	0.227
Qin Keqing	0.0044	1.148	0.030
Jia Yingchun	0.0042	1.987	0.336
Concubine Zhao	0.0039	1.590	0.203
Jia Yuanchun	0.0034	1.580	0.139
Jia Xichun	0.0031	1.744	0.227
Zijuan	0.0024	1.320	0.093
Jia She	0.0023	1.567	0.150
Jia Qiaojie	0.0000	1.078	0.033
Miaoyu	0.0000	1.265	0.053
Mingyan	0.0000	1.119	0.030

assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of a given node than equal connections to low-scoring nodes. As with degree centrality, this measure does not follow the same pattern as ρ . While some well connected members of the population are ranked high in this measure (e.g. Grandmother Jia), nodes with high degree centrality also seem to have

a high eigenvector centrality.

The standard deviation of the parameters of the Symmetric Star Model was estimated using the bootstrap technique. Using the frequency table, F , discussed in Section 1.4.1, we sampled 5000 datasets by taking T samples from F where each group was selected with probability equal to its frequency in F . In general, the standard deviation was low. This was particularly true for $\hat{\rho}$ where the maximum standard deviation was 0.0173. Table 2.5 presents the standard deviation of \hat{A} at different percentiles.

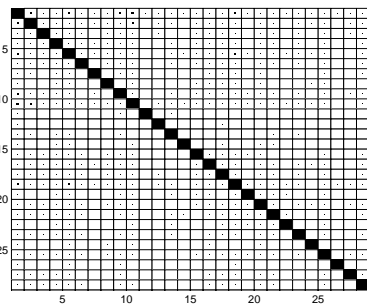
Table 2.5: Percentiles of Standard Deviation in \hat{A} estimated by SSM for *Dream of the Red Chamber*

Percentile	Max	95 %	75 %	Med	25 %	5 %	Min
StDev	0.2696	0.1025	0.0374	0.0100	0.0000	0.0000	0.0000

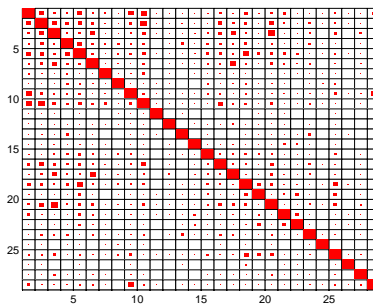
Alternative Visualization

As an alternative to the gray scale visualization presented above, we offer a technique that represents the relationship strength by the cell area plotted. In Figure 2.7, the width and height of each plotted cell is equal to the relationship strength. Therefore, if a pair of nodes shares a strong relationship, the cell will be almost filled and if the relationship is weak the cell will be almost empty. Throughout this section, we represent co-occurrence with the color black, while the half weight index is red and the SSM is green. Characters are presented in the same order as in Table 2.4.

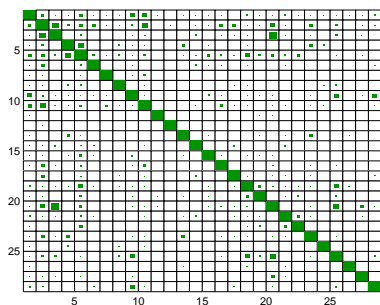
In Figure 2.8, we take advantage of this plotting technique to contrast different estimates by overlaying them on each other. In each cell, the smaller value is plotted in the center of the larger value. This means that a red square with a black square inside of it represents a half weight index estimate that exceeds the co-occurrence estimate.



(a) Co-Occurrence Matrix



(b) Half Weight Index



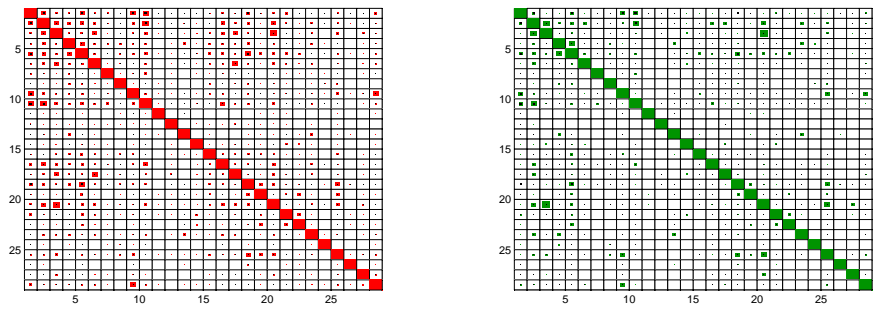
(c) SSM Adjacency Matrix

Figure 2.7: Estimates for *Dream of the Red Chamber* Represented by Plot Area

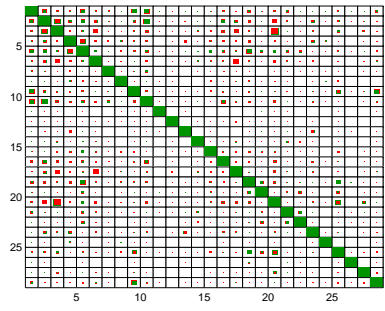
Main Characters

One of the main themes of the *Dream of the Red Chamber* is the love story surrounding the protagonist Jia Baoyu (1st character in Figure 2.5) and two potential fiances. These are the sickly Lin Daiyu (2nd character) and the “ideal” Xue Baochai (3rd character). Although Jia Baoyu shares a special bond with Lin Daiyu and has no significant emotional connection to Xue Baochai, he is ultimately tricked into marrying Xue Baochai.

In Table 2.6, we present the relationships between these two girls and the other characters as estimated by the co-occurrence matrix, half weight index, and SSM. As mentioned earlier, the SSM shows a sparse structure. When the elements of \hat{A} equal zero, they coincide with relationships in the novel where characters seldom co-occur.



(a) Co-occurrence overlaid with HWI (b) Co-occurrence overlaid with SSM



(c) HWI overlaid with SSM

Figure 2.8: Overlaid Estimates for *Dream of the Red Chamber* (O-Black, HWI-Red, SSM-Green)

From the novel, Lin Daiyu is a sensitive and sickly girl who prefers to be alone. By contrast, Xue Baochai is a social and calculating girl. She is extremely good at interpersonal communication, especially with the protagonist’s mother (Lady Wang) and grandmother (Grandmother Jia). These significantly different personalities are clearly represented by the SSM estimator while the other estimators do not identify this difference.

Xue Baochai, generally, has much stronger relationships with other characters except for three: Jia Baoyu (the protagonist), Miaoyu (a nun with a similar personality to Lin Daiyu) and Zijuan (a maid of Lin Daiyu). The co-occurrence matrix and half-weight index fail to show such a clear pattern.

Table 2.6: Relationships of Lin Daiyu and Xue Baochai to other characters in *Dream of the Red Chamber*

	Co-Occurrence Matrix (O)		Half Weight Index (H)		Symmetric Star (\hat{A})	
	Lin Daiyu	Xue Baochai	Lin Daiyu	Xue Baochai	Lin Daiyu	Xue Baochai
Jia Baoyu	0.1728	0.1274	0.4563	0.3587	0.3113	0.2258
Lin Daiyu	1.0000	0.1109	1.0000	0.4866	1.0000	0.4072
Xue Baochai	0.1109	1.0000	0.4866	1.0000	0.4072	1.0000
Jia Yuanchun	0.0072	0.0050	0.0531	0.0449	0.0156	0.0228
Jia Tanchun	0.0439	0.0533	0.2490	0.3482	0.0915	0.4848
Shi Xiangyun	0.0590	0.0490	0.3273	0.3119	0.2194	0.2365
Miaoyu	0.0072	0.0036	0.0552	0.0337	0.0597	0
Jia Yingchun	0.0252	0.0274	0.1667	0.2141	0	0.2846
Jia Xichun	0.0187	0.0202	0.1313	0.1692	0.0102	0.2461
Wang Xifeng	0.0497	0.0526	0.1840	0.2131	0.0317	0.0697
Jia Qiaojie	0.0022	0.0022	0.0170	0.0208	0	0.0348
Li Wan	0.0367	0.0482	0.2086	0.3160	0.0580	0.3384
Qin Keqing	0.0007	0.0007	0.0052	0.0062	0	0
Grandmother Jia	0.0655	0.0648	0.2725	0.2985	0.1925	0.2820
Jia She	0.0065	0.0043	0.0449	0.0357	0	0
Jia Zheng	0.0122	0.0144	0.0701	0.0952	0.0143	0.0174
Jia Lian	0.0072	0.0036	0.0423	0.0245	0.0002	0.0073
Xiangling	0.0180	0.0252	0.1185	0.1961	0.0741	0.2344
Ping'er	0.0122	0.0209	0.0668	0.1306	0.0016	0.1643
Xue Pan	0.0043	0.0101	0.0292	0.0809	0	0
Granny Liu	0.0072	0.0050	0.0493	0.0411	0.0101	0.0113
Lady Wang	0.0490	0.0590	0.2248	0.3037	0.0224	0.2065
Aunt Xue	0.0302	0.0396	0.1806	0.2750	0.0479	0.1657
Hua Xiren	0.0403	0.0389	0.1938	0.2105	0.0283	0.1469
Qingwen	0.0166	0.0115	0.1020	0.0829	0.0155	0.0886
Yuanyang	0.0086	0.0101	0.0556	0.0763	0	0.0430
Mingyan	0.0007	0.0007	0.0053	0.0064	0	0
Zijuan	0.0317	0.0108	0.2184	0.0888	0.1775	0.0376
Concubine Zhao	0.0050	0.0058	0.0361	0.0495	0	0.0338

2.6.2 Dolphins network

In the preceding example, the number of observed groups exceeded the number of parameters. However, in some datasets, it is possible that a small number of observations are

available either due to the costs, or the time associated with collecting the data. In the following example, we focus on a dataset where the number of parameters significantly exceeds the number of observations.

Bejder et al. (1998) presents a dataset with observations of 18 Hector’s dolphins. With $n = 18$, there are 170 parameters which must be estimated; however, only 40 observations were taken on this population. We first present the co-occurrence matrix and half-weight index for the dataset in Figure 2.9.

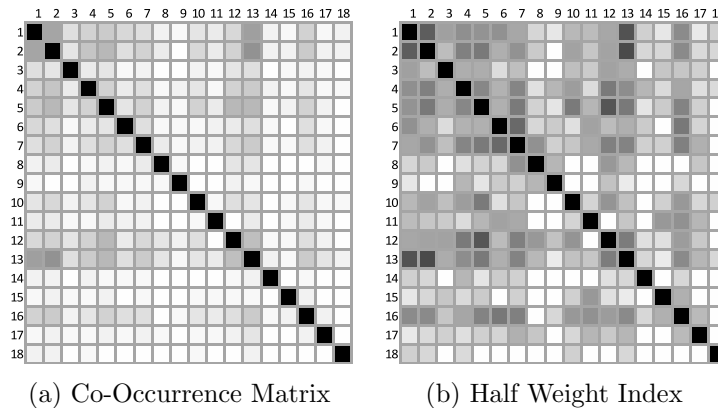


Figure 2.9: Conventional Estimates for Dolphins

Since there are more parameters than observations, we would expect that there would be multiple network structures which could explain the observed behavior. To demonstrate this, we applied Algorithm 1 using two different random number streams to obtain the two adjacency matrices shown in Figure 2.10 and found distinctly different structures. As an example, consider the relationships associated with node v_4 . In the left panel, node v_4 only has five relationships. Two are strong while the other three are more modest. However, in the right panel node v_4 has more and different relationships with a total of five strong relationships.

