UNIVERSITY OF CALIFORNIA

Los Angeles

Some Models in Relational Systems

A dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Philosophy
in Statistics

by

Aaron Danielson

2018

ABSTRACT OF THE DISSERTATION

Some Models in Relational Systems

by

Aaron Danielson
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2018
Professor Mark Stephen Handcock, Chair

This dissertation contains two distinct parts. The first part presents a theory for the observation process in surveys of human populations. Focus resides on the effect of survey instrument structure on the observed responses. If the effect of instrument structure on a survey question can be estimated and certain exchangeability conditions hold, sampling from the set of all possible survey instrument structures provides a way to assess the observed sample. This part then introduces probability distributions that aggregate over the set of possible instrument structures thereby defining a way to view data without an instrument structure. Methods to incorporate instrument structure are demonstrated in several examples including a survey of a large organization. This dataset requires new techniques to account for dependence between units in the population. The second part introduces a method to estimate latent networks of interacting nodes when a group-level response variable is observed. In large populations, the number of possible ties between nodes grows large. The models named *CoordiNet* and *TriadNet* place a prior over the dyads included in the model for the group-level response. Doing so induces parsimony in the estimated social structures. This method is applied to data from professional sports.

The dissertation of Aaron Danielson is approved.

Barbara S Lawrence

Frederic R Paik Schoenberg

Arash Ali Amini

Erin K Hartman

Mark Stephen Handcock, Committee Chair

University of California, Los Angeles

2018

DEDICATION

To Cara Tambellini for her unwavering love and support.

Contents

List of Figures

VITA

As an undergraduate Aaron majored in philosophy at Northwestern University. He continued at the University of Chicago to earn a Master's of Arts in the Humanities with a concentration in philosophy. After one year of teaching English at a high school outside of Nagasaki, Japan, Aaron returned to the University of Chicago to complete a Master's of Public Policy. From here, he moved to New York where he earned a Master's of Arts in economics. While attending UCLA, he worked for three seasons as the Assistant Director of Analytics at the Los Angeles Lakers.

Part I: Observation Effects in Survey Sampling

Chapter 1

Observation Effects in Survey Response

The first part of this dissertation addresses a type of incongruence between the phenomenon of scientific interest and the data ultimately recorded in a survey. Contingencies thought to be unrelated to the system under study can influence the survey process. For example, respondents interact with features of the survey instrument structure such as the order of responses and the order in which questions are asked. These interactions can alter the probability distributions over their responses. The observation of a phenomenon interferes with the system of scientific interest. If the interaction has a large effect on a response variable of interest, then the observation process biases the sampled data. If the probability distribution of the response variable can be expressed as a function of features of the observation process, then probabilities can be computed for counterfactual values of the observation process thereby generating the hypothetical population of probability mass functions associated with every value the features take. If survey designs assign different features, such as response or question orders, randomly, then causal interpretations of the effect of observational features on a response are warranted.

1.1 Observation vs. Data-Generation

Observation is the process by which a phenomenon becomes recorded data. Dependence between the random variables comprising these processes can bias the observed data according to the effect of features of observation variables. Let $\mathcal{Z}$ be a set of random variables associated with the observation process, possibly connected according to graph structure $G_{\mathcal{Z}}$. Observation variables describe features of the process by which data is collected. This set might contain features associated with response order, the time allotted to complete the survey, sampling mechanisms and nonresponse. Let $\mathcal{X}$ be a set of random variables associated with the data-generating mechanism, possibly connected according to graph structure $G_{\mathcal{X}}$. This set contains the variables thought to influence the value of the response if no observation were made such as attributes of the respondent and responses to other questions in the survey.

Under ideal scientific circumstances the processes are completely independent. That is, for all $X \in \mathcal{X}$ and $Z \in \mathcal{Z}$

$$p(X, Z) = p(X)p(Z).$$

Under these conditions realizations of the variables in the observation process do not influence the distributions of the random variables in the data-generating process. Hence the data generating process can be estimated without the random variables contained in $\mathcal{Z}$. The observation process is ignorable. But when variables in $\mathcal{Z}$ interact with or cause variables in $\mathcal{X}$, the observation and data-generating process must be modeled jointly.

As documented in Chapter 2, attributes of the survey instrument 's structure interact with the sampled individual's response process. Surveys can ask about two different types of information: data that could be externally verified and data that is inherently subjective. For example, suppose a survey asks two questions. How many computers do you own? And, how do you feel right now on a scale of one to five? The first question asks about a piece of information that could be theoretically verified without consulting the respondent. But, the latter question cannot. Interaction between of the observation and data-generating processes presents an interesting challenge for subjective questions. Suppose the order of choice varies across designs, and different designs of the survey are randomly assigned to respondents. If the presentation order of the responses interacts with the response process, the observation process is not ignorable. In the case of the number of computers, external validation is possible. But there is no straightforward way to validate the response. Responses to similar questions and actions can serve as proxies.

Because the subjective information cannot be verified externally, experiments are needed to test the effect of variation in the survey instrument structure on some result. Removing the dependence between the estimator and the instrument structure leaves a probability distribution without the artifacts of the observation process.

1.1.1 The Observation Process in Surveys

Observation is a process: the primal phenomenon must be organized into a logico-semantic representation such as a sequence of ordered responses indicating the respondent's level of interest in some topic. The representation must be measured via an instrument, the instrument must have a specific structure and sampled individuals must interact with the instrument. As depicted in the following flow chart, observation comprises the measurement and recording of a

representation of a phenomenon as it is perceived by the survey respondents.

$$\begin{aligned}\text{Phenomenon As-It-Exists} &\rightarrow \text{Phenomenon As-Can-Be-Represented}\\ &\rightarrow \text{Representation As-Can-Be-Surveyed}\\ &\rightarrow \text{Representation As-Can-Be-Perceived}\\ &\rightarrow \text{Representation As-Surveyed-In-Fact.}\end{aligned}$$

Realizations of each step in the process determine the observed data. The respondent's perceptions of the underlying phenomenon and mental state interact with the structure of the survey. Like the observer effect in physics, the features of the survey used to elicit responses can change the system under study. For example, unfamiliar or complicated survey designs may increase uncertainty in the mind of the respondent just as a thermometer may warm the liquid it is used to measure. But, unlike bodies of water or quantum particles, the perceptions of the humans under measurement also determine the responses. The respondent interprets the survey prompts in a way that need not match the interpretation intended by the survey designer. Hence all survey data depends on the representation of the phenomenon of interest, the instrument used to measure it and the perceptions of the sampled individuals. But, the dependence does not imply that variation across representations, instruments, mental states and perceptions is meaningful in a scientific context. When it is, the context of the data creation should be modeled explicitly.

New methods are needed to assess whether the realizations associated with the survey context are particularly extreme and to understand the range of possible outcomes under counterfactual circumstances. Not only does this help to get less obscured views of the scientific object of interest, it can reveal what type of designs can mitigate the observation effect. The results of explicit models for survey context can also reveal insight into the cognitive processes that alter responses. If this can be done, context-adjusted samples can be produced that might constitute better depictions of the real-world phenomenon scientists actually wish to study. The remainder of this section introduces a general graphical model for the survey context.

1.1.2 A Graphical Model for Survey Response

Let the observation process be the collection of random variables $\mathcal{O} = \{Z, E\}$ with dependence graph $G_{\mathcal{O}}$ and $\mathcal{X} = \{X, Q\}$ with dependence graph $G_{\mathcal{O}}$. If there is dependence between variables in two sets, then the observation process may intrude into the data-generating process. Aspects of the survey data creation process in general, and the network survey data creation process, in particular, can be depicted as a graphical model. Let Γ denote the representation chosen for the phenomenon of interest. The scientific community in which observation is embedded dictates a convention for the

phenomenon's representation. The convention reflects current theoretical understanding of the system, and practical issues associated with feasible measurement devices and tractability of data analysis. Let Z be a random variable describing features of the instrument used to collect the information required by a representation. Instruments include tools such as surveys and sensors (both biological and mechanical); features of instruments include their material composition and design.

Since surveys are completed by people, they contain artifactual information about the respondent. The way people answer survey questions depends on the state of their cognitive processes and their perceptions of themselves and the external world. The former represents the way the brain functions at the time of the survey; the latter describes their internal conceptions of themselves and their relations. For example, respondents distracted by the memory of something that occurred earlier in the morning may answer survey prompts differently than they would under other circumstances. The survey responses are therefore imbued with cognitive artifact. To given an example of the second concept, what constitutes "Interest" in a topic may differ between people. A person's available and relevant social context alters perception of the phenomenon of interest.

[55] defines the notion of psychological entropy to be "the experience of conflicting perceptual and behavioral affordances." The psychological discomfort felt by respondents is associated with a heightened "degree of perceptual and behavioral ambiguity within a situation." Represent the entropy of a respondent's mind by the random variable E. Different values of E alter the probability distribution over responses in a survey setting. For example, suppose objectively verifiable answers are available for the questions posed by a survey. When a higher entropy state is realized due to some event independent of the survey process, the subject responds with increased levels of inaccuracy. In cases where no such verifiable truth exists, such as how one person feels about another, the probability distribution exhibits higher entropy thereby becoming less predictable.

Observation of the phenomenon of scientific interest may require that respondents perceive the questions in a way that is similar to one another, and to the scientific community. When perceptions of the content of the questions differ, and the difference in understanding is sufficiently large to change a sampled unit's response, models of the data-generating process suffer from omitted variable bias. If there is dependence between variables in two sets, then the observation process may intrude into the data-generating process.

The level of cognitive entropy, E, determines the amount of variance in their reports. If cognitive entropy correlates with ambiguity and novelty, then features of the survey instrument can affect the probability measure over E. A novel or counterintuitive survey structure increases the probability of less reliable responses (if they could be verified).

Assume E and Γ are independent given Z. The way in which a field of inquiry represents phenomena should not depend on the particular state of E at the time and place of measurement. Moreover, given the features of an instrument, Z, cognitive entropy does not depend on the representation. But, as previously mentioned, under some circumstances the constitution of the instrument may be associated with E. Complex or unorthodox instrument designs induce greater ambiguity in the respondent's mind.

Scientists set the survey design Z *ex ante*; that is, they select an instrument to fit the needs of the data creation environment. The features of the survey affect the resulting data directly by framing the choice environment of the respondent and indirectly by altering the respondent's cognitive entropy. Some instrument designs may couple with certain representations to have large effects on the resultant data; others can have relatively little effect. Lastly, the instrument used to measure a phenomenon certainly depends on its representation. By setting a representation scientists identify a subset of instruments concomitant with the representation. A possible theoretical framework for observation and data-generating processes in surveys appears in figure 1.1. If a variables belongs to the observation process its node is colored green; if a node belongs to the data-generating process, it is shaded blue. The left panel displays a model in which the cognitive entropy and instrument are marginally independent. The right panels depicts a slightly more complicated scenario; it assumes the instrument structure causes changes in the probability distribution over cognitive entropy. The third panel presents in which the variables associated with the data-generating process cause changes to the probability distribution over cognitive entropy. Suppose the model in the left panel holds and the representation for the phenomenon is held fixed, $\Gamma = \gamma_0$, where γ_0 denotes a conventional representation. Then the probability of this relational system is

$$p(Q, Z, E|X = z, \Gamma = \gamma_0) = p(Q|Z, E, X)p(E)p(Z|\Gamma = \gamma_0).$$

Similarly, suppose Z is fixed $Z = z_0$ so that z_0 are the features associated with a conventional instrument to observe the phenomenon given the conventional representation $\Gamma = \gamma_0$. Then the probability of the observed data given the observation process is just

$$p(Q|Z = z_0, \Gamma = \gamma_0) = \sum_{e \in \mathcal{E}} p(Q|Z = z_0, E = e, \Gamma = \gamma_0)p(E|Z = z_0, \Gamma = \gamma_0),$$

Since E cannot be fixed, it is marginalized out of the expression. But, if a model for E is not tractable, then the survey responses conflate the uncertainty associated with the phenomenon of interest with the uncertainty associated with cognitive entropy of each of the respondents among other possible confounders.