One reason that we select this dataset is to re-enforce the point made by Voelkl et al. (2011) to pay attention to the size of raw datasets in network analysis. In fact, this dolphin

dataset and similar datasets have been studied in several literatures of network analysis and used as benchmarks in community detection (Bejder et al., 1998; Girvan and Newman, 2002; Lusseau et al., 2003). However, the networks used in these literatures are constructed by ad-hoc measures such as the half-weight index. Researchers may analyze these networks without being aware that the raw data contains too few observations for satisfactory estimation of the latent networks.

Despite the instability of the estimates in Figure 2.10, both demonstrate a great deal of sparsity because every zero that appears in the co-occurrence matrix, O , also appears in \hat{A} . Interestingly, the sparsity of \hat{A} does not come solely from the sparsity of O , i.e., $\hat{A}_{ij} = 0$ does not imply that $O_{ij} = 0$. In the dolphin example, O has only 33 elements above the diagonal which have a value of zero, but there are 74 zeros in \hat{A}_1 and 77 zeros in \hat{A}_2 . Since we do not apply a regularization method (any type of L_1 penalty) in the objective function, we call this phenomenon a *self-sparsity* property of the SSM. The source and mathematical explanation of self-sparsity is discussed in the next section. However, we will provide some intuition and elementary observations of this property first.

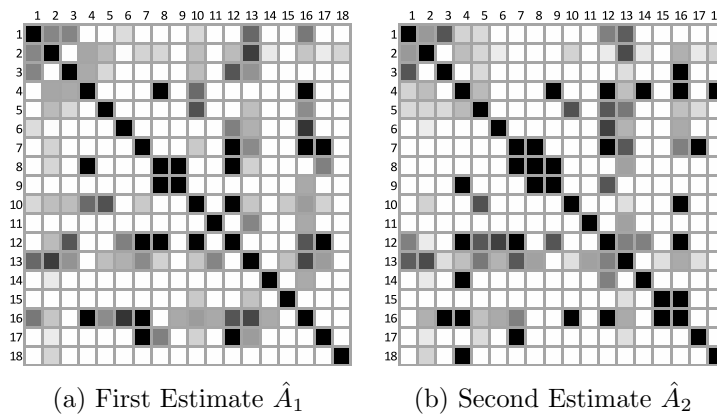


Figure 2.10: Adjacency Matrix Estimates for Dolphins

This sparsity is not simply present in the estimated adjacency matrices, but also appears in $\hat{\rho}$. In Table 2.7, the $\hat{\rho}$'s estimated with the two adjacency matrices in Figure 2.10 are given

along with the number of times that each node appears in the dataset. In each estimate, there are several $\hat{\rho}_i$'s which are equal to zero, but these do not occur at the same nodes. Moreover, there is no apparent threshold for the number of observations below which all $\hat{\rho}_i$ are equal to zero. Consider v_{12} , which appears 15 times and yet $\hat{\rho}_{12}^{(1)} = 0$. Similarly, v_4 appears 12 times and v_6 only appears 8 times; but in both estimates of $\hat{\rho}$, node v_6 is more likely to be the center of a group than v_4 .

Table 2.7: Two Different estimates of ρ for Dolphins

i	Nodes																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
$\sum G_i$	20	24	7	12	18	8	9	5	2	9	3	15	24	2	2	11	4	2
$\rho^{(1)}$	0.132	0.302	0.075	0	0.091	0.05	0	0.025	0	0.12	0	0	0.105	0	0	0.075	0.025	0
$\rho^{(2)}$	0.147	0.295	0	0	0.179	0.05	0.025	0	0.025	0	0	0.05	0.204	0	0	0.025	0	0

2.6.3 Self-sparsity

In Section 2.6.2, we introduced an interesting property of the Symmetric Star Model estimators which we refer to as *self-sparsity*. When T is small relative to n , the model tends to produce a sparse adjacency matrix. This sparsity in A is a result of sparsity either in G itself or in ρ , hence the name.

To see why this sparsity exists, recall the M-step for maximum likelihood of A (2.27) and ρ (2.35), as well as the E-step in (2.37).

Self sparsity derives primarily from the fact that $\rho_x = 0$ implies that $\mathbb{P}(S_x^{(t)} = 1 | G^{(t)}, A) = 0$ for all t . This is simply an application of (2.37).

To begin, notice that $\rho_x = 0$ is an absorbing state. That is, when $\rho_x^{(k)} = 0$ for iteration k , then $\rho_x^{(k+1)} = 0$. Since $\rho_x^{(k)} = 0$ implies $\mathbb{P}(S_x^{(t)} = 1 | G^{(t)}, A) = 0$ for all t , in (2.35), the numerator will sum to zero and we will have $\rho_x^{(k+1)} = 0$.

Next, observe that if $\rho_x = 0$, then the values of the x^{th} row of A are irrelevant. Technically, they can assume any value since the numerator of (2.37) will always equal zero. One might be tempted to assign $A_{xy} = 0$ for all y when $\rho_x = 0$; but under the Symmetric Star Model, notice that the form of (2.27) is changed to:

$$\hat{A}_{xy} = \frac{\sum_t G_x^{(t)} \mathbb{P}(S_y = 1|G^{(t)})}{\sum_t \mathbb{P}(S_y = 1|G^{(t)})}. \quad (2.39)$$

(2.39) is the asymmetric form of (2.27) which would be unidentifiable if used in general (identifiability will be discussed in greater detail in Chapter 4). This means that under the SSM, when $\rho_x = 0$, the elements of the x^{th} row of A are simply reflections of their symmetric counterparts.

There are several specific situations which will lead to a sparsity in A .

Case 1: $G_i^{(t)} G_j^{(t)} = 0$ for all t

This situation exists when two nodes never co-occur. It is natural that when two nodes (v_i and v_j) are never observed together, we would estimate their preference for each other (i.e. A_{ij}) to be zero.

Observe that from (2.37), $G_i^{(t)} = 0$ implies $\mathbb{P}(S_i = 1|G^{(t)}) = 0$. This means that if either $G_i^{(t)} = 0$ or $G_j^{(t)} = 0$, then the numerator of (2.27) will equal zero for observation t . Therefore, if $G_i^{(t)} G_j^{(t)} = 0$ for all t , then $A_{ij} = 0$.

Case 2: $\rho_i = \rho_j = 0$

$\rho_i = \rho_j = 0$ implies that $\mathbb{P}(S_i = 1|G^{(t)}) = \mathbb{P}(S_j = 1|G^{(t)}) = 0$ for all t . This means that the numerator and denominator of (2.27) will both equal zero. Since $A_{ij} = \frac{0}{0}$ is undefined and, more importantly, can take on any value without effecting the likelihood function, we define $A_{ij} = 0$.

Case 3: $\hat{\rho}_i = O_i$

From Appendix B, we have the following equation for the probability that node v_i is observed:

$$O_i = \sum_k \rho_k A_{ki} \quad (2.40)$$

If the probability that a node is acting as a group leader is equal to the probability that the node is observed, $\hat{\rho}_i = O_i$, then we have:

$$O_i = \rho_i + \sum_{k \neq i} \rho_k A_{ki} \quad (2.41)$$

$$0 = \sum_{k \neq i} \rho_k A_{ki}. \quad (2.42)$$

Therefore, for every $k \neq i$ either $\rho_k = 0$ or $A_{ki} = 0$.

Case 4: When ρ_i is small

Even when ρ_x is close, but not exactly equal, to zero the SSM will tend to produce a sparse result for A . To see why this is true, recall that if either $G_y^{(t)} = 0$ or $\mathbb{P}(S_x = 1|G^{(t)}) = 0$ for all t , then $A_{xy} = 0$. Since ρ_x is close to zero, we will have $\mathbb{P}(S_x = 1|G^{(t)}) = 0$ for most observations.

2.7. Conclusion

To our best knowledge, Star Models introduce an innovative approach to social network inference. By defining a model-based generating mechanism to link the latent network structure to observed grouped data and applying an EM algorithm, we can estimate network structure.

Not only are the estimators easy to calculate in a reasonable amount of time, but the estimators have a practical interpretation. The parameter ρ measures the leadership or assertiveness of a population member. A_{ij} measures the popularity or probability that a member of the population will be included in a group.

The Star Models compare favorably against existing techniques which measure social behavior. Since the co-occurrence matrix and half weight index lack a generating mechanism to connect them to the observed grouped data, these measures miss important features of a community's social behavior.

By applying the Symmetric Star Model to the 18th century Chinese novel *Dream of the Red Chamber*, we demonstrate that the SSM is able to detect important differences in the relationships between characters of a complex story. We also demonstrated that when the number of observations is small compared to the number of individuals in a population, as in the case of the dolphin dataset, $\{\hat{A}, \hat{\rho}\}$ shows signs of instability and “self-sparsity”.

Chapter 3: Inducing Sparsity with Penalized Rho Star Model

3.1. Introduction

In the last chapter, we saw that, for small datasets, it is possible to find multiple solutions to the Symmetric Star Model (SSM) which had similar likelihoods (see Figure 2.10). This is particularly true when the number of observations, T , is less than number of parameters to be estimated, d ; however, it is most dramatic when the number of observations is less than the number of members in the population, $T < n$.

If we consider the classic linear regression model, we can see that this behavior is not particularly surprising. Suppose we attempt to perform a linear regression on d parameters with T observations where $T < d$. This situation would be under determined and would produce multiple equally likely sets of parameters. Similarly, we think of the network inference problem with a low number of observations as an under determined system. One plausible solution is to reduce the number of parameters in the SSM.

Moreover, reducing the number of parameters sometimes is necessary even when the uniqueness of a solution is not a problem. Table 2.3 shows that for a network with moderate size ($n = 50$), a huge number of observations is needed to achieve accurate estimation since the number of parameters has order $O(n^2)$. In this chapter, we make an assumption of sparsity on parameters and propose a method based on regularization of ρ .

3.1.1 Motivation

The classic method of dealing with such a situation in linear regression is to apply the ‘least absolute shrinkage and selection operator’ (LASSO) (Tibshirani, 1996). This technique and its variants have been applied in a number of different ways to reduce the number of non-zero regression coefficients in a linear model.

Taking such an approach reduces the variance of individual parameter estimates by allowing bias into the model. LASSO is able to shrink some parameters exactly to zero, which has two advantages. First, this approach is useful for model selection in identifying the most relevant parameters. Second, applying such shrinkage can, in fact, improve the predictive ability of a model despite the added bias.