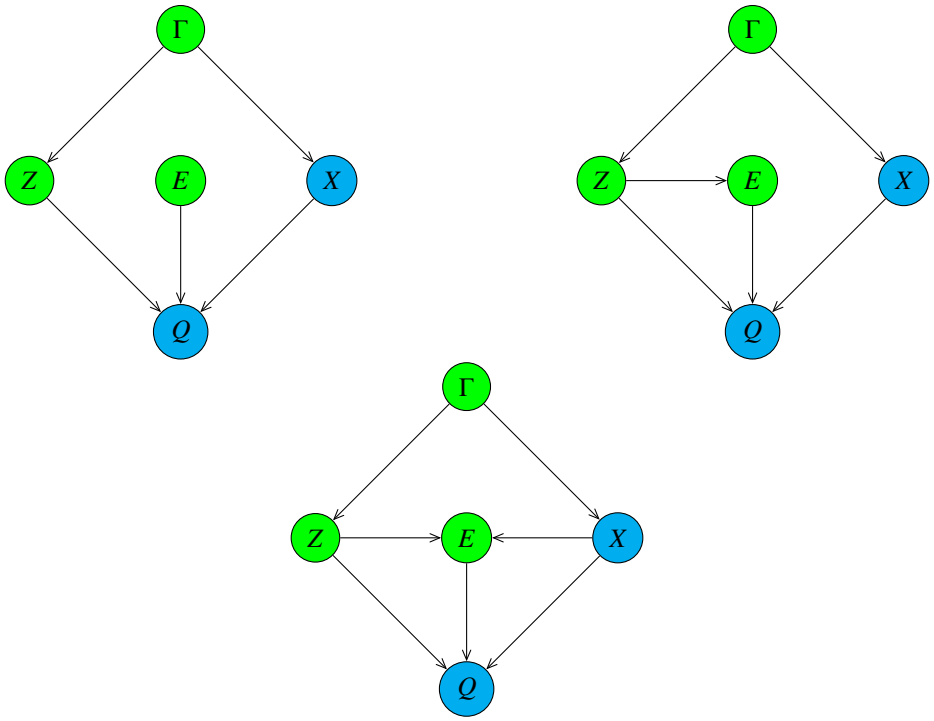

Figure 1.1: Observation and Data-Generating Processes as a Graphical Model.

1.2 Survey Data without an Instrument Structure

How would individuals respond to questions if data were collected without a survey instrument? Consider a survey of N subjects asking them to choose their favorite word from a list of options. Repeat the experiment $|\mathcal{Z}|$ times where each $z \in \mathcal{Z}$ is a distinct ordering of the options. Further suppose each z is sampled L times. The order in which the various instrument structures were presented to respondents would influence their choices. Suspending disbelief for a moment, suppose this is not the case and define

$$p(Q^z = q) = p(Q = q|Z = z) = \frac{1}{L}\sum_{l=1}^{L} \mathbb{I}\{Q_l = q|Z = z\} \tag{1.2.1}$$

to be the probability of observing data Q with instrument structure z. The family of the probability mass function $\{p(Q^z = q)\}_{z\in\mathcal{Z}}$ enables scientists to ask questions whether the differences in mass functions is significant. These structures alter the data by changing the state of cognitive processes in the mind of the respondent. If such instrument structures exist, they may reveal insights as to how respondents formulate responses to network surveys. Moreover, the outlying data may provide different information about the object of interest. One way to compute this is to consider the average KL-divergence

$$\frac{1}{|\mathcal{Z}|}\sum_{z'\in\mathcal{Z}} KL(p(Q^z)\|p(Q^{z'})) = KL\left(p(Q^z)\middle\|\prod_{z'\in\mathcal{Z}} p(Q^{z'})^{\frac{1}{|\mathcal{Z}|}}\right). \tag{1.2.2}$$

The right side of the term on the right side is the geometric mean of the probabilities $\left\{p(Q^{z'} = y)\right\}_{z' \in \mathcal{Z}}$ meaning that the average KL-divergence is a measure of distance from the probability of agreement between all z'. As an alternative define the expected probability that $Q = q$ over the instrument structure. Since each z appears L times, they are equally likely. The expected probability that $Q = q$ is therefore

$$p(Q^{\bar{z}}) = p^*(Q = q) = \frac{1}{|\mathcal{Z}|L} \sum_{z \in \mathcal{Z}} \sum_{l=1}^{L} \mathbb{I}\{Q_l = q | Z = z\}, \tag{1.2.3}$$

and the KL-divergence between $p(Q^z)$ and the expected probability is

$$KL(p(Q^z) || p(Q^{\bar{z}})) = \sum_{q \in Q} p(Q^z = q) \log\left(\frac{p(Q^{\bar{z}} = q)}{p(Q^z = q)}\right). \tag{1.2.4}$$

These two methods are not equal. Chapter 3 explores differences in the two means and their relation to the KL divergence measures.

Of course, this thought experiment is not possible to perform. The order in which replications are introduced to the respondent affects the downstream observations. More likely, the experiment could be run a single time in which the possible survey structures are randomly assigned to the respondents, or in which a single survey instrument structure is assigned to every respondent.

1.2.1 Exchangeability

Suppose only a sample of the possible instrument structures were implemented in a survey. In order to compute probability mass functions of survey responses conditioned on counterfactual instrument structures, the data must be exchangeable. The potential outcome Q^z is distributed independently from the value of Z used to observe the data (see [51]). Symbolically, for all $z \in \mathcal{Z}$, $Q^z \coprod Z$, so that

$$p(Q^z = q | Z = z', X = x) = p(Q^z = q | Z = z'', X = x)$$

for all $z', z'' \in \mathcal{Z}$. Then if the instrument structures used in the experiment were randomly sampled the probability mass functions can be interpreted as

$$p\left(Q^z = q | X = x\right) = p\left(Q^z = q | \operatorname{do}(Z = z), X = x\right).$$

Since Z is exchangeable, the effect of Z is causally identified if the positivity assumption holds. That is, there must be a positive probability of receiving each treatment. Thus Q^z can be interpreted as the counterfactual network whenever $z \neq z_{\text{obs}}$. As before, averaging over the counterfactual probabilities provides an estimate for the expected probability distribution over Q. Then the probability distribution for Q adjusted for the survey instrument structure is given by

$$p^*(Q = q) = \sum_{z \in \mathcal{Z}} p(Q = q | Z = z)\pi(Z = z) = \frac{1}{|\mathcal{Z}|} \sum_{z \in \mathcal{Z}} p(Q = q | Z = z). \tag{1.2.5}$$

When the survey instrument structure is not randomly assigned to subjects due to dependence between the observation and data-generating processes, methods from causal analysis identify instrument effects. For example, if a variable X associated with the data-generating process is thought to cause Z, then

$$p^*(Q = q) = \sum_{z \in \mathcal{Z}} p(Q = q | (Z = z))\pi(Z = z) = \frac{1}{|\mathcal{Z}|} \sum_{z \in \mathcal{Z}} \sum_{x} p(Q = q | Z = z, X = x). \tag{1.2.6}$$

Chapter 3 explores alternative aggregate measures such as mass functions proportional to the power means and the Wasserstein barycenter.

1.2.2 Aggregate Measures for the Survey Process

Throughout this section assume $\Gamma = \gamma_0$ where γ_0 is the conventional representation of the phenomenon of interest. Suppose scientists survey Q in every way it could be observed, with every possible instrument structure. The procedure generates a family of probability measures

$$\{p(Q = q | E = e, X = x, Z = z)\}_{z \in \mathcal{Z}}. \tag{1.2.7}$$

Each member provides a different lens through which humans view the primal phenomenon of interest. Average over the different instruments and define an aggregate probability measure over the collection. Assuming the space $\mathcal{Z}$ is finite, define

$$\begin{aligned} p^*(Q = q | E = e, X = x, \Gamma = \gamma_0) &= \sum_{z \in \mathcal{Z}} p(Q = q, Z = z | E = e, X = x, \Gamma = \gamma_0) \\ &= \sum_{z \in \mathcal{Z}} p(Q = q | E = e, X = x, \Gamma = \gamma_0, Z = z)\pi(Z = z | \Gamma = \gamma_0). \end{aligned} \tag{1.2.8}$$

The probability mass function $\pi(Z = z | \Gamma = \gamma_0)$ places equal prior weight on each distinct instrument consistent with the representation γ_0. The prior does not represent what design is likely to be used in the data created process, rather it assigns probability to each of the ways the data could have been observed.

Note that E depends on Z, but changes in E due to changes in Z are not included. This model conflates changes in the level of cognitive entropy with changes in uncertainty due to Q under different Z. It does not account for different levels of cognitive entropy determined by different designs. Hence if cognitive entropy is thought to have a negligible effect on the resulting data no matter the instrument structure, then 1.2.8 is a sensible choice. However, if a model for E given Z is available, the probability

$$p^*(Q = y, E = e|\Gamma = \gamma_0) = \sum_{z \in \mathcal{Z}} p(Q = q|E = e, \Gamma = \gamma_0, Z = z)p(E = e|Z = z)\pi(Z = z|\Gamma = \gamma_0). \tag{1.2.9}$$

can be used in its place. Here two the two sources of variation under different Z are modeled explicitly, but the uncertainty in Q and E are conflated. If the cognitive entropy plays a large role in the behavior of the respondents, it should be controlled in a similar manner to Z. Supposing E is finite, define the refined probability to be

$$p^*(Q = q|R = r_c) = \sum_{e \in E} \sum_{z \in \mathcal{Z}} p(Q = q|E = e, Z = z)p(E = e|Z = z)\pi(Z = z|R = r_c). \tag{1.2.10}$$

This is the probability of observing realizations of Q absent a particular survey instrument structure and cognitive entropy. Although the focus of this paper remains 1.2.8, methods to estimate this alternative probability measure are discussed in 4.

In typical situations it is not feasible to measure with every possible instrument. By modeling the data as a function of the instrument's features, counterfactual data can be simulated and compared to the data created by instruments actually used. This important step relies on exchangeability assumptions inherited from the causal inference tradition. If statistics derived from the observed data appear to be extreme relative to statistics derived from the counterfactual collection, then the particular instrument may have created data that does not reflect the distribution of data sets that could be created by the set of all instruments. One must assess whether inferences made from such data sets reflect properties of the scientific phenomenon of interest.

If use of these methods demonstrates that features of the observation process affect survey results to large degrees, then it is possible observation effects exist in many social systems. The process at work in surveys represents a particular example of a larger phenomenon.

Chapter 2

Accounting for Observation Effects in Subjective Surveys

2.1 Introduction

Surveys of subjective information differ from surveys of objective information in an important way: reports about attitudes, preferences and impressions of the world do not refer to an unambiguous ground truth. If a respondent is asked to report income, a birthdate or counts of some addictive behavior then, theoretically, these responses can be compared to correct values. Inherently subjective surveys are therefore impossible to validate relative to an external standard. Thus, a survey respondent cannot report subjective impressions incorrectly. It is well known that features of the survey process such as the survey instrument's structure affect the responses of sampled individuals. When the responses can be modeled as a function of instrument structure, the population of possible survey responses can be generated. This paper introduces a method to evaluate the extent to which a survey instrument structure produces samples very different from the overall population. Then a method to aggregate over the population of counterfactual responses to produce a distribution of responses free from the influence of a particular instrument structure is presented.

The structure of this paper proceeds as follows. Section 2.2 argues that surveys of subjective information are especially vulnerable to features of the survey instrument structure. Examples from existing survey data demonstrate response and question order effects. These previous studies illustrate how reports of subjective impressions vary with features of instrument structure. Section 2.3 introduces a general framework for surveys of subjective information. This approach enables the researcher to assess the impact of the survey instrument structure relative to a population of possible structures. Finally, a method to estimate a probability mass function aggregating over all possible instrument

structures is introduced. Section 2.4 presents a method to model the entire survey jointly. Counterfactual methods can be applied to aggregate over possible assignments of instrument structure and assess the extremity of the survey instrument assigned in the observed data. Section 2.5 addresses incorporating time limits on a survey as an additional dimension of instrument structure.

2.2 Survey Instrument Structure and the Observation Process

Many random variables affecting survey data have been identified within the total survey error paradigm. [45] reviews its history. Within this framework a taxonomy of errors is enumerated. Error associated with survey instrument structure nests within this larger conceptual structure. The authors write that, "Instead of concentrating on error models that have nice estimation properties, we should focus on error models that better reflect the phenomenon being modeled." This is the approach taken in this paper. Rather than estimating error, interested is shifted towards a better understanding of the data generating and data observation processes. The concept of an error implies a systematic deviation from a true value. When surveys ask questions about objective information, theoretically verifiable facts, respondents can answer correctly or incorrectly. For example, if asked to report the number of cups of coffee purchased per week, researchers could compare the survey responses to an electronic purchase history. Objective survey questions refer to answers that could be known with the use of the appropriate archival technologies. But, when surveys inquire about subjective information, there exists no externally verifiable ground truth to which the response can be compared.

[94] posits a theory supporting the present thesis that subjective and objective survey questions differ in a fundamental way. Under their model respondents, "carry around in their heads a mix of only partially consistent ideas and considerations. When questioned, they call to mind a sample of these ideas, including an oversample of ideas made salient by the questionnaire and other recent events, and use them to choose among the options offered. But their choices do not, in most cases, reflect anything that can be described as true attitudes; rather, they reflect the thoughts that are most accessible in memory at the moment of response." The authors contend individuals often have conflicting opinions on an array of topics. Survey instrument structure biases cognitive processes towards a subset of the respondents' ideas. They argue systematic variation arises from the artifactual effects of the survey instrument structure on responses. The present argument pushes strengthens these claims. Questions about subjective perceptions cannot have error, rather the subjective report exists within the context in which it is observed. Changes in the context of observation beget changes in the survey responses. Since responses can only be collected within a survey context, the dependence between context and response is unavoidable.

2.2.1 Examples of Response Effects

This section presents several examples of response effects owing to features of instrument structure. [30] states, "A survey instrument is a tool for consistently implementing a scientific protocol for obtaining data from respondents. For most social and behavioral surveys, the instrument involves a questionnaire that provides a script for presenting a standard set of questions and response options." The particular form these questions and response options take comprise the survey instrument structure. When the artifacts of the instrument structure interact with respondents' cognitive processes, omission of the association between features of the instrument structure and the surveyed data leads to a misspecified model of the data generating process.

In the following examples, the presentation order of questions or responses varies. Then the researcher can perform statistical tests to determine the degree of difference between the responses under different survey structures. The focus of this paper is the effect of structure, not content. Structure involves *how* humans observe a phenomenon while content involves *what* humans observe. Changing the wording of a question changes the scientific object of interest; the surveys elicits different information from respondents. But, changing the presentation order of responses elicits information about the same content in a different way. Response changes persist across types of surveys whether they are conducted face-to-face, by mail or electronically. This is important to note as simply changing the survey method does not ensure that response effects do not bias the resulting data. Response effects are ubiquitous in surveys of inherently subjective information.

[73] reports several instances in which contingencies of sampling, such as the order in which questions appear, influence the behavior of survey respondents. In September of 1997 pollsters interviewed 1002 potential voters regarding their impressions about the honesty and trustworthiness of Bill Clinton and Al Gore. Respondents were asked whether each person is "honest and trustworthy." When they were asked about Clinton first 50% said yes to Clinton and 60% said yes to Gore. When asked in the opposite order, 57% said yes to Clinton and 68% to Gore. The probability of observing this difference in support for Clinton given the two instrument structures is approximately 0.027 while the corresponding number for Gore is approximately 0.010. Other studies reported in the paper also produce large discrepancies when the survey instrument structure varies. The company that administered the poll averaged the results across the two survey instrument structures so that

$$\Pr\{i \text{ affirms } x\} = \frac{\Pr\{i \text{ affirms } x | Z = z_1\} + \Pr\{i \text{ affirms } x | Z = z_2\}}{2}$$

with equal weights since half of the sample responded to each structure. As the survey elicited subjective opinions, a ground truth is not a coherent concept.

Order	Form A	Form B
1	More difficult to obtain	Easier to obtain
2	Stay the same	Stay the same
3	Easier to obtain	More difficult to obtain

Table 2.1: Actual Survey Instrument Structures Implemented in Mail Survey

[11] provides another example as to how survey design can alter the process by which the object of scientific interest becomes data. The authors present results of experiments on taxonomic and componential organization using balanced incomplete block designs. Test subjects were asked to identify the most dissimilar word contained in each of the triads presented to them. Dimension reduction via nonmetric multidimensional scaling was used with estimates of the similarity between words as inputs. The second of the authors' two goals for the paper is to "discuss the structural problems which may arise with the use of" balanced incomplete block designs. They write, "Clearly, the similarity measure for any two words depends upon context, i.e. in the case of the triads test the three words. For this reason, it is important that triads tests be done on words which are from the same semantic domain, and which are at the same level of contrast in that domain." They find certain randomizations had a real effect on the resulting analysis. For example the triad $\{\text{corn}, \text{peas}, \text{lima beans}\}$ leads to what the authors call a structural problem with the design. The three form a complete triad in the beans that there are strong dyadic associations between each. If another vegetable were chosen from a different domain in place of one of the three, this object would be judged as most distinct. Hence the specific random assignment of triads altered the observation process.

As reported in [3], response effects appear in mail surveys. Under these conditions respondents are free to complete the survey in any order. Moreover, they may reread questions and change answers after taking time to deliberate. The authors note mild question order and response order effects. Most notably, subjects were asked whether divorce should be easier to obtain, more difficult to obtain, or stay the same as it is now. Choices were presented in ascending and descending orders according to 2.1. The paper reports 57.9% of respondents chose the "more difficult" option when the option to make divorce more difficult to obtain was presented before the option to make divorce easier to obtain, while 47.4% chose this option under the alternative design. To generate an estimate of the probability distribution over beliefs the data could be pooled so that

$$p(Q = q) = p(Q = q|A)p(A) + p(Q = q|B)p(B)$$

where $p(A)$ and $p(B) = 1 - p(A)$ are the marginal probabilities of the assigned instrument structure. Since the two designs were assigned randomly, the estimate easier to obtain, while 47.4% chose this option under the alternative design. To generate an estimate of the probability distribution over beliefs the data could be pooled so that

$$p^*(Q = q) = \frac{p(Q = q|A) + p(Q = q|B)}{2}$$

is the probability distribution unaffected by the specific realization of random assignment.

[63] investigates response-order effects in the General Social Survey of 1984. These are defined as "changes in the answers to closed-ended survey questions produced by varying the order in which response options are presented." Respondents were asked about 13 qualities regarding children. After seeing the list, respondents were asked to name most three desirable qualities, and from those three, to choose one as the most desirable. Analogously, respondents were asked for the three least important and the least important among those three. They conclude responses to the questions about child qualities are partially determined by the order of possible responses. In particular, placing a quality among the first three choices increased the probability that it was chosen as one of the three most desirable traits. For example when the trait 'Manners' was placed first rather than last, it was chosen as a top three quality in 16.3% more surveys. When the trait 'Honest' was placed second rather than 12$^{\text{th}}$, it was chosen as a top three quality in 17.3% more surveys.

The two designs were implemented in the study are depicted in 2.2. Since the order was reversed, the qualities at the beginning and the end have significant variation across the two designs. But, unlike the Clinton-Gore example, the set of possible survey instrument structures is much larger than these two. There are 13! ways to list the options for survey respondents. Averaging over the two observed structures models

$$p(Y_i = k|Z = z_1 \text{ or } Z = z_2).$$

If there exists a model for the data taking the ordering information as data, probability distributions associated with counterfactual survey instrument structures can be estimated. Then the true counterpart to the Clinton-Gore averaging is given by

$$\frac{1}{|\mathcal{Z}|} \sum_{z \in \mathcal{Z}} p(Y_i = k|Z = z)$$

where $\mathcal{Z}$ is the set of possible instrument structures. Doing so requires a method to generate the set of counterfactual probability mass functions. Note that the 13 traits comprising the list of alternatives admits admits no natural or pre-

Order	Design 1	Design 2
1	Manners	Studious
2	Success	Interested
3	Honest	Considerate
4	Clean	Responsible
5	Judgment	Obey
6	Control	Amicable
7	Role	Role
8	Amicable	Control
9	Obey	Judgment
10	Responsible	Clean
11	Considerate	Honest
12	Interested	Success
13	Studious	Manners

Table 2.2: Actual Survey Instrument Structures Implemented

ferred order; neither of the two forms is more natural than the other. Then any counterfactual order would have been as plausible as the two observed survey instrument structures. This empirical fact lends credence to exchangeability assumptions. That is, responses to the survey prompted with counterfactual instrument structures are independent of the observed survey instrument structure.

Electronic surveys are also subject to response effects. [68] provides an additional example of an intrusive observation process. Individuals were surveyed about their opinions regarding the response to Hurricane Katrina. Two versions of survey were used: one listing ranked alternatives in ascending order and another listing them in descending order. For example the first question in the survey asked about respondents' interest in the events surrounding Hurricane Katrina. For a subset of questions the alternative responses were listed in one of the two variations shown in 2.3. These are the two orders that preserve the relative ranking between alternatives. The authors demonstrate that the format assigned to respondents has a significant effect on the number of questions for which respondents chose one of the top two levels. Novel ordering response effects for this study are presented in 2.3.

Order	Form A	Form B
1	Extremely interested	Not interested at all
2	Very interested	Slightly interested
3	Moderately interested	Moderately interested
4	Slightly interested	Very interested
5	Not interested at all	Extremely interested

Table 2.3: Actual Survey Instrument Structures Implemented

2.3 Counterfactual Survey Responses

This section proposes a method to evaluate the sensitivity of the observed survey responses to changes in the instrument structure using the electronic survey introduced in 2.2. Respondents received either form A or B. 2.3 details the two structures presented for question one. In one version, the choices are in ascending order; in the other they are presented in descending order. Both structures preserve the ordinal nature of the data while altering the order of presentation. The experiment records for each respondent $i = 1, \ldots, N$ the response given A or the response given B, but not both. Figure 2.1 shows the difference in responses between the two groups A Chi-Squared test of the contingency table associated with 2.1 concludes the test statistic is as large at the observed statistic in 5.6% of samples under the null hypothesis of independence. This result suggests meaningful variation exists in the responses between the two groups. Form A lists the possible choices in descending order and places relatively more mass on the two largest values. The responses of the individuals who received form B place relatively more mass on the bottom three levels - which appeared closer to the beginning of the list of choices on their survey. The forms were assigned randomly to individuals. Hence the potential outcome Q^z is distributed independently of Z, the actual survey instrument structure, for all $z \in \mathcal{Z}$ ([51]). Then for $i = 1, \ldots, N$ and $z_i \in \{A, B\}$

$$p(Q_i^{z_i} = q_i | Z_i = A) = p(Q_i^{z_i} = q_i | Z_i = B).$$

As a consequence the probability mass functions governing the response to the question can be computed for counterfactual survey instrument structures. Consider the average of the two probability mass functions

$$p(Q_i = q_i) = \frac{1}{2}\left(p(Q_i^A = q_i) + p(Q_i^B = q_i)\right).$$

The resulting probability mass function equally weights the contribution of all potential outcomes: estimation of the probability distribution over a respondent's selection accounts for all the possible ways the survey data could have been generated. The instrument structure assigned to a respondent biases the response. Use of form A leads to overes-

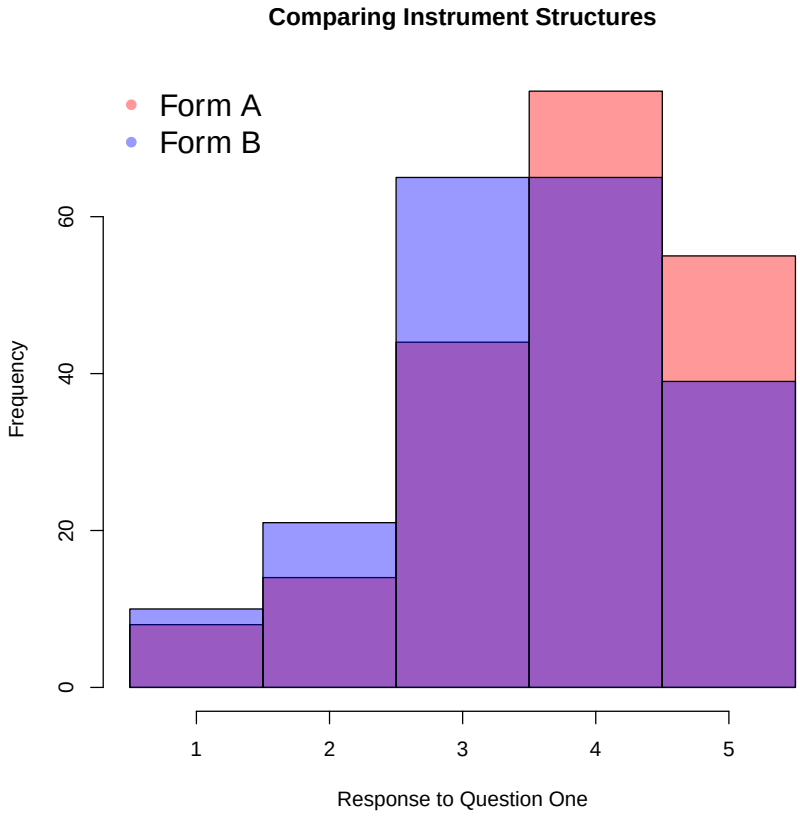

Figure 2.1: Histogram of Responses by Survey Form.

timates of interest in hurricane Katrina while form B leads to underestimates. Under this distribution Q_i is distributed independently of the survey instrument structure. If this is the true scientific object of interest, then averaging removes the bias introduced by the survey instrument structure. In the theory advanced by [94], considering all possible survey instrument structures leads to surveys in which the respondent oversamples from every subset of ideas made more salient by a particular survey instrument structure. Averaging corrects for this oversampling.