Our basic goal is to introduce greater sparsity into A . However, as will be seen later, it is not enough to penalize the elements of A towards zero.

In the classic application of LASSO, an L_1 penalty is applied to the magnitude of the parameters of the model,

$$\min_{\beta} SSE(\beta) + \eta \sum_{j=1}^p |\beta_j| \quad (3.1)$$

where η is a tuning parameter which forces the parameters β towards zero more strongly as it increases.

If η is increased until $\sum_{j=1}^p |\beta_j| = 0$, the model will be reduced to the null model. In the linear regression situation, the null model is simply the mean value of the response variable.

However, in the network inference situation, nodes should behave independently in the null model. That is, nodes appear in a group independently of the presence of other nodes. Therefore, attempting to make A sparse by penalizing the sum of the elements of A does not cause the Star Model to approach the null model. Since $A_{ij} = 0$ implies v_i will never approach v_j if v_i is the central node of the group, this violates the independence assumption. The null model will be discussed in more detail in Chapter 4.

From the discussion on self-sparsity, we see that sparsity in ρ will lead to sparsity in the adjacency matrix, since if both ρ_i and ρ_j equal zero, A_{ij} will not affect the likelihood of SSM. Based on this, we propose a method which further shrinks more elements of ρ towards zero. We refer to this as the *Penalized Rho Star Model* (PRSM).

The motivation for shrinking some ρ 's to zeros is obvious in the case of $T < n$ because it is impossible that every member of the population has been observed to be the center

of at least one group; therefore, some members of the population must have $\rho_i = 0$. Even in the case of the dolphin dataset where $n = 18$ and $T = 40$, it is very unlikely that every node has formed a group.

It is also useful to recognize that in the unpenalized SSM, ρ_i has an upper and lower bound. The upper bound is O_i . The lower bound is $\frac{\sum_t G_i^{(t)} \prod_{j \neq i} 1 - G_j^{(t)}}{T}$. The existence of a lower bound is important because for nodes which appear as singletons, it is impossible for their ρ to be exactly zero.

Therefore, it is not enough to apply an ad-hoc threshold for the number of times a node must be observed for its ρ to be non-zero.

3.2. Methodology

3.2.1 Linear Penalty

To penalize ρ , we apply a linear penalty based on the probability that node v_i is observed.

$$\eta \sum_i \frac{\rho_i T}{\sum_t G_i^{(t)}} = \eta \sum_i \frac{\rho_i}{O_i} \tag{3.2}$$

where η is a tuning parameter.

This penalty has a nice interpretation. Nodes which are observed very rarely have a stronger penalty applied to their ρ than nodes which are observed more often. That is, if a node is unlikely to occur, we assume that it is also unlikely to act as a leader in the population. Conversely, the more often a node is observed, the more likely that they are exerting a leadership role over the population. An additional benefit of the linear penalty is that it makes the optimization problem convex with respect to ρ .

3.2.2 Formulation of Optimization Problem

Our optimization problem can be formulated as follows:

$$\begin{aligned}
& \underset{A, \rho}{\text{minimize}} && \Lambda(G|A, \rho) = -\mathcal{L}(G|A, \rho) + \eta \sum_i \frac{\rho_i T}{\sum_t G_i^{(t)}} \\
& \text{subject to} && \left(\sum_k \rho_k \right) - 1 = 0. \\
& && A_{ij} - A_{ji} = 0, i = 1, \dots, n, j = 1, \dots, n. \\
& && -\rho_k \leq 0, i = 1, \dots, n. \\
& && -A_{ij} \leq 0, i = 1, \dots, n, j = 1, \dots, n. \\
& && A_{ij} - 1 \leq 0, i = 1, \dots, n, j = 1, \dots, n.
\end{aligned} \tag{3.3}$$

where $\eta \geq 0$.

Note that in this formulation, we have explicitly included $0 \leq \rho_i \leq 1$ and $0 \leq A_{ij} \leq 1$. These constraints were present in the formulation used in Chapter 2; however, the Expectation Maximization algorithm enforced them without special consideration. In the following sections, we will actively enforce the constraint that ρ_i is positive.

3.2.3 Algorithm

In Chapter 2, we were able to estimate A and ρ simultaneously using the EM algorithm. If we try estimating both parameters using the EM algorithm with the added penalty, we violate the constraint $\sum_i \rho_i = 1$.

To address this issue, we break A 's and ρ 's update into two separate tasks. We will continue to use the EM algorithm to update A while using a fixed ρ ; then we will hold A fixed and use convex optimization with inequality constraints to solve for ρ .

Estimating A with Fixed ρ

For fixed ρ , the derivative of (3.3) with respect to A_{xy} is the same as in Chapter 2; therefore, the estimator is the same as the one that we derived in the SSM (2.27).

Estimating ρ with Fixed A

Optimizing a penalized likelihood is challenging because the objective function is a high-dimensional non-concave function with singularities (Fan and Li, 2001). However, fixing A reduces (3.3) to a concave function with inequality constraints.

$$\begin{aligned} \underset{A, \rho}{\text{minimize}} \quad & \Lambda(G|A, \rho) = -\mathcal{L}(G|A, \rho) + \eta \sum_i \frac{\rho_i T}{\sum_t G_i^{(t)}} \\ \text{subject to} \quad & \left(\sum_k \rho_k \right) - 1 = 0, \\ & -\rho_k \leq 0 \quad \forall k \end{aligned} \tag{3.4}$$

To solve for ρ , we use the Interior Point Method. The Interior Point Method is a technique for solving optimization problems that have inequality constraints. Instead of solving the Karush-Kuhn-Tucker (KKT) conditions for the inequality constraints, we add a logarithmic penalty to the objective function such that as the estimated term approaches its boundary, the objective function tends towards positive infinity (Boyd and Vandenberghe, 2011).

Since (3.4) and the log penalty are convex in ρ , we can solve this problem numerically using Newton's Method.

We implement the Interior Point Method using MATLAB's `fmincon` function and provide the gradient of the objective function to reduce runtime.

Practical Considerations

There are three practical considerations which have to be addressed within our algorithm.

First, we have to find a reasonable starting point. Rather than selecting a random starting point, we begin at the solution to the unpenalized SSM.

Second, there is no guarantee that the algorithm will converge to a global optimum. In fact, it is possible that there are multiple local optimums in the parameter space. Suppose

we have two local optimal solutions $\{\rho_1, A_1\}$ and $\{\rho_2, A_2\}$. If, by holding A_1 fixed, we get ρ_2 and by holding A_2 fixed, we get ρ_1 , the algorithm can oscillate between equally satisfactory solutions. In this case, the objective function will fail to improve with each iteration. To handle this, we allow the algorithm to search five alternative solutions before declaring the current best solution optimal.

Finally, occasionally `fmincon` will produce a solution which is infeasible (e.g. $\rho_i = 0$ for some ρ_i which appears as a singleton). When this happens, we reject the solution and randomly select a new starting point.

```

Data:  $G, \eta$ 
Result:  $\hat{A}, \hat{\rho}$ 
Initialize:
bestOBJ=  $\infty$ 
counter= 1
 $\{A^{(0)}, \rho^{(0)}\} = \text{SSM}(G)$ 
while counter < 6 do
     $\rho^{(m)} = \text{fmincon}(G, \eta)$ 
     $A^{(m)} = \text{SSM}(G, \rho^{(m)})$ 
    currentOBJ=OBJFun( $G, A^{(m)}, \rho^{(m)}$ )
    if currentOBJ < bestOBJ then
        bestOBJ=currentOBJ
         $A_{opt} = A^{(m)}$ 
         $\rho_{opt} = \rho^{(m)}$ 
        counter= 1
    end
    counter=counter+1
end

```

Algorithm 2: Penalized Rho Algorithm

3.3. Selecting Tuning Parameters

The value of η is selected using the Bayesian Information Criterion (BIC).

The BIC is frequently used when a model is fitted using maximization of a log-likelihood (Hastie et al., 2009). The generic form of BIC is:

$$BIC = -2\mathcal{L} + (\log T)d \tag{3.5}$$

where \mathcal{L} is the log-likelihood, T is the number of observations, and d is the number of parameters. BIC tends to penalize complex models more heavily, giving preference to simpler models in selection.

In previous sections, we have calculated the number of parameters in the SSM using (2.16). This is the number of unique symmetric elements in A , $\binom{n}{2}$, and the number of independent parameters in ρ , $(n - 1)$. However, as the number of non-zero elements in ρ decreases from n , the structure of A changes and d will not follow this equation. If we let n_o be the number of non-zero elements in ρ , the form of A becomes:

$$\begin{array}{c|c} X & Y \\ \hline Y' & 0 \end{array}$$

where X is an $n_o \times n_o$ symmetric matrix and Y is an $n_o \times (n - n_o)$ matrix and 0 is an $(n - n_o) \times (n - n_o)$ matrix of zeros.

As a result, the formula for the number of parameters becomes:

$$d = \underbrace{\binom{n_o}{2}}_X + \underbrace{n_o(n - n_o)}_Y + \underbrace{(n_o - 1)}_\rho \quad (3.6)$$

Figure 3.1 shows how the number of parameters changes with n_o for a population of $n = 20$ individuals. When n_o is close to n , the reduction in parameters is not very significant. However, as the number of non-zero nodes decreases, the number of parameters begins to decline rapidly. This suggests that while PRSM will be beneficial in simplifying models when a significant number of ρ 's are reduced to zero, there will be little benefit if G contains many singletons which cannot be reduced to zero.

3.4. Simulation Studies

In previous sections, we used the mean absolute error (MAE) (2.38) as a measure of the overall accuracy of an estimate of A . We will continue to do so, but we will also begin to

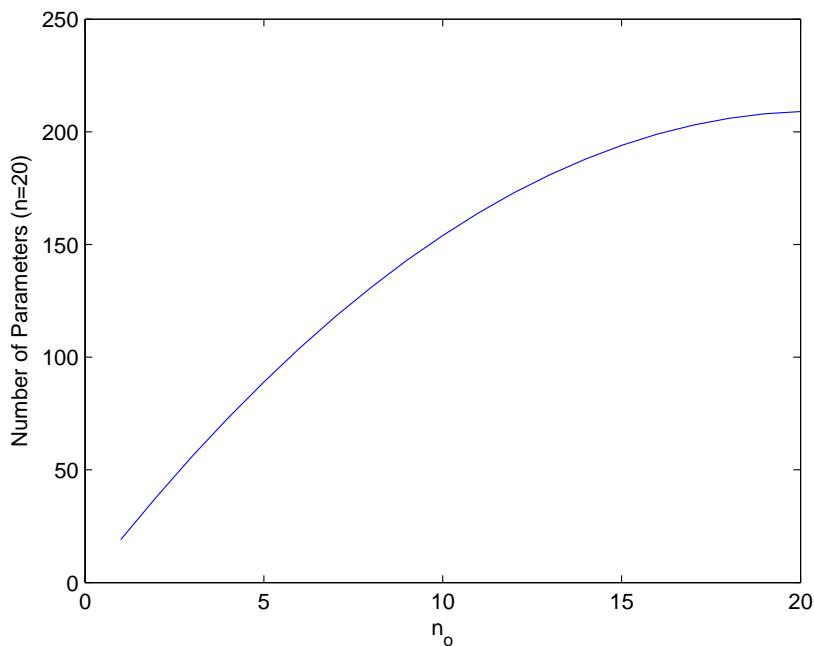


Figure 3.1: Number of Parameters as n_o Changes

focus on the estimated value of n_o .

3.4.1 Toy Example

Consider the following toy situation. We have an adjacency matrix represented in Table 3.1 where there are only two members of the population exerting leadership. We have a limited set of 20 observations shown in Table 3.2. Notice that node v_4 is not a leader, but it is very popular.

Table 3.3 shows the effect of applying Algorithm 2 with increasing η until the number of non-zero elements of ρ has been reduced to $n_o = 2$. Notice that when we apply SSM, the model estimates that nodes v_1 through v_4 all have non-zero ρ . This makes sense when we consider that each of these nodes appear in over half of the observations (see Table 3.2).

As η becomes larger, the first node to be forced to zero is v_3 . This reduction is achieved

Table 3.1: Example of a Population With Sparse ρ 's

ρ	i	j						
		1	2	3	4	5	6	7
0.5	1	1.0000	0.7854	0.0000	0.9063	0.0000	0.0000	0.7452
0.5	2	0.7854	1.0000	0.8324	0.8817	0.5885	0.8594	0.0000

Table 3.2: Frequency Table Low Observations Example

A							Frequency	Elements
1	2	3	4	5	6	7		
1	0	0	0	0	0	0	1	1
1	0	0	1	0	0	0	1	2
1	1	0	0	0	0	1	1	3
1	1	0	1	0	0	0	1	3
0	1	1	1	1	0	0	2	4
1	1	0	1	0	0	1	3	4
1	1	0	1	0	1	0	1	4
1	1	1	1	0	0	0	1	4
1	1	0	1	1	1	0	1	5
1	1	1	0	1	1	0	1	5
1	1	1	1	0	1	0	5	5
1	1	1	1	1	0	0	1	5
1	1	1	1	1	1	0	1	6
Number of Times Observed							20	
18	18	11	17	6	9	4		

with a relatively low value of $\eta = 20$. When η is increased to 500, the PRSM identifies the sparsity which was built into the original model and the Bayesian Information Criterion achieves its minimum.

Table 3.3: Bayesian Information Criterion as η Increases

η	j							LL	BIC	Parameters
	1	2	3	4	5	6	7			
0	0.3466	0.5224	0.0816	0.0494	0.0000	0.0000	0.0000	-54.7256	172.3616	21
1	0.3477	0.5201	0.0779	0.0542	0.0000	0.0000	0.0000	-54.7273	172.3650	21
2	0.3521	0.4538	0.0735	0.1207	0.0000	0.0000	0.0000	-54.6992	172.3089	21
5	0.3545	0.4610	0.0652	0.1193	0.0000	0.0000	0.0000	-54.7228	172.3559	21
10	0.3536	0.5333	0.0520	0.0611	0.0000	0.0000	0.0000	-54.8207	172.5517	21
20	0.3544	0.4556	0.0000	0.1900	0.0000	0.0000	0.0000	-55.2724	161.4722	17
50	0.3577	0.4853	0.0000	0.1569	0.0000	0.0000	0.0000	-55.3119	161.5512	17
100	0.3586	0.5385	0.0000	0.1028	0.0000	0.0000	0.0000	-55.6926	162.3126	17
200	0.3715	0.5446	0.0000	0.0840	0.0000	0.0000	0.0000	-55.7212	162.3698	17
500	0.3428	0.6572	0.0000	0.0000	0.0000	0.0000	0.0000	-57.8849	151.7186	12
1000	0.3428	0.6572	0.0000	0.0000	0.0000	0.0000	0.0000	-57.8853	151.7193	12
2000	0.3429	0.6571	0.0000	0.0000	0.0000	0.0000	0.0000	-57.8856	151.7199	12
5000	0.3429	0.6571	0.0000	0.0000	0.0000	0.0000	0.0000	-57.8853	151.7195	12

In addition to the effects of PRSM on ρ , we consider what happens to A as η increases. In Table 3.4, we can see that the basic structure of the true adjacency matrix is present with minimal penalty. Further, in Table 3.5 we can see that the same structure is preserved and that there is no significant distortion of individual elements (i.e. A_{14}) when $\eta = 500$.

Table 3.4: Estimated Adjacency Matrix for $\eta = 1$

ρ	i	j						
		1	2	3	4	5	6	7
0.3477	1	1.0000	0.8817	0.0000	0.7096	0.0000	0.0000	0.5768
0.5201	2	0.8817	1.0000	0.8280	0.9125	0.3240	0.8674	0.0000
0.0779	3	0.0000	0.8280	1.0000	1.0000	1.0000	0.0000	0.0000
0.0542	4	0.7096	0.9125	1.0000	1.0000	0.9515	0.0000	0.0000

One might observe that the estimates in Tables 3.4 and 3.5 are not very accurate when compared to the true adjacency matrix (Table 3.1); however, there are only 20 observations,

Table 3.5: Estimated Adjacency Matrix for $\eta = 500$

		j						
ρ	i	1	2	3	4	5	6	7
0.3428	1	1.0000	0.8000	0.0000	0.7083	0.0000	0.0000	0.5834
0.6572	2	0.8000	1.0000	0.8369	0.9239	0.4565	0.6847	0.0000

so it is not surprising that the estimates are not very accurate.

3.4.2 BIC for Increasing η with Fixed Number of Parameters

In Table 3.3, we notice that when the number of parameters decreases, it does so with a sudden drop. This results in a sudden drop in the BIC (3.5). However, observe that, in general, the lowest value of η that produces a reduction in parameters seems to also be the η associated with the lowest BIC. That is, BIC tends to increase as η increases when the number of parameters remains the same. This is a natural result of (3.5).

3.4.3 Estimating Parameters for Large Sparse Networks ($n = 50$)

In the previous chapter, we observed that as the number of observations increases, the MAE of the estimated A improves. However, for situations where there is a large number of nodes, the improvement rate can be very slow. Since PRSM does not seem to distort the individual elements of A , we now show that for sparse latent network structure, the estimates converge to the true adjacency matrix much faster.

To simulate this situation, we focus on a system of 50 nodes where $n_o = 8$. In initial testing, a penalty of $\eta = 500$ was found to consistently produce minimal BIC solutions. For each scenario, we produced 100 sets of observations and estimated $\{\hat{A}, \hat{\rho}\}$ for each set of observations. We then calculated descriptive statistics for MAE and n_o .

The results of this experiment are shown in Table 3.6. As expected, the accuracy of

both the SSM and PRSM improves as the number of observations increases. This is true of both \hat{A} and the estimated number of non-zero elements in ρ .

More importantly, PRSM converges faster than SSM to the correct \hat{n}_o . For example, with only 500 observations, PRSM performs as well as SSM with 4000 observations. Also, we observe that when the SSM and PRSM estimate the same value for \hat{n}_o , they tend to estimate the same degree of error in \hat{A} . Finally, when there are sufficient observations for the SSM to correctly identify the number of non-zero parameters in ρ , PRSM doesn't appear to provide much improvement in estimation.

Table 3.6: Average and Standard Deviation of Mean Absolute Error as Observations Increase

Obs	Symmetric Star Model				Penalized Rho Star Model			
	Avg MAE	StDev MAE	Avg n_o	StDev n_o	Avg MAE	StDev MAE	Avg n_o	StDev n_o
200	0.0421	0.0046	28.21	2.3151	0.0390	0.0059	22.28	6.4135
500	0.0393	0.0040	26.52	2.4842	0.0299	0.0078	15.46	5.2309
1000	0.0386	0.0048	25.28	2.2160	0.0256	0.0114	11.50	4.6046
2000	0.0367	0.0049	21.75	2.2264	0.0216	0.0112	10.37	4.4099
4000	0.0302	0.0055	16.47	2.1342	0.0185	0.0123	9.18	3.1699
10000	0.0222	0.0068	9.49	1.1055	0.0167	0.0122	8.16	0.3685
20000	0.0172	0.0069	8.20	0.4020	0.0144	0.0102	8.24	1.4079
50000	0.0116	0.0060	8.00	0.0000	0.0123	0.0111	8.00	0.0000

3.5. Data Analysis

3.5.1 Penalized Rho Star Model for Dolphins

The initial motivation for the PRSM was to introduce sparsity on ρ to reduce the number of parameters in the SSM in situations where the number of observations is small compared

to the number of members in the population. In the last chapter, we saw this in the dolphin data where there are 18 individuals but only 40 observations.

This situation and the dataset itself deserve some attention before considering the estimates derived from the PRSM. First, the dataset has four nodes which appear as singletons (Nodes v_1, v_2, v_5 , and v_{13} are indicated in the following tables by placing them in parentheses.) Since nodes which appear as singletons cannot have $\rho_i = 0$, the minimal value of n_o is 4. Applying (3.6), we will have to estimate at least 65 parameters, which still exceeds our observations.

In Tables 3.7 and 3.8, we give the results of using two different random number streams for the data evaluation. The first interesting thing to notice about these two tables is that they achieve their minimum BIC at the same η . This seems to be a pattern of the PRSM and is not fully understood.

Secondly, notice that the two highlighted solutions tend to agree closely with each other. There is disagreement on only three nodes, $\{v_4, v_7, v_{10}\}$, and these are nodes with relatively low values. This is much better performance than in Table 2.7 where there are 7 nodes of disagreement and significant differences in estimated values.

Finally, notice that the number of estimated parameters is far above the number of observations or even the minimal expected number of parameters. This suggests that while applying PRSM to this small dataset should stabilize the estimate of \hat{A} , we should not expect the effect to be perfect. This can be observed in Figure 3.2. While the solutions are more consistent than in Figure 2.10, there is still some instability.