Under certain conditions, averaging also reduces the variance in the estimates of the probability mass functions. Variation in $p_{\hat{\theta}}(Q_i = q_i|z)$ arises from uncertainty regarding the estimator $\hat{\theta}$. If the estimators $\hat{\theta}_z$ for $z \in \mathcal{Z}$ are independent of one another, then

$$\begin{aligned}\text{var}\left(p(Q_i = q_i)\right) &= \frac{1}{4}\left(\text{var}\left(p(Q_i^A = q_i)\right) + \text{var}\left(p(Q_i^B = q_i)\right)\right) \\ &\leq \max\left\{\text{var}\left(p(Q_i^A = q_i)\right), \text{var}\left(p(Q_i^B = q_i)\right)\right\}\end{aligned}$$

so that if

$$\begin{aligned}\text{var}\left(p(Q_i^A = q_i)\right) &< 3\text{var}\left(p(Q_i^B = q_i\right) \\ \text{var}\left(p(Q_i^A = q_i)\right) &< 3\text{var}\left(p(Q_i^A = q_i\right)\end{aligned}$$

then

$$\text{var}\left(p(Q_i = q_i)\right) \leq \min\left\{\text{var}\left(p(Q_i^A = q_i)\right), \text{var}\left(p(Q_i^B = q_i)\right)\right\}.$$

That is, the variance of the aggregated probability mass function is always less than the most variable mass function. And, when the variances of the potential outcome mass functions are sufficiently similar, aggregation reduces the uncertainty in the estimate. In general, though, the model parameters do depend upon one another, and the inequalities do not apply.

In this analysis the parameter θ is assumed to be invariant across instrument structures, $\theta_z = \theta$ for all $z \in \mathcal{Z}$. The probability respondent i chooses level k is a function of the estimated parameters. To model the choice as a propor-

tional odds regression set

$$\begin{aligned} p(Q_i = k|z_i, x_i; \theta) &= p(Q_i \leq k|z_i, x_i; \theta) - p(Q_i \leq k-1|z_i, x_i; \theta) \\ &= F(\alpha_{k,k+1} - (z_i, x_i)^\top(\beta_z, \beta_x)) - F(\alpha_{k-1,k} - (z_i, x_i)^\top(\beta_z, \beta_x)) \end{aligned} \tag{2.3.1}$$

(2.3.2)

where F is the Cauchy distribution function with center zero and scale one,

$$F(v) = \frac{1}{\pi}\arctan(v) + \frac{1}{2}.$$

The parameter vector θ subsumes the cut-point parameters α, the parameters associated with survey instrument structure β_z and the parameters associated with respondent attributes β_x. Figure 2.2 displays the coefficients with 95% confidence intervals. The parameter 'Form' is an binary variable indicating whether the respondent received form

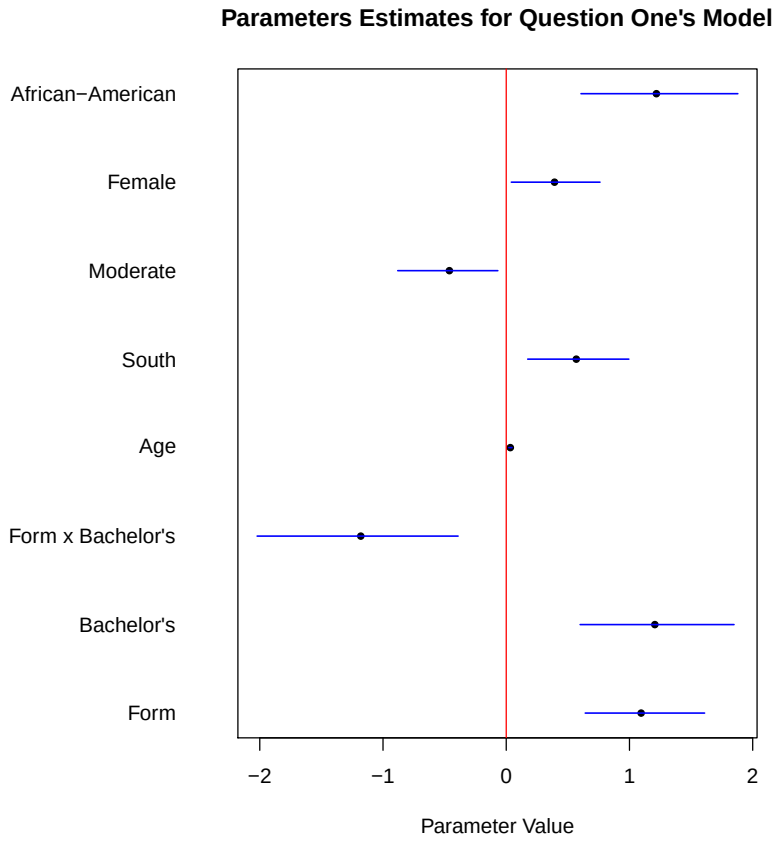

Figure 2.2: Parameter Estimates with Confidence Intervals for Question One's Model.

A; 'Bachelor's' indicates whether the respondent attained a Bachelor's degree; 'Age' indicates the respondent's age;

'South' is an binary variable indicating whether the respondent lives in the American South; 'Moderate' is a binary variable indicating the respondent identifies as a political moderate; 'Female' indicates whether the respondent is female and 'African-American' indicates whether the respondent is African-American. Corroborating results in [68], ordering effects disappear for respondents with a higher level of education. This can be seen by noting the sign and magnitude of the coefficients for 'Form' and 'Form x Bachelor's.'

2.3.1 Individual Responses

Define the aggregate probability mass function to be the arithmetic mean of the probability mass function conditioned on the actual and counterfactual instrument for respondent i,

$$p^*(Q_i = k|x_i;\theta) = \frac{1}{2}\sum_{z\in\{A,B\}} p(Q_i = k|z, x_i;\theta) \tag{2.3.3}$$

with the empirical counterpart

$$p^*(Q_i = k|x_i;\hat{\theta}) = \frac{1}{2}\sum_{z\in\{A,B\}} p(Q_i = k|z, x_i;\hat{\theta}). \tag{2.3.4}$$

Since the maximum likelihood estimates of θ are asymptotically normal, the variance of $p^*(\widehat{Q_i = k}|x_i;\hat{\theta})$ can be approximated via the Δ-method as

$$\text{var}(p^*(Q_i = k|x_i;\hat{\theta})) \approx \left(\nabla_\theta \sum_{z\in\mathcal{Z}} p(Q_i = k|z, x_i;\hat{\theta})\right)^\top \Sigma(\hat{\theta})\nabla_\theta \sum_{z\in\mathcal{Z}} p(Q_i = k|z, x_i;\hat{\theta}) \tag{2.3.5}$$

where $\Sigma(\hat{\theta})$ is the estimated covariance matrix associated with the maximum likelihood estimate $\hat{\theta}$. The estimator of 2.3.3 converges in distribution to a Gaussian distribution with mean 2.3.4 and variance Bootstrapping provides an alternative method to compute variances. Sampling respondents from the data with replacement and computing the parameter estimate R times obtains estimates $(\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(R)})$. For each sample $r = 1, \ldots, R$, compute the fitted probability mass function conditioned on the actual instrument $p(Q_i = q_i|x_i, z = z_{\text{obs}};\hat{\theta}^{(r)})$, the fitted probability mass function conditioned on the counterfactual instrument $p(Q_i = q_i|x_i, z = z_{\text{cf}};\hat{\theta}^{(r)})$ and the aggregate probability mass function $p^*(Q_i = q_i|x_i;\hat{\theta}^{(r)})$. Figure 2.3 shows bootstrapped estimates of the median and 95% confidence intervals of the actual and counterfactual distributions for 25 randomly selected respondents. The blue letter in the plot - an indication the version of the form assigned for the survey - marks the median probability for level two, "Slightly interested." Figure 2.4 shows the actual and averaged distributions. Note the bootstrapped sample variance of the averaged distribution is less than the actual distribution for 242 of 397 respondents; a paired t-test indicates the mean difference is less for the averaged distributions.

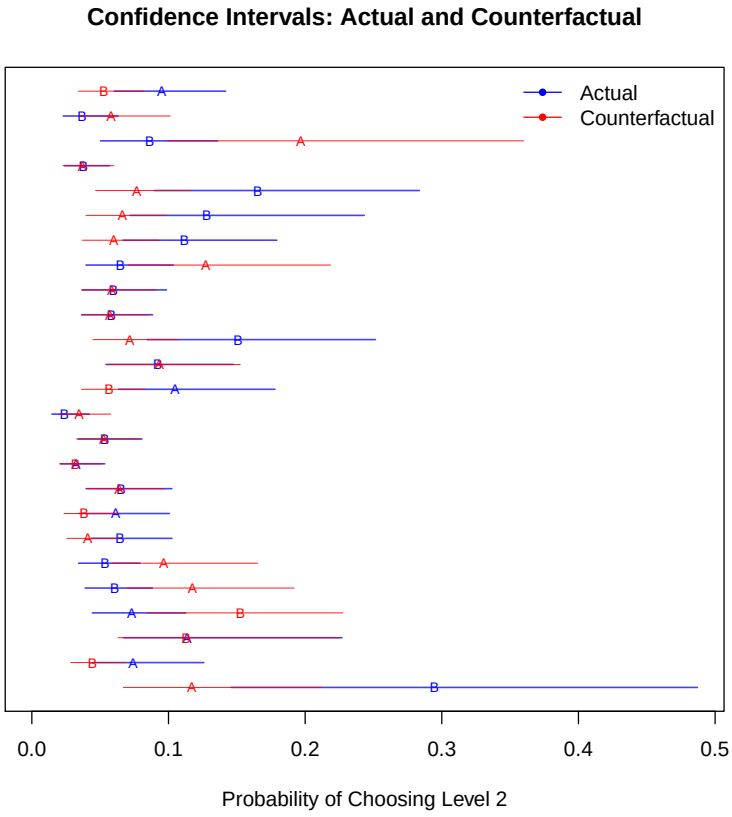

Figure 2.3: Bootstrapped Probability Estimates with Confidence Intervals.

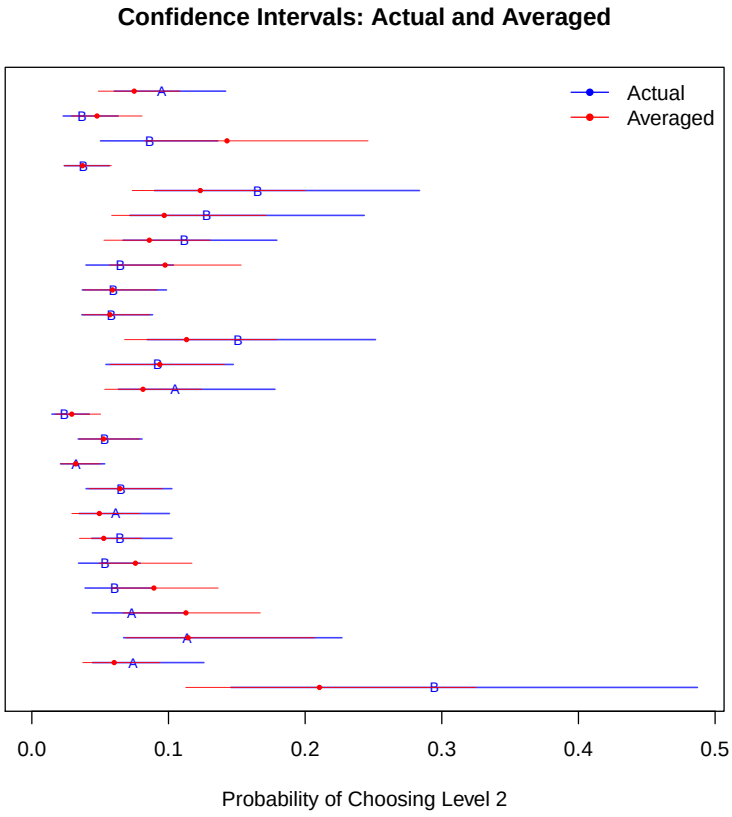

Figure 2.4: Bootstrapped Probability Estimates with Confidence Intervals.