Table 3.7: $\hat{\rho}_1$ for Dolphin Data

η	ρ																				LL	BIC	r_o	Param.
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18						
0	0.1260	0.2527	0.0250	0.0750	0.1592	0.0500	0.0000	0.0000	0.0250	0.0000	0.0000	0.0000	0.2371	0.0000	0.0000	0.0250	0.0000	0.0000	-269.1730	939.7741	10	134		
1	0.2063	0.2226	0.0197	0.0000	0.2007	0.0472	0.0238	0.0000	0.0174	0.0172	0.0000	0.0519	0.1688	0.0000	0.0000	0.0244	0.0000	0.0000	-275.2341	975.8621	11	142		
2	0.1252	0.3041	0.0215	0.0000	0.1986	0.0443	0.0227	0.0000	0.0133	0.0000	0.0000	0.0247	0.2221	0.0000	0.0000	0.0236	0.0000	0.0000	-269.1257	939.6796	10	134		
5	0.1907	0.3173	0.0385	0.0585	0.0949	0.0000	0.0020	0.0015	0.0000	0.0780	0.0011	0.0025	0.1519	0.0008	0.0008	0.0465	0.0143	0.0008	-314.1080	1128.5032	16	167		
10	0.2253	0.2982	0.0265	0.0399	0.1891	0.0299	0.0000	0.0103	0.0000	0.0000	0.0000	0.0230	0.1304	0.0000	0.0000	0.0188	0.0086	0.0000	-273.8450	973.0839	11	142		
20	0.1377	0.3546	0.0087	0.0000	0.1632	0.0199	0.0112	0.0000	0.0025	0.0000	0.0000	0.0000	0.3022	0.0000	0.0000	0.0000	0.0000	0.0000	-277.3189	899.1470	8	115		
50	0.0875	0.4429	0.0042	0.0000	0.0906	0.0054	0.0000	0.0000	0.0011	0.0045	0.0000	0.0000	0.3551	0.0000	0.0000	0.0077	0.0011	0.0000	-298.1092	997.6465	10	134		
100	0.0701	0.4749	0.0023	0.0052	0.1012	0.0112	0.0000	0.0015	0.0000	0.0000	0.0000	0.0080	0.3256	0.0000	0.0000	0.0000	0.0000	0.0000	-293.4648	961.3962	9	125		
200	0.0541	0.5967	0.0012	0.0028	0.0540	0.0029	0.0000	0.0008	0.0003	0.0000	0.0000	0.0045	0.2730	0.0000	0.0000	0.0097	0.0000	0.0000	-299.9348	1025.2637	11	142		
500	0.0293	0.5624	0.0005	0.0000	0.0266	0.0012	0.0009	0.0000	0.0001	0.0003	0.0000	0.0019	0.3768	0.0000	0.0000	0.0000	0.0000	0.0000	-311.0611	1023.5503	10	134		
1000	0.1431	0.2494	0.0453	0.0000	0.2597	0.0500	0.0000	0.0000	0.0250	0.0250	0.0000	0.0000	0.1274	0.0000	0.0000	0.0500	0.0250	0.0000	-271.2443	943.9167	10	134		
O_i	(20)	(24)	7	12	(18)	8	9	5	2	9	3	15	(24)	2	2	11	4	2						

Table 3.8: $\hat{\rho}_2$ for Dolphin Data

η	ρ																		LL	BIC	n_o	Param.
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18				
0	0.1082	0.3522	0.0750	0.0000	0.0889	0.0500	0.0000	0.0250	0.0000	0.1215	0.0000	0.0000	0.0793	0.0000	0.0000	0.0749	0.0250	0.0000	-274.6306	950.6894	10	134
1	0.1141	0.3557	0.0232	0.0000	0.2164	0.0502	0.0239	0.0000	0.0174	0.0000	0.0000	0.0249	0.1499	0.0000	0.0000	0.0243	0.0000	0.0000	-277.5394	956.5070	10	134
2	0.0357	0.1695	0.0607	0.0000	0.2021	0.0677	0.0688	0.0000	0.0136	0.0625	0.0166	0.0512	0.2480	0.0000	0.0000	0.0000	0.0037	0.0000	-290.3162	1026.9964	12	149
5	0.1482	0.3691	0.0539	0.0229	0.1670	0.0384	0.0000	0.0000	0.0079	0.0187	0.0000	0.0000	0.1169	0.0000	0.0000	0.0441	0.0130	0.0000	-269.4204	964.2348	11	142
10	0.1300	0.3378	0.0133	0.0000	0.1318	0.0295	0.0160	0.0000	0.0046	0.0312	0.0000	0.0674	0.2384	0.0000	0.0000	0.0000	0.0000	0.0000	-271.2002	943.8285	10	134
20	0.1059	0.3620	0.0082	0.0302	0.1663	0.0204	0.0000	0.0000	0.0025	0.0113	0.0000	0.0000	0.2933	0.0000	0.0000	0.0000	0.0000	0.0000	-279.1031	932.6727	9	125
50	0.1118	0.4548	0.0043	0.0090	0.1289	0.0103	0.0000	0.0027	0.0011	0.0000	0.0000	0.0000	0.2612	0.0000	0.0000	0.0158	0.0000	0.0000	-279.7590	960.9460	10	134
100	0.1136	0.3518	0.0008	0.0052	0.1027	0.0056	0.0000	0.0009	0.0005	0.0000	0.0000	0.0000	0.4143	0.0000	0.0000	0.0045	0.0000	0.0000	-291.6426	984.7133	10	134
200	0.0535	0.4168	0.0000	0.0050	0.0311	0.0029	0.0000	0.0007	0.0003	0.0017	0.0000	0.0000	0.4857	0.0000	0.0000	0.0023	0.0000	0.0000	-298.7058	998.8397	10	134
500	0.0109	0.5123	0.0005	0.0000	0.0135	0.0012	0.0004	0.0000	0.0000	0.0000	0.0000	0.0019	0.4594	0.0000	0.0000	0.0000	0.0000	0.0000	-314.3525	973.2142	8	115
1000	0.0150	0.0080	0.0000	0.0000	0.0026	0.0006	0.0000	0.0002	0.0000	0.0004	0.0000	0.0000	0.9727	0.0000	0.0000	0.0005	0.0000	0.0000	-340.9349	1161.1869	14	160
O_i	(20)	(24)	7	12	(18)	8	9	5	2	9	3	15	(24)	2	2	11	4	2				

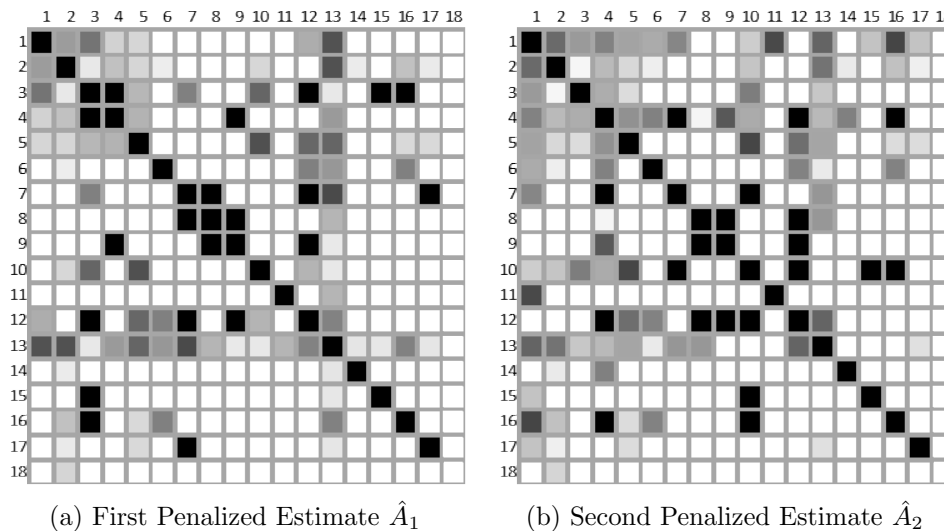


Figure 3.2: Penalized Adjacency Matrix Estimates for Dolphins ($\eta = 20$)

3.5.2 Penalized Rho Star Model for *Dream of the Red Chamber*

In addition to being useful for handling datasets where the number of observations is smaller than the number of individuals, the Penalized Rho Star Model can be useful for simplifying a model when the size of the population is large.

To demonstrate this, we apply the Penalized Rho Star Model to a reduced dataset from *Dream of the Red Chamber*. Since the original dataset focuses on the 29 most central characters, it contains a large number of singletons and pairs.

Singletons are important because groups consisting of a single node can only occur if the observed node is the central node of the group. Therefore, the ρ associated with a singleton cannot be exactly equal to zero. All but four characters appear as singletons.

Pairs have a similar effect on the Star Model. Since every pair must have at least one member with a non-zero ρ , datasets with a large number of unique pairs can be constrained in the number of nodes whose ρ can be penalized to zero.

To apply the Penalized Rho Star Model to the *Dream of the Red Chamber* dataset, we

remove all singletons and pairs. This dataset reduction is strictly for purposes of illustrating the behavior of Penalized Rho Star Model when simplifying a model where the population size is large.

The results of this simulation are shown in Tables 3.9 and 3.10. One might expect to get results that are inconsistent with the story using the reduced dataset. For the Symmetric Star Model (i.e. $\eta = 0$), we observe ρ values which are distinctly different than those in Table 2.4.

In particular, the main character is estimated to no longer play a central role. This is a result of the removal of singletons and pairs. 35% of the original observations of Jia Baoyu occur when he is alone or with only one other character. However, as the penalty is increased, the number of the parameters decreases and the centrality of the main character reasserts itself.

Since our interest lies in reducing the complexity of the dataset, we will focus on the solution with the lowest number of parameters, $\eta = 500$. At this level of penalty, eleven individuals are estimated to have $\rho = 0$.

Figure 3.3 shows the estimated relationships between characters based on the reduced dataset. As before, characters have been sorted by ρ in descending order.

Clearly, this estimate is different from the estimate shown in Figure 2.5c. However, the difference does not appear to destroy the social structure of the data. Important characters like Grandmother Jia and Lady Wang remain important and trivial characters are ranked low. Lin Daiyu still has fewer and weaker relationships than Xue Baochai. The resulting estimate also remains sparse with high contrast between relationships.

This initial experimentation suggests that for datasets with many individuals appearing in large groups, the Penalized Rho Star Model has potential for reducing model complexity without sacrificing inferential ability.