2.3.2 Pooling Respondents

The electronic survey was sufficiently anonymous to make the stable unit treatment value assumption (SUTVA). That is, assume

$$p(Q_i^{z_j} = q_i | Z_j = A) = p(Q_i^{z_i} = q_i | Z_j = B)$$

for i and $j \neq i$ in $1, \ldots, N$. By pooling across the respondents, the set of possible instrument structures becomes $\mathcal{Z} = \{A, B\}^N$. The population average

$$p^*(\mathbf{Q} = \mathbf{q} | x; \hat{\theta}) = \frac{1}{N} \sum_{i=1}^{N} p^*(Q_i = q_i | x_i; \hat{\theta}) \tag{2.3.6}$$

aggregates the probability mass function of each unit in the sample. Similar inferential conditions apply to the population average. The distribution of the overall average is asymptotically normal with mean 2.3.6 and variance

$$\text{var}(p^*(\mathbf{Q} = \mathbf{q} | x; \hat{\theta})) \approx \left(\nabla_\theta p^*(\mathbf{Q} = \mathbf{q} | x; \hat{\theta})\right)^\top \Sigma(\theta) \nabla_\theta p^*(\mathbf{Q} = \mathbf{q} | x; \hat{\theta}).$$

Bootstrapped estimates of the response probabilities for levels two through five are presented in 2.5. Each panel shows a bootstrapped distribution for the aggregate probability mass function defined by

$$p^*(\mathbf{Q} = \mathbf{q} | x; \hat{\theta}^{(r)}) = \frac{1}{N} \sum_{i=1}^{N} p(Q_i = q_i | z, x; \hat{\theta}^{(r)})$$

where $r = 1, \ldots, R$ indexes the bootstrap replications. The proportions observed in the original survey are denoted by the red line. The observed values for levels three and five deviate from the mean of the bootstrapped distribution by more than .05. The observed data is an unlikely realization given the responses to the question are distributed according to the aggregate probability mass function.

2.3.3 Extreme Instrument Structures

In the original survey, respondents were randomly assigned form A with some probability π_A. Sampling from this space creates counterfactual treatments to all respondents with the same marginal probabilities. A counterfactual estimate of the population-wide probability mass function given a vector z of instrument structures can be compared to the set of all such estimates by computing

$$\sum_{z' \in \mathcal{Z}: z' \neq z} d(p(Q = q | z = (z_1, \ldots, z_N)), p(Q = q | z' = (z'_1, \ldots, z'_N))), \tag{2.3.7}$$

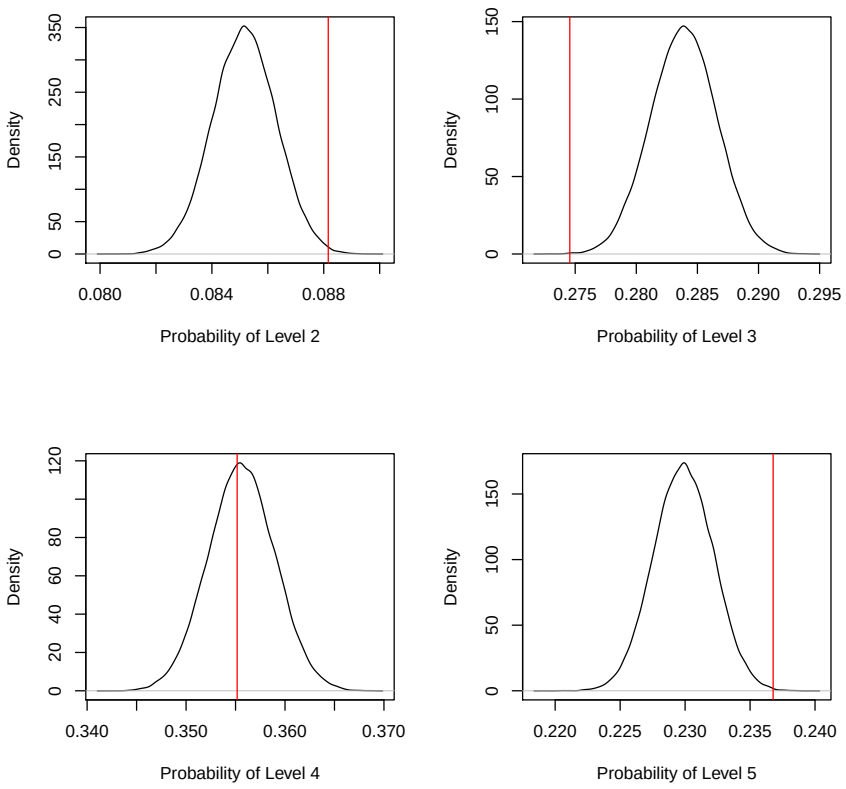

Figure 2.5: Density Estimate of the KL Divergences.

the total distance of $p(Q = q|z = (z_1, \ldots, z_N))$ from all the alternatives. The sum indicates the extremity in the estimate relative to all others in the collection. Since this is also a function of the estimator $\hat{\theta}$, a further application of the Δ-methods yields an approximation of its asymptotic distribution. Figure 2.6 displays a non-parametric density computed with the KL divergences between probability mass functions conditioned on z and their geometric mean. The fitted mass function evaluated with the actual survey instrument structure is much closer to the majority of the

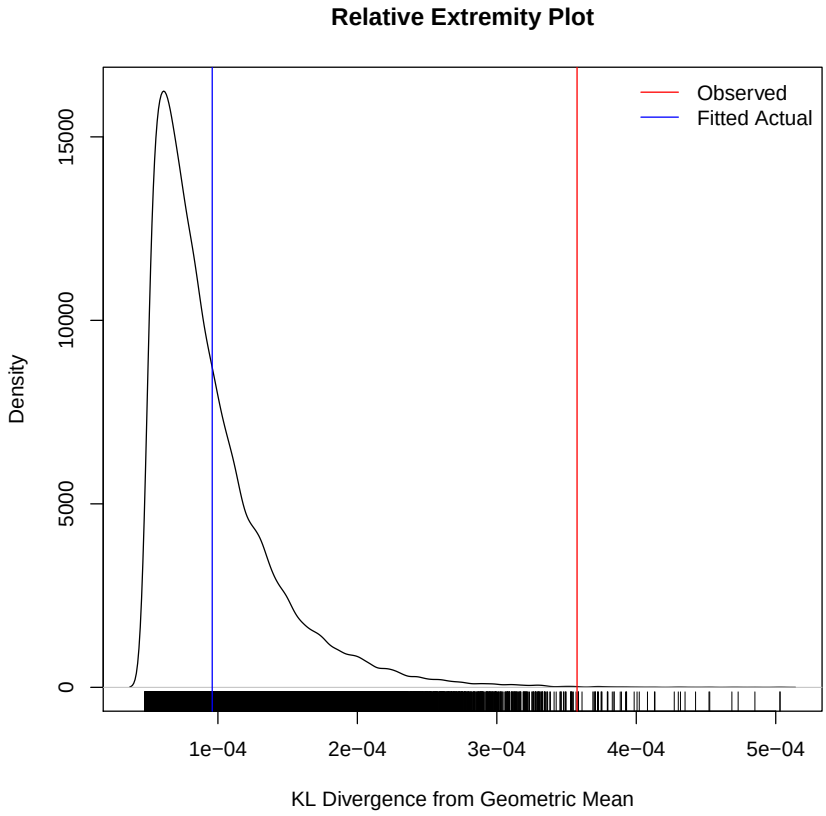

Figure 2.6: Density Estimate of KL distances.

survey designs than the observed probabilities. The response generating process cannot be perfectly modeled; this method works better when the responses can be explained well by the variables in the observation and data-generating processes. The observed distribution's distance from the majority of probability mass functions reflects two sources of variation: variation due to the actual survey instrument structure and non-systematic variation at the unit level. The latter combines the association of unmeasured variables and idiosyncratic noise.

Setting the distance function in 2.3.7 to be the KL divergence implies a common reference point. Observe

$$\frac{1}{|\mathcal{Z}|}\sum_{z'\in\mathcal{Z}}\mathrm{KL}(p(Q^z)\|p(Q^{z'})) = \mathrm{KL}\left(p(Q^z)\middle\|\prod_{z'\in\mathcal{Z}}p(Q^{z'})^{\frac{1}{|\mathcal{Z}|}}\right).$$

The right side of the term on the last line is the geometric mean of the probabilities $\left\{p(Y^{z'}=y)\right\}_{z'\in\mathcal{Z}}$ meaning that the average KL-divergence is a measure of distance from the probability of agreement between all z'. The most outlying member of the collection is defined as the probability mass function association with

$$z^* = \underset{z\in\mathcal{Z}}{\mathrm{argmax}}\,\mathrm{KL}\left(p(Q^z)\middle\|\prod_{z'\in\mathcal{Z}}p(Q^{z'})^{\frac{1}{|\mathcal{Z}|}}\right).$$

Then the ratio

$$\frac{\mathrm{KL}\left(p(Q^z)\middle\|\prod_{z'\in\mathcal{Z}}p(Q^{z'})^{\frac{1}{|\mathcal{Z}|}}\right)}{\mathrm{KL}\left(p(Q^{z^*})\middle\|\prod_{z'\in\mathcal{Z}}p(Q^{z'})^{\frac{1}{|\mathcal{Z}|}}\right)}$$

measures the relative extremity of Q^z. Note that by the AM-GM inequality, for all $q\in\mathcal{Q}$,

$$\prod_{z'\in\mathcal{Z}}p(Q^{z'}=q)^{\frac{1}{|\mathcal{Z}|}} < \frac{1}{|\mathcal{Z}|}\sum_{z'\in\mathcal{Z}}p(Q^{z'}=y),$$

so that the sum over all elements in $\mathcal{Q}$ is less than one. Normalizing defines the probability mass function proportional to the geometric mean of the collection of counterfactual probability mass functions,

$$\Psi(Q=q) = \frac{\prod_{z'\in\mathcal{Z}}p(Q^{z'}=q)^{\frac{1}{|\mathcal{Z}|}}}{\sum_q\prod_{z'\in\mathcal{Z}}p(Q^{z'}=q)^{\frac{1}{|\mathcal{Z}|}}}.$$

Then

$$\mathrm{KL}\left(p(Q^z)\|\Psi\right) = \log(K) + \mathrm{KL}\left(p(Q^z)\middle\|\prod_{z'\in\mathcal{Z}}p(Q^{z'})^{\frac{1}{|\mathcal{Z}|}}\right)$$

where

$$K = \sum_y\prod_{z'\in\mathcal{Z}}p(Q^{z'}=q)^{\frac{1}{|\mathcal{Z}|}}.$$

Then the elements that maximizes (minimizes) the distance from the unnormalized version will also maximize (minimize) the distance from the normalized version.

Using the Wasserstein distance between functions implicitly compares the probability mass functions to their Wasserstein barycenter. The total Wasserstein distance from all others is given by

$$\sum_{z' \neq z} d(p_z, p'_z)^p = \sum_{z' \neq z} \inf_{\pi(z,z')} \sum_x \sum_y d(x,y)\pi(x,y).$$

The probability mass function closest to the $\{p(Q|z\}_{z \in \mathcal{Z}}$ is the Wasserstein barycenter. While

$$\underline{\mathrm{p}}(Q = q) = \underset{p(Q|z_1),\ldots,p(Q|z)}{\operatorname{argmin}} \sum_{z' \neq z} d(p_z, p'_z)^p \tag{2.3.8}$$

is the member of the collection closest to the other functions. This suggests a different diagnostic: rank the members of the collection according to their Wasserstein distance from the Wasserstein barycenter. Observe that if $\underline{\mathrm{p}}(Q = q)$ solves 2.3.8, then it is also the member of the collection closest to the Wasserstein barycenter.

2.3.4 Range of Possible Survey Outcomes

The experiments already considered preserve the observed distribution of respondents assigned to form A and B. But, the probability of assignment to each of the two groups is a feature of survey design. Removing the constraints augment the set of possible survey instrument structure to be $\{A, B\}^N$. The surveys in which everyone receives A and everyone receives B belong to this set. Sampling π_A from the uniform distribution and then drawing form assignments generates the full set of possible survey instrument structures. Figure 2.7 shows the density estimates for a sample of assignment vectors. In each of the four panels, the red line indicates the observed probability, the green line represents the probability had all been given form A and the purple presents the probability of response had all respondents been assigned to form B. The distance between the green and purple lines measures the range of possible probabilities achievable by different designs. The probability distributions exhibit significant variation; for example, the percentage of respondents indicating moderate interest ranges from about 21 under the all form A design to 35 under the all form B design. These plots illustrate the extent to which the subjective opinions of human respondents can be manipulated via alternative survey structures. As in the case in which the number of respondents receiving form A and B are fixed, a measure of extremity can be defined.

2.4 A Multi-Question Framework

Unlike paper-based surveys, electronic surveys enforce the direction in which respondents complete the questionnaire; the questions must be answered in the order in which they appear. Any such temporally-ordered survey can be

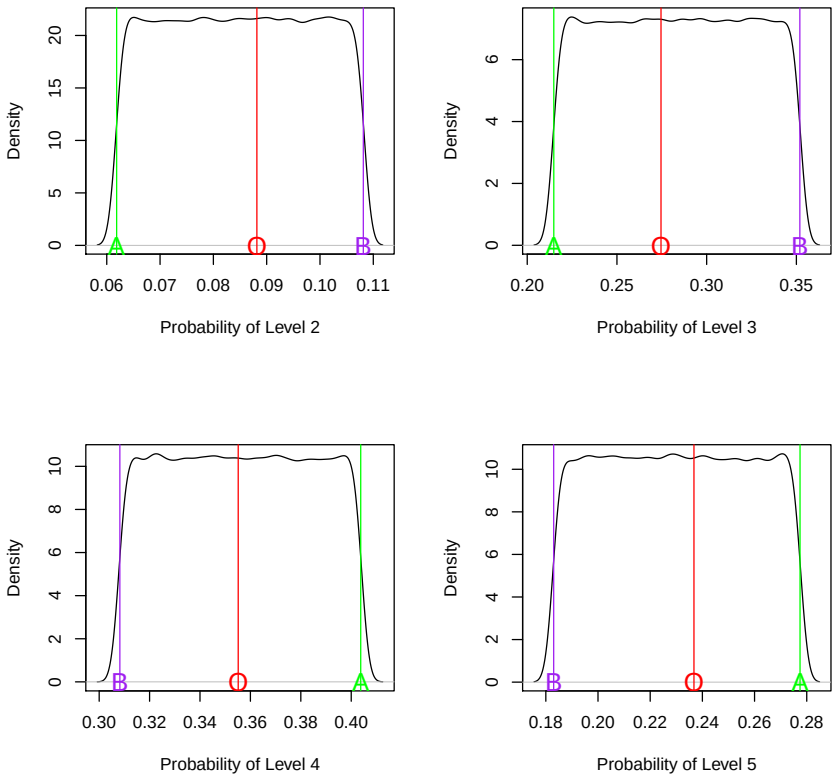

Figure 2.7: Estimates of the Aggregated Overall Mass Function.

q1/q2	1	2	3	4	5
1	0.025	0.008	0.005	0.003	0.005
2	0.035	0.038	0.010	0.003	0.003
3	0.005	0.076	0.164	0.030	0.000
4	0.003	0.003	0.149	0.197	0.005
5	0.000	0.000	0.013	0.081	0.144

Table 2.4: Observed Joint Distribution of Questions One and Two.

decomposed as

$$p(\mathbf{Q} = \mathbf{q}) = p(\mathbf{Q}_1 = \mathbf{q}_1, \ldots, \mathbf{Q}_M = \mathbf{q}_M) = \prod_{m=1}^{M} p(\mathbf{Q}_m = \mathbf{q}_m | \mathbf{q}_{m'}, m' < m).$$

The choices made in the preceding questions affect the responses to subsequent questions. Table 2.4 depicts a simple illustration of this phenomenon. The joint probability density of questions one and two from the survey exhibits dependence. For example, 56.7% of respondents supplied the same answers to questions one and two. If the response to question two can be modeled in terms of the response to question one and the response to question one depends on the presentation order of the alternatives, then response to question two indirectly depends on the presentation of question one. Survey instrument effects in one part of the survey can spill over to affect other responses. Questions one and two can be modeled as

$$p(\mathbf{Q}_1, \mathbf{Q}_2 | z, x) = p(\mathbf{Q}_2 | \mathbf{Q}_1, z, x) p(\mathbf{Q}_1 | z, x).$$

The mass functions for Q_1 and Q_2 depend on the respondent level attributes x. Figure 2.8 depicts the causal relations assumed to exist as a direct acyclic graph. Randomization in the assignment of Z to respondents ensures the causal effect of Z on Q_1 is identified. Identification of the causal effect of Z on Q_2 requires conditioning on Q_1 and X. Doing so blocks the two backdoor paths between Z and Q_2 ([80]). Hence probability mass functions defined with counterfactual values of Z can be computed and compared. Since question two is ordered, the same probability distributions are used as question one save for the parameters. Figure 2.9 shows the parameter estimates with 95% confidence intervals of the marginal distributions of questions one and two for choices three, four and five. Bootstrapped estimates of the joint probability distribution are computed as

$$p^*(\mathbf{Q}_1 = \mathbf{q}_1, \mathbf{Q}_2 = \mathbf{q}_2 | x; \hat{\theta}_1^{(r)}, \hat{\theta}_2^{(r)}).$$

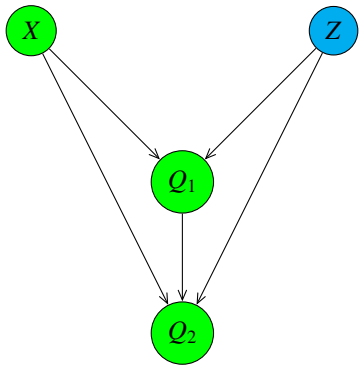

Figure 2.8: Survey Process Model

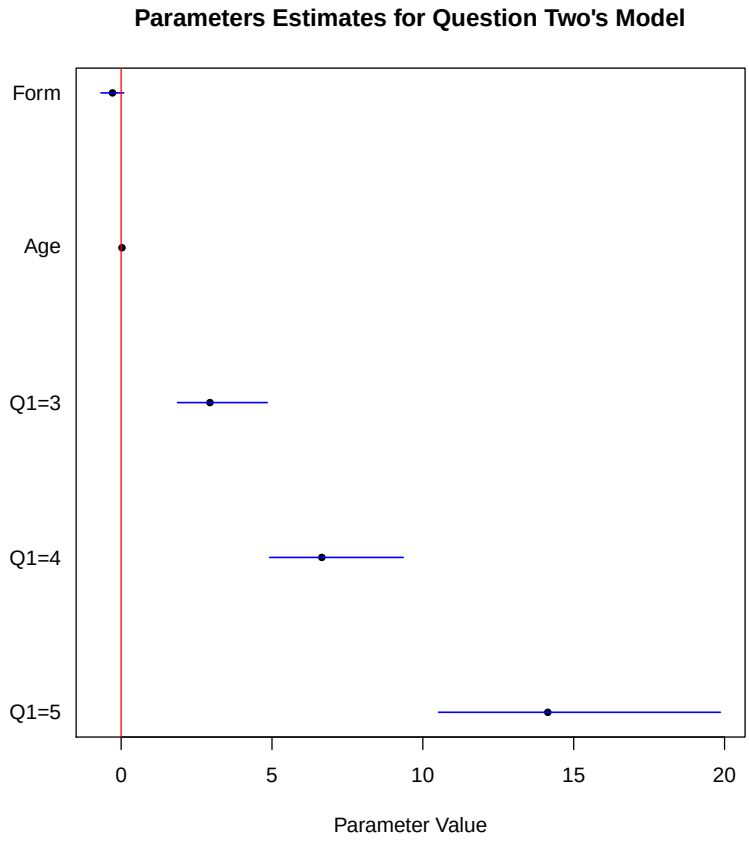

Figure 2.9: Parameter Estimates with Confidence Intervals for Question Two's Model.

Figure 2.10 plots the aggregate marginal probability distribution for question one by the aggregate marginal probability for question two for a sample of counterfactual survey instrument structures with 95% confidence ellipses. The large variation in the responses to question two of the survey reflect the two sources of variation: variation due to the effect of Z on the response to question two and the variation in question one due to Z. In the actual survey the presentation order

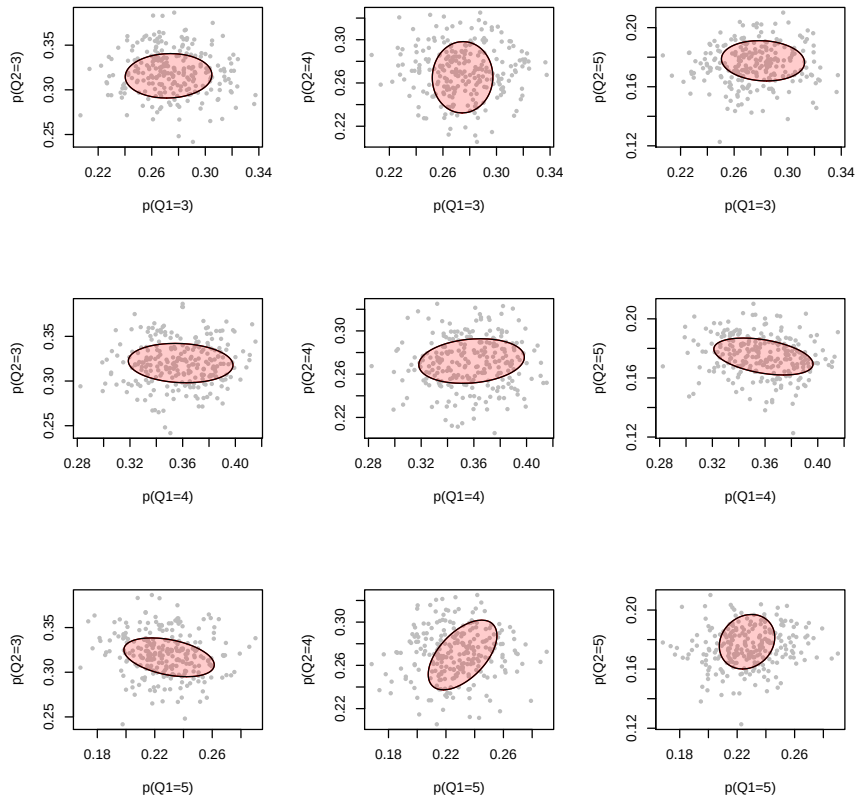

Figure 2.10: Aggregate Probability Estimates with 95% Confidence Ellipses for Question One and Two.

did not vary across questions. That is, $(z_1, z_2) = (z, z)$. But, theoretically, the presentation could have varied; it would have been possible to alternate the order of responses across consecutive questions. Exchangeability in this scenario requires the potential outcomes of questions one and two are distributed independently from the survey instrument actually administered to respondents.

$$p(\mathbf{Q}^{(z_1,z_2)}|z' = (z'_1, z'_2)) = p(\mathbf{Q}^{(z_1,z_2)}|z'' = (z''_1, z''_2))$$

for all $z', z'' \in \{A, B\}^{2N}$. This assumption implies the same model would have been estimated had instrument structure varied across questions. The graphical model depicting the survey process changes to the DAG in Figure 2.11. The universe of possible outcomes expands greatly.

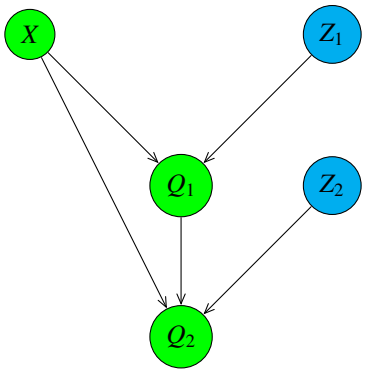

Figure 2.11: Counterfactual Survey Process Model

More generally, this procedure offers a method to compute probability distributions for all questions in the survey. The set of all probability mass functions Algebraically the counterfactual joint probability mass function is given by

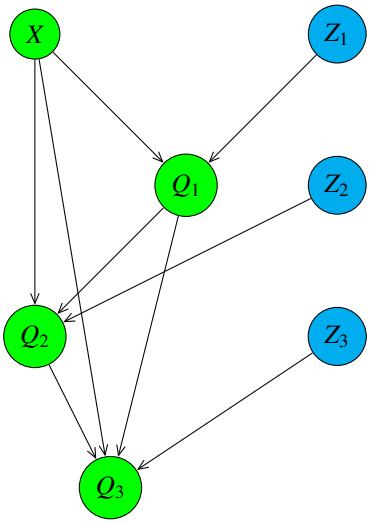

Figure 2.12: Counterfactual Survey Process Mode of Length Three.

$$p(\mathbf{Q} = \mathbf{q}|z = (z_1, \ldots, z_M), x) = \prod_{m=1}^{M} p(\mathbf{Q}_m|z_M, x, \mathbf{q}_{m'}, m' < m).$$

The variation introduced by instrument structures propagates through the survey. Then extremity measures for the entire survey instrument can be computed, as can the range of possible surveys under the set of possible structures $\{A, B\}^{MN}$.