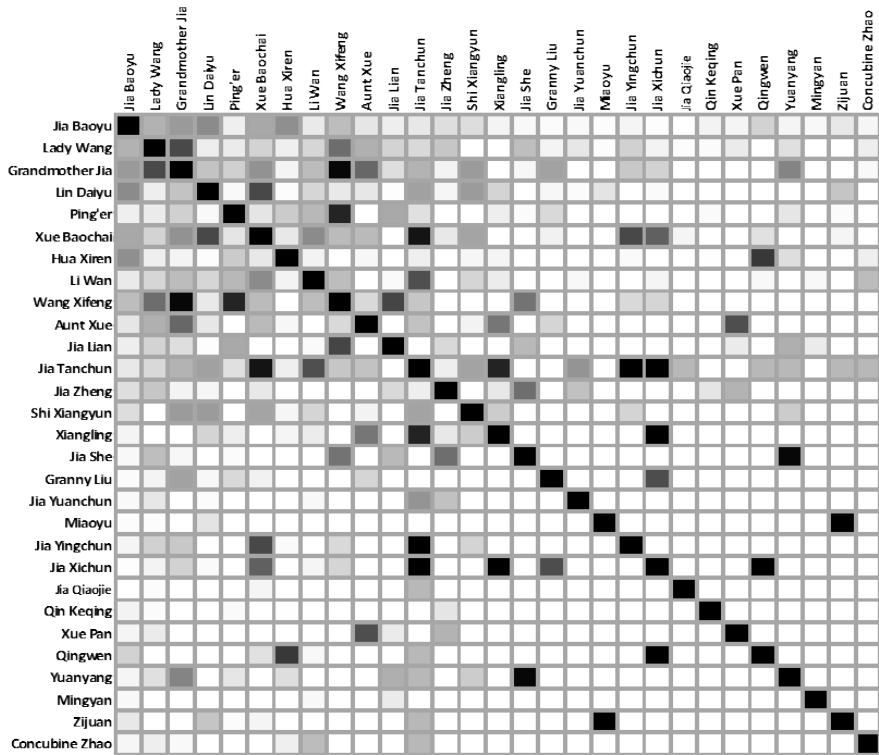


Figure 3.3: Estimated Relationships for *Dream of the Red Chamber* ($\eta = 500$)

Table 3.9: $\hat{\rho}$ for Reduced Dream of the Red Chamber Data

l	ρ													LL	BIC	n_o	Param.		
	Jia Baoyu	Lin Daiyu	Xue Baochun	Jia Yuanhun	Jia Tanchun	Shi Xiangyun	Miaoyu	Jia Yingchun	Jia Xichun	Wang Xifeng	Jia Qiaojie	Li Wan	Qin Keqing					Grandmother Jia	Jia She
0	0.000	0.208	0.040	0.005	0.019	0.027	0.000	0.007	0.002	0.001	0.000	0.076	0.016	0.009	0.020	-6127.077	15027.550	25	424
1	0.005	0.196	0.054	0.002	0.024	0.025	0.000	0.000	0.002	0.001	0.000	0.072	0.014	0.012	0.016	-6154.612	15010.669	23	413
2	0.000	0.202	0.034	0.004	0.013	0.028	0.000	0.007	0.000	0.114	0.000	0.069	0.012	0.009	0.016	-6162.051	15097.499	25	424
5	0.000	0.207	0.050	0.005	0.000	0.020	0.000	0.000	0.000	0.000	0.000	0.074	0.010	0.088	0.010	-6170.885	14945.101	21	398
10	0.000	0.215	0.042	0.004	0.010	0.027	0.000	0.009	0.001	0.002	0.000	0.078	0.010	0.009	0.018	-6132.159	14965.763	23	413
20	0.005	0.195	0.082	0.000	0.035	0.019	0.000	0.002	0.001	0.002	0.000	0.066	0.006	0.033	0.009	-6161.199	15063.090	24	419
50	0.028	0.220	0.000	0.002	0.024	0.008	0.000	0.001	0.000	0.154	0.000	0.054	0.000	0.221	0.008	-6280.133	15261.711	23	413
100	0.000	0.239	0.050	0.000	0.023	0.008	0.000	0.000	0.000	0.002	0.000	0.049	0.002	0.188	0.001	-6205.036	14954.533	20	389
200	0.525	0.093	0.037	0.000	0.000	0.004	0.000	0.000	0.000	0.129	0.000	0.021	0.000	0.038	0.002	-6370.652	15344.633	21	398
500	0.741	0.049	0.020	0.000	0.003	0.002	0.000	0.000	0.000	0.011	0.000	0.014	0.000	0.050	0.000	-6545.446	15497.991	18	368
1000	0.838	0.030	0.025	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.006	0.000	0.083	0.000	-6783.078	16221.814	22	406
2000	0.946	0.002	0.007	0.000	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.008	0.000	0.003	0.000	-6960.617	16661.925	24	419
5000	0.977	0.004	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.002	0.000	-7334.618	17501.503	28	433
10000	0.986	0.002	0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.006	0.000	-7396.048	17199.195	18	368
O_i	453	246	232	23	126	104	12	62	42	260	4	129	19	279	46				

Table 3.10: $\hat{\rho}$ for Reduced Dream of the Red Chamber Data

l	ρ														LL	BIC	n_o	Param.	
	Jia Zheng	Jia Lian	Xiangling	Ping'er	Xue Pan	Granny Liu	Lady Wang	Aunt Xue	Hua Xiren	Qingwen	Yuanyang	Mingyan	Zhijuan	Concubine Zhao					
0	0.042	0.036	0.015	0.091	0.005	0.013	0.147	0.043	0.148	0.020	0.003	0.003	0.005	0.000	-6127.077	15027.550	25	424	
1	0.045	0.040	0.016	0.090	0.004	0.014	0.154	0.041	0.148	0.019	0.000	0.000	0.007	0.000	-6154.612	15010.669	23	413	
2	0.039	0.026	0.014	0.063	0.005	0.008	0.116	0.041	0.139	0.028	0.000	0.002	0.007	0.001	-6162.051	15097.499	25	424	
5	0.037	0.032	0.007	0.094	0.006	0.007	0.142	0.040	0.153	0.012	0.001	0.000	0.003	0.001	-6170.885	14945.101	21	398	
10	0.038	0.036	0.014	0.092	0.006	0.011	0.154	0.046	0.156	0.020	0.000	0.000	0.004	0.000	-6132.159	14965.763	23	413	
20	0.038	0.035	0.005	0.079	0.003	0.009	0.166	0.032	0.149	0.021	0.003	0.000	0.004	0.000	-6161.199	15063.090	24	419	
50	0.000	0.006	0.007	0.040	0.000	0.001	0.067	0.000	0.152	0.005	0.001	0.001	0.000	0.000	-6280.133	15261.711	23	413	
100	0.019	0.018	0.002	0.079	0.001	0.001	0.140	0.021	0.154	0.004	0.000	0.000	0.001	0.000	-6205.036	14954.533	20	389	
200	0.004	0.000	0.002	0.023	0.000	0.001	0.081	0.005	0.035	0.000	0.000	0.000	0.000	0.000	-6370.652	15344.633	21	398	
500	0.002	0.003	0.000	0.024	0.000	0.000	0.058	0.005	0.018	0.000	0.000	0.000	0.000	0.000	-6545.446	15497.991	18	368	
1000	0.002	0.000	0.010	0.000	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.000	0.000	0.000	-6783.078	16221.814	22	406	
2000	0.000	0.001	0.000	0.006	0.000	0.000	0.021	0.001	0.002	0.000	0.000	0.000	0.000	0.000	-6960.617	16661.925	24	419	
5000	0.000	0.000	0.000	0.009	0.000	0.000	0.001	0.002	0.002	0.000	0.000	0.000	0.000	0.000	-7334.618	17501.503	28	433	
10000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.001	0.000	0.000	0.000	0.000	0.000	-7396.048	17199.195	18	368	
O_i	74	81	49	129	32	39	231	110	184	86	59	11	47	31					

3.6. Conclusion

In previous work, it has been shown that datasets with relatively low numbers of observations tend to produce unstable \hat{A} . In this chapter, we leveraged the self sparsity property of the SSM to propose a method for applying a linear penalty to ρ to induce sparsity in \hat{A} .

This approach corresponds to an application of LASSO in a linear regression setting and has a number of desirable properties that can be applied to small datasets.

First, elements of ρ that are close to zero are penalized towards zero, thereby reducing the complexity of the model. The benefits of this reduction in complexity are measured using the Bayesian Information Criterion. We further saw that the optimal level of the penalty η was the smallest value that reduced the non-zero elements of ρ to their smallest value.

Second, it was shown that the number of observations needed to estimate the true adjacency matrix is reduced. That is, when working with small datasets that are the result of sparse latent network structures, the PRSM converges to the true adjacency matrix faster than the SSM.

Finally, even when there are too few observations to support the necessary number of parameters, the PRSM induces more stability into solutions than we observed in the SSM. This stability manifests itself both in ρ and A .

Chapter 4: Future Work

In this chapter, we will discuss three avenues for future work and finish with a general discussion of applications that can expand on existing work. In the first section, we will present an alternative representation of the Star Models as a finite mixture model of multivariate Bernoulli random variables and provide a motivation for applying this model. This will lead to a discussion of the conditions for the model to be identifiable.

In the second section, we will propose a more general mechanism of group formation and discuss techniques for solving the problem.

In the third section, we propose some relationships which may be useful for efficiently enumerating connected networks.

4.1. Identifiability of Star Models

4.1.1 Definition of Multivariate Bernoulli Random Variables

In this section, we reintroduce the Star Models as a finite mixture model of multivariate Bernoulli random variables.

A multivariate Bernoulli random variable can be thought of as a set of different coins which are tossed simultaneously so that each coin's result can be uniquely observed. Therefore, the parameters of an n sized multivariate Bernoulli random variable are $\pi = \{\pi_1, \dots, \pi_n\}$ where π_i does not necessarily equal π_j . The parameter π_i is the probability that coin i will come up as a head.

The probability mass function of the multivariate Bernoulli distribution is (Dai et al., 2013)

$$\mathbb{P}(G|\pi) = \prod_{t=1}^T \prod_{i=1}^n \pi_i^{G_i^{(t)}} (1 - \pi_i)^{(1-G_i^{(t)})} \quad (4.1)$$

It is important to note that the multivariate Bernoulli is *not* a Binomial distribution because each coin can have a different value of π_i . And it is not a Multinomial distribution where $\sum_i \pi_i = 1$ and only one coin can be observed to be a head at one time.

4.1.2 Definition of a Finite Mixture Distribution

Suppose a random variable, X , takes on values in the sample space \mathcal{X} , and that its distribution can be represented by a probability density (or mass) function of the form:

$$p(x) = \rho_1 f_1(x) + \cdots + \rho_k f_k(x) \quad (4.2)$$

where

$$\rho_i > 0 \quad \forall i \quad (4.3)$$

$$\sum_i \rho_i = 1 \quad (4.4)$$

and

$$f_i(x) \geq 0 \quad \forall i \quad (4.5)$$

$$\int_{\mathcal{X}} f_i dx = 1 \quad \forall i \quad (4.6)$$

In such a case, we say that X has a *finite mixture distribution* and that $p(x)$ is a *finite mixture density function*.

The parameters ρ_1, \dots, ρ_k are called the *mixing weights* and $f_1(x), \dots, f_k(x)$ are the *component densities* of the mixture (Titterington et al., 1985).

From these definitions, it is very easy to see that the Star Models are, in fact, finite mixtures of multivariate Bernoulli random variables.

4.1.3 Identifiability of Finite Mixtures of Multivariate Bernoulli Random Variables

In general a finite mixture of multivariate Bernoulli random variables is not identifiable (Teicher, 1961). However, this shortcoming does not prevent such models from being useful in practice. When dealing with classification problems where the researcher only has to identify which component density an observation came from, this type of mixture can be effectively used (Carreira-Perpinan and Renals, 2000). In such a situation, the individual parameters of the multivariate Bernoulli random variables are not of interest. Of course, this presents a challenge in network inference because we are specifically interested in the individual parameters.

4.1.4 Identifiability

A basic requirement for any model is *identifiability*. For Star Models, this means for any two sets of parameters $\{A, \rho\}$ and $\{A^*, \rho^*\}$:

$$\mathbb{P}(G = g|A, \rho) = \mathbb{P}(G = g|A^*, \rho^*) \quad \forall g \implies A = A^*, \rho = \rho^*. \quad (4.7)$$

Unlike the Known Star Model, if both A and ρ are allowed to take arbitrary values, the Star Model is unidentifiable.

This can be demonstrated by the following simple counterexample. Consider a network of size n with A defined as follows: $A_{12} = 1$ and all the other off-diagonal components are 0. Further, let $\rho_i = 1/n$ for all i . Given this, the probability of G has the form:

$$\mathbb{P}(G = g|A, \rho) = \begin{cases} \frac{1}{n} & g = \{1, 1, 0, \dots, 0\}, \\ \frac{1}{n} & \{g : g_1 \neq 1, \sum_i g_i = 1\}, \\ 0 & \text{otherwise.} \end{cases}$$

There are an infinite number of parameters yielding the same distribution, but a simple alternative set of parameters $\{A^*, \rho^*\}$ is given as follows. Let $\rho^* = (0, 2/n, 1/n, 1/n, \dots, 1/n)$. $A_{21}^* = 1/2$ and all the other off-diagonal components of A^* are 0. Obviously, we have $\mathbb{P}(G = g|A, \rho) = \mathbb{P}(G = g|A^*, \rho^*)$. This counterexample demonstrates that Star Models are not identifiable without an additional condition like the symmetry condition presented in Chapter 2.

4.1.5 Necessary and Sufficient Condition for Identifiability

To understand what is required to make a finite mixture of multivariate Bernoulli random variables identifiable, we turn to Yakowitz and Spragins (1968) which provides the following theorem:

Theorem 1. \mathcal{H} is identifiable if and only if \mathcal{F} is a linearly independent set over the field of real numbers, \mathbb{R} .

The symmetric condition introduced in Section 2.4 satisfies this condition in a manner that is not fully understood. A more thorough understanding of this theorem could lead to a less restrictive condition on the Star Models.

4.2. General Grouping Model

Throughout this dissertation, we have shown how the relationships between nodes could be inferred from grouped data by limiting the set of possible subgraphs generating the observed group. Furthermore, we showed how the work of Rabbat et al. (2006) places a different restriction on the subgraphs than Star Models.

A natural extension of this work is to relax the restrictions on the set of possible graphs. One way to do this is to simply let the observed group, $G^{(t)}$, be the result of any connected subgraph on the nodes in $G^{(t)}$.

Such a model would allow a process of group formation which we call the *General Grouping Model* (GGM). Under this model of group formation, we assume that there is a latent symmetric weighted adjacency matrix, A , where A_{ij} is the probability that the relationship between nodes i and j is active for sample t .

We attempt to apply the Expectation Maximization algorithm to optimize the likelihood function. For large groups, the E-step is difficult to calculate exactly, but could be approached using Monte Carlo Expectation Maximization (MCEM).

4.3. Enumeration of Connected Graphs

In the previous section, we would need to efficiently estimate the probability that a group was connected. One way to approach this would be to develop a technique to enumerate all connected graphs.

Here, we introduce a simple method for uniquely identifying every possible network on an arbitrary set of nodes. Once this indexing system is established, we show how some properties of the numbering system lead to a very easy technique for finding connected graphs.

This work is inspired by *An Atlas of Graphs* (Read and Wilson, 2004). The authors of this curious work present pictures of over 10,000 graphs along with tables giving the number of graphs with a certain property and tables of parameters associated with many of the pictured graphs.

One drawback of the Atlas is the numbering system that it uses for graphs. To begin, the Atlas categorizes graphs based on different properties. There are chapters containing trees, regular graphs, planar graphs, etc. In each chapter, graphs are given a prefix to indicate the property that is the focus of the chapter. For instance, unicyclic graphs have

the prefix “U-” while 3-connected plane graphs have the prefix “Tc-”. Within each chapter, graphs are numbered sequentially based on the number of vertices in the graph.

The first shortcoming of this numbering system is that there are multiple graphs which show up in different chapters with different numbers. For instance, G_{18} , P_4 , and C_1 all have the same adjacency matrix.

The second drawback is that it is impossible to draw a graph given its graph number. Consider the tree T_{168} ; all we know from the graph number is that the graph is a tree; we don’t know the number of vertices or the degree. Instead, we have to cross reference this number to a table.

The final drawback is that given an arbitrary adjacency matrix, there is no index to find the picture of that graph.

As an alternative, we propose the following indexing system. Suppose we have the symmetric unweighted adjacency matrix, A , of a simple graph. We adopt the labeling convention that the first row and column of the adjacency matrix are associated with node v_n and the last row and column are associated with v_1 .

We first reduce A to its upper triangular form. For an $n \times n$ adjacency matrix, this will result in a triangular matrix with $\sum_{i=1}^{n-1} i$ terms.

Then we convert the upper triangle into the *binary form* of the network, B , by putting the first row of the matrix into the binary number, then the next, etc.

Table 4.1: Adjacency Matrix for N_{100}

	v_5	v_4	v_3	v_2	v_1
v_5	1	0	0	0	1
v_4	0	1	1	0	0
v_3	0	1	1	1	0
v_2	0	0	1	1	0
v_1	1	0	0	0	1

Table 4.2: Upper Triangular Matrix for N_{100}

	v_5	v_4	v_3	v_2	v_1
v_5		0	0	0	1
v_4			1	0	0
v_3				1	0
v_2					0
v_1					

As an example of this process, consider the adjacency matrix in Table 4.1. We convert this to the upper triangular form shown in Table 4.2. From here it is very simple to see that this has the binary form shown below. Further, since this binary number has a decimal value of 100, we call this network N_{100} .

$$B_{100} = 0001100100$$

In order to make the network index more universal, we allow a single network to be applied to multiple numbers of vertices. For instance, N_1 is simply a network that connects nodes v_1 and v_2 . If we are considering a graph with 15 nodes, N_1 would still connect nodes v_1 and v_2 , but it would not connect any other nodes.

This aspect of our indexing system leads to the following definition:

Definition 4.1. The *basis* of a network is the minimum number of vertices on which the network can be defined.

Proposition 4.2. A network can only be connected in its basis.

Suppose a network, N_x , has basis n but is represented on $n + 1$ nodes. From the binary form, we can easily see that the $n + 1$ row of the adjacency matrix must consist of only zeros. Therefore, node v_{n+1} is not connected to any other nodes in the network and N_x is not connected.

We believe that leveraging this relationship along with several others can lead to an efficient algorithm for enumerating all connected networks on a graph with n nodes. Since

these relationships are built on a unique indexing system, they can be stored dynamically for reuse once the attributes of a given network have been established.

4.4. Additional Work

There are several extensions of the existing work which are possible.

4.4.1 Singleton Free Datasets and Alternative Handling of Singletons

There are some datasets which can never contain a singleton. For example, in an e-mail dataset each observation is an interaction between two individuals, even though additional individuals may be involved and the primary members of the interaction may be ambiguous. We can modify Star Models so that a pair of individuals generates the group, then adds members to the group. This would increase the dimension of ρ to $m = \binom{n}{2}$ pairs and the dimension of A to $m \times n$. This would be a problematic formulation for large groups; however, the Penalized Rho technique could be applied to simplify the model.

Additionally, the implicit assumption in the PRSM that a singleton is acting as a “leader” may have to be modified. As was seen in the Dolphin dataset, since four individuals appeared as singletons, there was no way to penalize all of the elements of ρ to zero. It is possible that the generating mechanism could be modified to account for this behavior differently.

4.4.2 Effect of Removing Individuals

This work can also be used to quantify the effect of changes in a population. For instance, it is common practice to separate disruptive children in classrooms. A legitimate research area would measure whether such actions change the social behavior of the remaining children. An important aspect of this is that we would have to compare old observations ignoring the removed member and observations taken after the individual is removed.

The dataset from the *Dream of the Red Chamber* provides an opportunity to begin this research because the first 80 chapters of the dataset occur before the death of a main

character.

4.4.3 Behavior Transitions Across Time

In the famous karate club example (Zachary, 1977), Zachary was looking at the social interactions of a group as it underwent a social fracture. Similarly, research on captive primates has focused on techniques for integrating existing populations to establish stable social conditions (Schel et al., 2013). There are likely ways to relax the assumption of observation independence to model social behavior through time.

4.4.4 Social Group Evolution

Snijders et al. (2010) has proposed a Markov Chain approach to group formation through time. It would be worthwhile to consider ways in which the Star Models could be incorporated into an evolutionary model.

4.4.5 Non-Member Grouping Effects

There is clear evidence that social behavior can be dependent on some condition other than the members of the population. For example, the size of chimpanzee groups increases when female members of the population are in estrus. Therefore, it might be worthwhile to incorporate additional covariates into the Star Model to further simplify the model.

4.4.6 Measurement Error

Because much of the social activity data is collected by hand, a practical concern for researchers is determining if measurement error is occurring. This can come in different forms. First, one researcher can misidentify an individual member of the population. If there were a training set of data developed by expert researchers, it could be compared against the data collected by new researchers to check for errors.

Another source of work related to measurement error is to deal with the issue of unobserved members. For example, suppose we were using photographs taken by a group of

humans. There is an important member of the group who is unobserved, the person taking the picture. In fact, it is likely that this person is the central node of the observation and exerts considerable influence on the make up of the group.

A final area of measurement error that should be addressed is overlapping groups. As group size increases, it becomes more and more likely that the mode of group formation includes some element of two groups coming together. Addressing this will increase the complexity of the problem; however, this complexity may be mitigated by leveraging the General Grouping Model discussed above.

4.4.7 Population Bounding

At present there are two techniques for bounding a population (Laumann et al., 1989; Wasserman and Faust, 1994), the *realist* and the *nominalist* approach. Under the realist approach, members of groups are identified based on group membership as perceived by the actors themselves (e.g. college men are included in the population based on their membership in a specific fraternity). In the nominalist approach, the population is defined based on the analytical needs of the researcher, even though the researcher may perceive no clear social boundary (e.g. co-authorship datasets).

Research on the effect of measurement errors would likely create insights into the proper way to determine the population boundary and the effect of making mistakes in boundary selection.

4.5. Outreach

Thus far in this work, we have conducted only limited outreach to other researchers who may have an interest in these techniques. In the coming months, I intend to coordinate presentation of this material with the following organizations:

- United States Military Academy Network Science Center
- Naval Postgraduate School Operations Research Department

- Center for Army Analysis
- RAND
- Institute for Defense Analysis
- Army Research Laboratory
- INFORMS

Additionally, we intend to make a submission to the editors of the Encyclopedia of Social Network Analysis and Mining.

Finally, we will look towards developing a website that includes links to our work, datasets, and (possibly) downloads of code.

Appendix A: Notation

Table A.1 provides a list of quantities addressed in this dissertation along with the notation that is used for each quantity. A complete explanation of each quantity is reserved for the section of the dissertation where the quantity is introduced.