Integrating question order effects in the framework presents an interesting application. The model in 2.12 is altered to accommodate the new question order. Let z_q denote the question order; the set of question orders is the set of all

permutations of length M. If the order of questions varies in the data the aggregate probability mass function

$$p^*(\mathbf{Q} = \mathbf{q}|x) = \frac{1}{|\mathcal{Z}|M!} \sum_{z_q} \sum_{z \in \mathcal{Z}} p(Q = q|z = z = (z_1, \ldots, z_M), z_q, x)$$

averages over both question and response orders.

2.5 Discussion

Much occurs before data is ever analyzed. A scientific phenomenon of interest must be conceived in some way. This conception is given a logico-semantic representation. This representation must be measured with an instrument in some larger context. The various choices made prior to the analysis shape the observed values. Whether this has any practical effect on the conclusions (or inferences) made from the data depends on the sensitivity of the values with respect to the range of possible choices at each level. Survey data constitutes an interesting case study: the object that is measured consciously participates in the measurement. In these cases the observation and data-generating processes cannot be separated. This paper provides a graphical model encapsulating random variables belonging to the data-generating and observation processes. Using counterfactual logic, a method is proposed to average over the different contexts under which respondents could have completed the survey. The information in a collection of models with counterfactual survey designs is aggregated to form a single probability function. This can be used to observe the sample free from the hue cast by a particular instrument structure. Then the paper suggests a diagnostic to analyze the extremity of a survey instrument's structure.

Survey instrument structure matters. When scientists write a survey, they intend to measure the distribution of subjective ideas in the minds of respondents. They do not seek to measure the opinions of a group of respondents as restricted or influenced by a particular way of presenting the survey. This paper has shown that reports are sensitivity to instrument structure across different instrument types: face-to-face, mail and electronic surveys. When the instrument structure varies across respondents and assumptions about exchangeability are plausible, counterfactual survey responses can be generated and compared to the observed data. Empirical examples computed using the post Katrina survey conducted by Knowledge Networks in 2006 demonstrates strong response order effects. Since the survey enforced respondents to complete it sequentially, each question can be modeled in terms the previous questions, features of instrument structure and respondent-level attributes. Variation in the joint probability mass function over elements in the space of survey instrument structure can be assessed.

2.5.1 Counterfactual Survey Time Limits

The survey studied in this paper also contains data on the length of time required for the respondent to complete the survey. The electronic format did not place an explicit constraint on the amount of time allowed to complete the survey. If exchangeability assumptions are warranted, then time limits can be construed as another form of survey instrument structure. Then for survey instrument structure for a single individual is the vector

$$(z_1, \ldots, z_M, t).$$

If respondents decided how much time to allocate to the survey prior to completion, then the causal effect of time on the response can be computed. This can also hold if the amount of time allocated to each is decided upon reading the prompt. If, however, the time spent answering questions is a consequence of the survey responses, then

$$p(Q^{z,t}|z', t') \neq p(Q^{z,t}|z', t'').$$

Additionally, since the survey was web-based, the amount of time required for a respondent to complete the survey was recorded. Estimates of the survival curve are presented in 2.13. Since some rule regarding the amount of time allotted for completion must be chosen, including the choice of no time limit, this feature of the data collection process can also be characterized as survey instrument structure. Hence the survey instrument structure subsumes the ordering of alternatives and the amount of time provided to respondents.

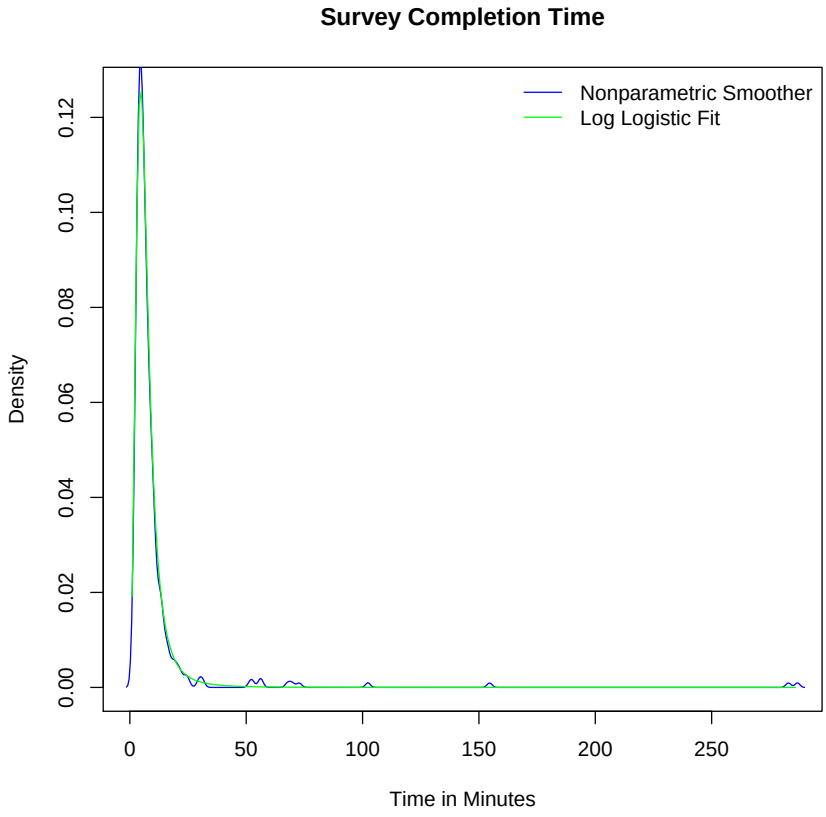

Figure 2.13: Results of Maximum Likelihood Estimation.

Chapter 3

The Aggregation of Measures

3.1 Introduction

In some scientific settings a collection of probability mass functions over a finite set of states must be aggregated. In chapters 2 and 5 collections of probability mass functions conditioned on counterfactual instrument structure are analyzed. In both cases the surveyed information depends upon the survey instrument. Then the responses do not represent the primary object of scientific interest, the responses independent of a particular sampling instrument. To average over the effects of the survey instrument structure, a method is needed to produce a probability mass function independent of a particular survey instrument structure. Other common needs for such a procedure include the aggregation of a collection of experts and the creation of informed prior distributions.

This chapter explores several methods to do just this in the absence of training data. Two options receive significant attention: the arithmetic average and the Wasserstein barycenter. To do this criteria for the comparison of alternative measures are defined. It is shown that the arithmetic average of a set of probability mass functions is optimal in the sense that it is the closest to the component. Moreover, this measure is the only aggregating measure to give equal weight to each measure for every collection of measures. This egalitarian property is a stronger version of positive responsiveness - the aggregating measure is monotonically increasing in each of its arguments. Although it is a natural choice, if the members of the collection belong to an exponential family, the egalitarian aggregating measure is not conjugate. The probability distribution of the sum of members from an exponential family is not a member of that exponential family.

An alternative to the egalitarian measure is the Wasserstein barycenter. The barycenter when the distance between

any pair of distinct elements is 0 or 1 is explored. The barycenter solutions satisfy boundedness (and therefore unanimity), but they are not responsive. Boundedness ensures the aggregate measure places no more or less probability than the members of the collection do individually.

A third class of aggregating measures are the measures proportional to the power means. The geometric mean belongs to this class. In applications, the power mean can confer computational benefit. Moreover, when the collection of probability mass functions share a common exponential family, the (weighted) geometric results in a conjugate aggregating measure. Although the aggregating measure proportional to the geometric mean is exchangeable and unanimous, it is not positively responsive nor is it bounded.

By introducing weights to the class of geometric means an optimal approximation to the egalitarian measure can be computed that is conjugate to the collection. This is achieved by minimizing the KL divergence (or other measures of distance between probability mass functions) relative to the egalitarian measure subject to the boundedness conditions and, possibly, the requirement that the aggregating measure is no more informed than the egalitarian measure in the sense that the entropy of the aggregating measure is at least as large as the entropy of the egalitarian measure. Unlike other aggregating measures proportional to a power mean, the geometric mean admits a probabilistic interpretation; it is the normalized probability that the component measures agree on subsets of the outcome space. This is the probability of unanimity.

3.2 The Egalitarian Measure

Consider a collection of probability measures $\{P_\omega\}_{\omega\in\Omega}$ over a vector of random variables $(Y_1, \ldots, Y_K)$. Let Ψ be a probability measure over the same set of random variables that aggregates the information contained in the collection $\{P_\omega\}_{\omega\in\Omega}$. A natural solution, as presented in [86], is to define Ψ to be the measure nearest to the members of the collection of measures. Let M be the $N \times K$ matrix with the j^{th} column equal to $(p_j(1), \ldots, p_j(K))$. Choose $\Psi \in \Delta_{K-1}$ to minimize the squared distance from M,

$$\sum_{j=1}^{N}\sum_{k=1}^{K}(\Psi(k) - p_j(k))^2.$$

The solution, given by the first-order conditions of the objective function

$$\mathcal{L}(\Psi, \lambda) = \sum_{j=1}^{N}\sum_{k=1}^{K}(\Psi(k) - p_j(k))^2 - \lambda(\sum_{k=1}^{K}\Psi(k) - 1),$$

is to use the average value for each k in the support,

$$\Psi(k) = \frac{1}{N}\sum_{j=1}^{N} p_j(k). \tag{3.2.1}$$

In words, simply choose the aggregate measure to be the expected probability with respect to some measure over subsets in Ω. In some cases a measure over ω may be known, such as the percentage of a certain type of sensor, or can be estimated using data. But, many applications admit no obvious choice. These situations require a more general formulation. This choice is the most natural definition and the subject of the next section. Equation 3.2.1 satisfies many important properties. In particular, it satisfies unanimity.

Definition 1. *A probability measure Ψ aggregating the mass functions $\left\{p_j(1), \ldots, p_j(K)\right\}_{j=1}^{N}$ satisfies unanimity at state $k \in \Omega$ if*

$$p_1(k) = p_2(k) = \cdots = p_N(k) = p(k)$$

implies

$$\Psi(k) = p(k).$$

Suppose N measures place probability $\frac{1}{3}$ on the outcome that it rains on a given day. Under what conditions should the aggregate probability differ from $\frac{1}{3}$? The evidence, represented by the probability mass functions, does not suggest a larger or smaller probability is warranted. If an aggregating measure is unanimous at $K-1$ states, then it is called unanimous. Since it is an average of the of the N probability mass functions, when all probability mass functions agree on the probability assigned to state k,

$$\Psi(k) = \frac{1}{N}\sum_{j=1}^{N} p_j(k) = p(k)$$

where $p(k)$ is the probability mass assigned to state k by all of the probability mass functions in the collection. Beyond unanimity, the egalitarian measure must be bounded by below and above by the minimum and maximum probability masses, respectively.