Table A.1: Notation Table

Quantity	Notation	Section
Number of Nodes	n	1.1
Full Set of Nodes	V	1.1
Specific Node (Node i)	v_i	1.1
Adjacency Matrix	A	1.1
Element of an Adjacency Matrix	A_{ij}	1.1
Erdős-Renyi Graph with Fixed Probability	$G(n, p)$	1.2
Erdős-Renyi Graph with Fixed Number of Edges	$G(n, m)$	1.2
Generic Probability	p	1.2
Number of Edges	m	1.2
Number of Observed Groups	T	1.4.1
Index of Observed Groups	t	1.4.1
Subset of the Population in Observation t	$V^{(t)}$	1.4.1
Full Set of Observed Groups	G	1.4.1
Single Group in Observation t	$G^{(t)}$	1.4.1
Indicator of node v_i in Observation t	$G_i^{(t)}$	1.4.1
Number of Individuals in Observation t	n^t	1.4.1
Unweighted Undirected Adjacency Matrix Connecting $G^{(t)}$	$G^{(t)}$	1.4.1
Compressed Set of Observed Groups	F	1.4.1
Co-Occurrence Count Matrix	$O\#$	1.5.3
Co-Occurrence Frequency Matrix	O	1.5.3
Threshold for Co-Occurrence Count Matrix	α	1.5.3
Half Weight Index	H	1.5.3
Message from v_i	e_i	1.5.3
Message from v_i to v_j	e_{ij}	1.5.3
Indicator Vector of Center Node of Group $G^{(t)}$	$S^{(t)}$	2.2
Probability that v_i is the center of Group $G^{(t)}$	ρ_i	2.2.1
Number of Parameters	d	2.3
Log-Likelihood	\mathcal{L}	2.3.1
A Specific Example of a Group	g	2.4
Indicator Function of Node v_y	$\gamma(G_y^{(t)})$	2.4.1
Probability Node v_i is Observed	O_i or π_i	2.6.3 or 3.1.1
Tuning Parameter	η	3.1.1
Objective Function	Λ	3.2.2
Number of Non-Zero Elements of ρ	n_o	3.3

Appendix B: Properties of Grouped Data Under Star Models

For any observed G , there are certain descriptive statistics which can be calculated directly from G . These include

- Frequency that node v_i is observed (O_i)
- Frequency that nodes v_i and v_j co-occur (O_{ij})
- Frequency that nodes v_i and v_j co-occur given one of them is observed (H_{ij})
- Average group size (\bar{n}_t)
- Average group size given node v_i is a member of the group ($\bar{n}_{t,i}$)

Clearly, some of these statistics (referred as empirical quantities later in this section) form the basis of the measures typically used to infer network structure. However, it is also possible to calculate the expected values of these statistics (referred as theoretical quantities) from the Symmetric Star Model.

Therefore, in this appendix, we present expressions for theoretical quantities from Star Models. It is worth noting that since the Known Star Model and the Symmetric Star Model have the same generating mechanism, the properties calculated below are not dependent on whether the central node is known or not.

- Probability that v_i is observed

$$\mathbb{P}(G_i^{(t)} = 1|A) = \sum_k \rho_k A_{ki} \quad (\text{B.1})$$

- Probability that v_i and v_j co-occur (O)

$$\mathbb{P}(G_i^{(t)} G_j^{(t)} = 1|A) = \sum_k \rho_k A_{ki} A_{kj} \quad (\text{B.2})$$

- Probability that v_i and v_j co-occur given one of them is observed (HWI)

$$\mathbb{P}(G_i^{(t)}G_j^{(t)} = 1 | G_i^{(t)} + G_j^{(t)} > 0, A) = \frac{2 \sum_k \rho_k A_{ki} A_{kj}}{\sum_k \rho_k A_{ki} + \sum_k \rho_k A_{kj}} \quad (\text{B.3})$$

- Average group size

$$E[n_t | A] = \sum_{i=1}^n E[n_t | S_i^{(t)} = 1, A] E[S_i^{(t)} = 1] = \sum_{i=1}^n \rho_i \sum_{j=1}^n A_{ij} \quad (\text{B.4})$$

- Average group size given v_i is a member of the group

$$\begin{aligned} E[n_t | G_i^{(t)} = 1, A] &= \sum_{k=1}^n \mathbb{P}(S_k^{(t)} = 1 | G_i^{(t)} = 1) E[n_t | S_k^{(t)} = 1] \\ &= \sum_{k=1}^n \frac{\rho_k A_{ki}}{\sum_l \rho_l A_{li}} \sum_{j=1}^n A_{kj} \end{aligned} \quad (\text{B.5})$$

Appendix C: Runtime of Star Models

One aspect of the Star Models that is of particular concern is the time that it takes to estimate the parameters for a dataset. To explore this, we varied the population size and number of observations as shown in Table C.1. These trials ran on a Intel Pentium CPU G2030 at 3.00 GHz with 4.00GB of RAM.

For each population, we produced a latent network structure with $n_o = n$, then generated six observation sets of differing length from that structure. We repeated this process 100 times for each population and report the average time to estimate the parameters.

Table C.1: Average Runtime in Seconds

	$n = 5$		$n = 10$		$n = 20$		$n = 30$	
Parameters	$d = 14$		$d = 54$		$d = 209$		$d = 464$	
Obs	SM	PRSM	SM	PRSM	SM	PRSM	SM	PRSM
200	11.5	118.4	17.0	348.8	28.5	555.6	35.9	929.7
500	11.0	102.8	15.9	351.4	35.6	761.8	70.9	1,869.6
1,000	10.1	104.4	15.1	314.8	39.6	886.0	114.0	3,240.7
2,000	10.3	90.9	13.8	299.5	44.9	1,166.1	184.4	4,968.5
5,000	10.1	84.6	13.5	265.6	58.8	1,277.6	364.3	9,376.0
10,000	10.1	76.4	14.0	261.8	59.0	1,527.9	912.1	23,794.9

The number of parameters, d , increases according to (2.16).

One of the most striking features of Table C.1 is that for datasets where the number of observations is on the order of ten times larger than the number of parameters to be estimated, the runtime appears to decrease. This improvement in runtime is likely caused by a reduction in the number of iterations necessary for convergence.

Bibliography

- Alderson, D. (2008). Catching the network science bug: Insight and opportunity for the operations researcher. *Operations Research*, 56(5):1047–1065.
- Barabasi, A. and Albert, R. (1999). Emergence of scaling in random graphs. *Science*, 286(5439):509512.
- Bejder, L., Fletcher, D., and Brager, S. (1998). A method for testing association patterns of social animals. *Animal Behavior*, 56:719–725.
- Boyd, S. and Vandenberghe, L. (2011). *Convex Optimization*. Cambridge University Press.
- Cairns, S. J. and Schwager, S. J. (1987). A comparison of association indices. *Animal Behavior*, 35.
- Carreira-Perpinan, M. A. and Renals, S. (2000). Practical identifiability of finite mixtures of multivariate bernoulli distributions. *Neural Computation*, 12:141–152.
- Choudhury, M., M., W. A., Hofman, J. M., and Watts, D. J. (2010). Inferring relevant social networks from interpersonal communication. *International World Wide Web Conference Committee*.
- Dai, B., Ding, S., and Wahba, G. (2013). Multivariate bernoulli distribution. 19(4).
- Dice, L. R. (1945). Measures of the amount of ecological association between species. *Ecology*, 26:297–302.
- Erdos, P. and Renyi, A. (1960). On the evolution of random graphs. 5.

- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456).
- Freeman, L. C., White, D. R., and Romney, A. K. (1989). *Research Methods in Social Network Analysis*. George Mason University Press.
- Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99:7821–7826.
- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
- Hawkes, D. (1974). *The Story of the Stone, or The Dream of the Red Chamber, Vol. 1: The Golden Days*. Penguin Classics.
- Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. Springer.
- Laumann, E. O., Marsden, P. V., and Prensky, D. (1989). The boundary specification problem in network analysis. *Research Methods in Social Network Analysis*.
- Lusseau, D., Schneider, K., Boisseau, O. J., Haasse, P., Slooten, E., and Dawson, S. M. (2003). The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations: Can geographic isolation explain this unique trait? *Behavioral Ecology and Sociobiology*, pages 396–405.
- MacCarron, P. and Kenna, R. (2013). Viking sagas: Six degrees of icelandic separation-social networks from the viking era. *Significance*, pages 12–17.
- McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. John Wiley and Sons, Inc.

- Michalski, R., Palus, S., and Kazienko, P. (2014). Matching organizational structure and social network extracted from email communication. *Encyclopedia of Social Network Analysis and Mining*.
- Moreno, J. L. (1934). *Who Shall Survive? A New Approach to the Problem of Human Interactions*. Nervous and Mental Disease Publishing Co.
- Newman, M. E. J. (2011). *Networks: An Introduction*. Oxford University Press.
- PsycNet, A. (2012). Review of the book *Who Shall Survive*. *PsycINFO Database*.
- Rabbat, M., Figueiredo, M., and Nowak, R. (2006). Network inference from co-occurrences.
- Read, R. C. and Wilson, R. J. (2004). *An Atlas of Graphs*. Clarendon Press.
- Schel, A. M., Rawlings, B., Claidiere, N., Wilke, C., Wathan, J., Richardson, J., Pearson, S., Herrelko, E., Whiten, A., and Slocombe, K. (2013). Network analysis of social changes in a captive chimpanzee community following the successful integration of two adult groups. *American Journal of Primatology*, 75:254–266.
- Snijders, T. A. B., Koskinen, J., and Schweinberger, M. (2010). Maximum likelihood estimation for social network dynamics. *The Annals of Applied Statistics*, 4:567–588.
- Teicher, H. (1961). Identifiability of mixtures. *The Annals of Mathematical Statistics*, 32(1):244–248.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288.
- Titterton, D. M., Smith, A. F. M., and Markov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley and Sons, Inc.
- Voelkl, B., Kasper, C., and Schwab, C. (2011). Network measures for dyadic interactions: Stability and reliability. *American Journal of Primatology*, 73:731–740.

Wasserman, S. and Faust, C. (1994). *Social Network Analysis: Methods and Applications*.
Cambridge University Press.

Yakowitz, S. J. and Spragins, J. D. (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39:209–214.

Zachary, W. W. (1977). An information flow model for conflicts and fission in small groups. *Journal of Anthropological Research*, 33:452–473.

Curriculum Vitae

Charles Weko is a Lieutenant Colonel in the United States Army where he is responsible for \$12.7 billion in pay and allowances for 530,000 Army Reserve and National Guard Soldiers. He received his Bachelor of Science in Electrical Engineering from Rose-Hulman Institute of Technology in 1995. He was certified as a Black Belt in Six Sigma in 2005. He received a Master of Business Administration degree from Grantham University in 2007 and a Master of Science in Operations Research from Naval Postgraduate School in 2009.

His work experience includes working as a Process Engineer for Solo Cup Company, command of the 298th Maintenance Company, and manager of the Fort Dix Clothing Issuing Facility. As an Army staff officer, he has worked on the Navy's integration of Unmanned Systems, Army Reserve Suicide Prevention, modeling the retrograde of equipment from Iraq, and analysis of Blue Force Tracker Data.