Definition 2. *A probability measure Ψ aggregating the mass functions $\left\{p_j(1), \ldots, p_j(K)\right\}_{j=1}^{N}$ $\{P_\omega(B)\}_{\omega\in\Omega}$ is said to be bounded if for all $k = 1, \ldots, K$,*

$$\Psi(k) \in \left[\min_j p_j(k), \max_j p_j(k)\right].$$

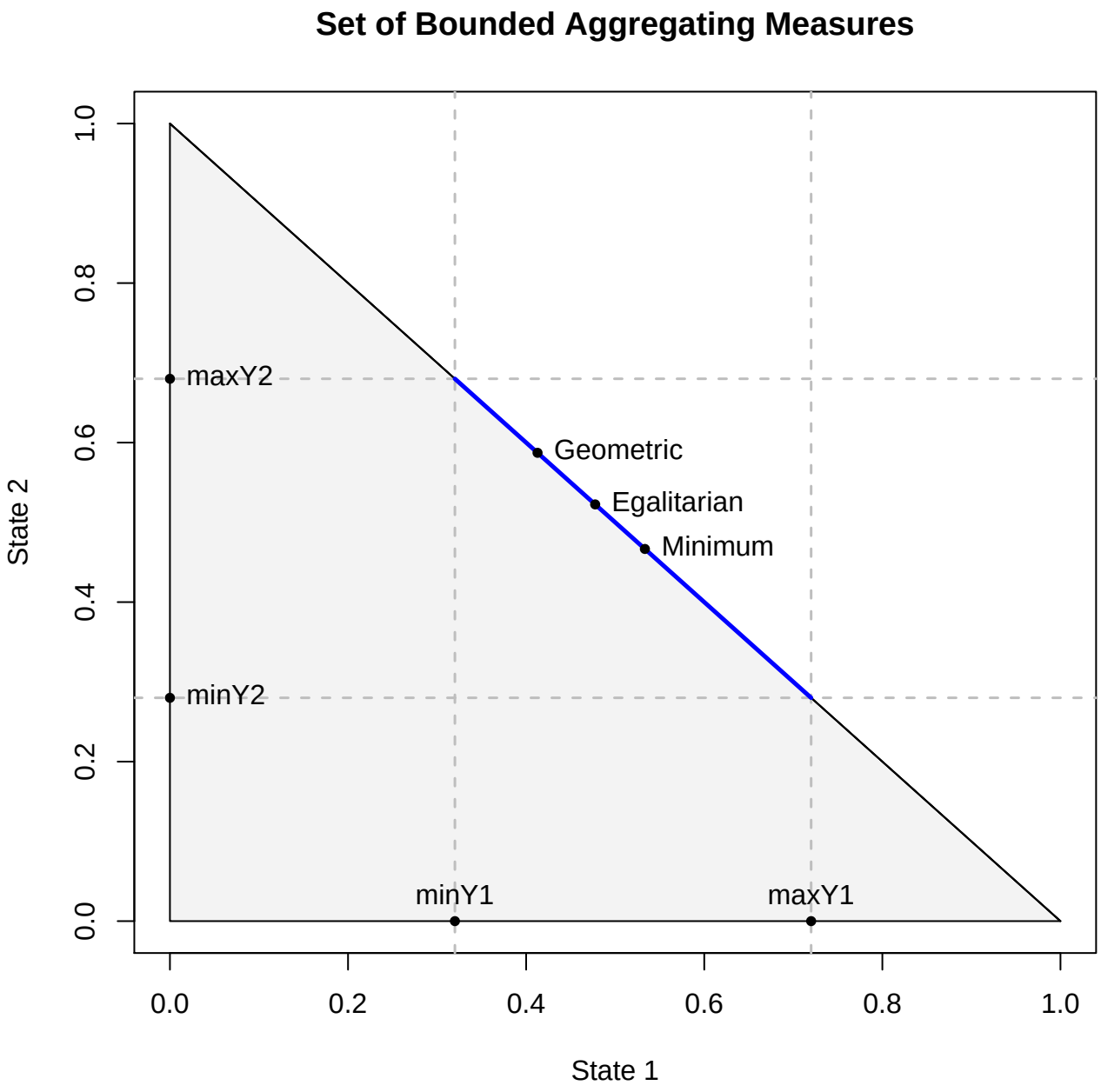

Figure 3.1: Region of bounded Aggregating Measures.

3.1 depicts the boundedness conditions. The blue line segment is the set of bounded aggregating measures. A bounded aggregate measure can be as pessimistic as the most pessimistic measure and as optimistic as the most optimistic measure. If the collection places mass on every element in the state space, then it cannot place mass outside of the union of the support of $\{p_j(k)\}$. This will not be true, in general, if there exists elements in the state space k such that $p_j(k) = 0$ for all $j = 1, \ldots, N$. Boundedness also implies unanimity.

Lemma 3.2.1. *If an aggregating measure Ψ is bounded, then it satisfies unanimity. Additionally,*

$$\operatorname{supp}(\Psi) \subseteq \bigcup_{j=1}^{N} \operatorname{supp}(p_j).$$

.

Proof. Assume the experts agree on a state k. Then boundedness implies

$$\kappa = \min_j p_j(k) \leq \Psi(k) \leq \max_j p_j(k) = \kappa.$$

Suppose $k \in \operatorname{supp} \Psi$. Then boundedness implies there exists some $j \in \{1, \ldots, N\}$ such that $p_j(k) > 0$ so that $k \in \operatorname{supp} p_j$. Hence k is contained the union of the support of each member of the collection. If $p_j(k) > 0$ for all j, then $\Psi(k) \geq \min_j p_j(k) > 0$ and

$$\operatorname{supp}(\Psi) = \bigcup_{j=1}^{N} \operatorname{supp}(p_j).$$

. □

Remark 1. *Lemma 3.2.1 provides a useful necessary condition: if Ψ does not satisfy unanimity, it cannot be bounded.*

Since any function that lies outside the range of the experts on a set of positive measure cannot be an adjusted measure, our definition excludes any non-unanimous measure. To see the use of this characterization suppose the proposed aggregating measure is

$$\Psi(k) = \frac{\exp\left\{\beta \frac{1}{N} \sum_j p_j(k)\right\}}{\sum_k \exp\left\{\beta \frac{1}{N} \sum_j p_j(k)\right\}}.$$

Then whenever the experts agree on an element k so that $p_j(k) = q(k)$,

$$\Psi(k) = \frac{\exp\{\beta q(k)\}}{\sum_k \exp\{\beta q(k)\}} \neq q(k)$$

for some k. Hence this choice is not bounded. Note the egalitarian measure is not the only bounded aggregating measure. One could choose to place weights on the individual probabilities such that

$$\Psi(k) = \frac{1}{N} \sum_{j=1}^{N} w_j p_j(k)$$

with $w \in \Delta_{N-1}$. Any choice of w gives rise to a bounded aggregating measure including placing all the weight on a

single expert. Such a construction ignores the information from the other $N-1$ experts.

Definition 3. *Say the measure Ψ aggregating over $\left\{p_j(k)\right\}_{j,k}$ is positively responsive to $j \in \{1,\ldots,N\}$ at state k if*

$$p'_j(k) - p_j(k) > 0 \implies \Psi'(k) - \Psi(k) > 0 \tag{3.2.2}$$

where Ψ' is the aggregating measure obtained by substituting $p'_j(k)$ for $p_j(k)$. If Ψ is differentiable with respect to p_j at k, then

$$\frac{\partial}{\partial p_j(k)}\Psi(k) > 0.$$

Positive responsiveness implies an increase in expert j's assessment of k is accompanied by an increase in $\Psi(k)$. An aggregating measure can also be negatively responsive. To see this let

$$\Psi(k) = \frac{1}{N}\sum_{j-1}^{N}(1 - p_j(k)).$$

Clearly, $\frac{\partial}{\partial p_j(k)}\Psi(k) < 0$ for all j and k. Positive responsiveness implies the aggregating measure moves in the same direction as component measures when they increase or decrease the mass placed on a given k. A stronger property of aggregating measures concerns how the rates of change relate to one another. The next concept defines such a property.

Definition 4. *The aggregating measure Ψ is said to be egalitarian at k if*

$$P'_i(k) - P_i(k) = P'_j(k) - P_j(k) = \epsilon > 0 \implies \Psi'(k) - \Psi(k) = \delta(\epsilon) > 0.$$

If Ψ is differentiable with respect to P_ω at B, then this can be expressed as

$$\frac{\partial}{\partial P_\omega(B)}\Psi(B) = \frac{\partial}{\partial P_\iota(B)}\Psi(B). \tag{3.2.3}$$

In words, each measure receives equal treatment at every point in the support. A weaker version of egalitarianism is possible. Let $\left\{p_j(k)\right\}$ and $\left\{q_j(k)\right\}$ be two probability mass functions and j, j' two distinct indices in $1,\ldots,N$. Whenever $p_j(k) = q_{j'}(k)$ and $q_{-j}(k) = q_{-j'}(k)$, then

$$p'_j(k) - p_j(k) = q'_{j'}(k) - q_j(k) \implies \Psi'(k;p) = \Psi'(k;q);$$

and, if the aggregating measure is differentiable

$$\frac{\partial}{\partial p_j(k)}\Psi(k;p) = \frac{\partial}{\partial q_{j'}(k)}\Psi(k;q).$$

This weaker form of egalitarianism is equivalent to saying that Ψ is invariant under permutations of the collection of measure. Component measures do not always receive the same weight, but they receive the same marginal weight whenever the two components agree and the remaining components are permutations of one another. To see an example of such an aggregating measure consider an aggregating measure proportional to the geometric mean

$$\Psi(k) = \frac{\left(\prod_j p_j(k)\right)^{1/N}}{\sum_k \left(\prod_j p_j(k)\right)^{1/N}}.$$

Clearly, this measure is invariant under permutations of the collection, but it is not egalitarian. The properties and egalitarianism imply the egalitarian aggregating measure. This is demonstrated in the following proposition.

Proposition 3.2.2. *Suppose Ω is finite and the aggregating measure Ψ for $\{P_i\}$ is bounded and egalitarian. Then*

$$\Psi(B) = \frac{1}{N}\sum_{i=1}^{N} P_i(B). \tag{3.2.4}$$

Proof. Since Ψ is bounded, it must be unanimous. By unanimity for all B,

$$\Psi(B|P_i(B) = 0, i \in [N]) = 0.$$

Let any $P_i(B)$ increase to $\epsilon_1 > 0$. Then $\Psi(B) = \delta(\epsilon_1)$. If another $P_j(B)$ increases to $\epsilon_2 > 0$, then $\Psi(B) = \delta(\epsilon_1) + \delta(\epsilon_2)$. This implies for any $\{P_i(B)\}$,

$$\Psi(B) = \sum_{i=1}^{N} \delta(P_i(B)).$$

Now suppose each $P_i(B) = q(B)$, then

$$q(B) = N\delta(q(B)) \implies \delta(q(B)) = \frac{q(B)}{N}.$$

Hence

$$\Psi(B) = \sum_{i=1}^{N} \delta(P_i(B)) = \frac{1}{N}\sum_{i=1}^{N} P_i(B).$$

If all measures are absolutely continuous,

$$\psi(y) = \frac{1}{N} \sum_{i=1}^{N} p_i(B)$$

almost everywhere. □

If a collection of independent experts receive the same weight in every part of the support, then the aggregating measure is simply their average. Of course, this is the same result obtained by minimizing the distance of $\Psi(k)$ from each of a set of finite measures $p_j(k)$ shown in the introduction. The egalitarian aggregate measure has yet another motivation. In the absence of training data to guide the aggregation of a set of probability mass functions, the researcher can appeal to the principle of insufficient reason as described in [4]. They write, "in the absence of evidence to the contrary, all possibilities should have the same initial probability." Placing a uniform prior over the different mass functions implies the egalitarian aggregating measure

$$\Psi(k) = \sum_{j=1}^{N} p_j(k)\pi(j) = \frac{1}{N} \sum_{j} p_j(k).$$

3.3 The Power Means as Aggregating Measures

This section considers a general class of aggregating measures defined by setting Ψ to be proportional to a power mean of the collection of mass functions $\left\{p_{j\cdot}\right\}$. The power mean aggregating measures are unanimous, responsive and exchangeable measures. The unweighted power means, characterized by a parameter r, satisfy one-side of the boundedness constraints. For r sufficiently close to one, the power mean aggregating measure is bounded. Conditions for additional regions are given. If the members of the collection have weights, they can be chosen to approximate the egalitarian aggregating measure. This holds because the power means are continuous functions. In addition, if certain constraints hold on the probability matrix M, the power measure will be bounded for all r sufficiently large. These means there will be bounded aggregating measures formed by power means with very different properties. The next two subsections show properties that a function of probability measures must satisfy to produce an aggregating measure.

3.3.1 Function of the Component Measures

Suppose the aggregate probability of set $B \in \mathcal{B}$ is a function of the collection of measures. Let the aggregate measure be

$$\Psi(B) = F(P_1(B), \ldots, P_N(B))$$

where $F : [0,1]^N \to [0,1]$. To be an aggregate measure, Ψ must be a measure, and therefore, countably additive. This means

$$\Psi(\cup_i E_i) = \sum_i \Psi(E_i).$$

For this to hold,

$$F\left(\sum_i P_1(E_i), \ldots, \sum_i P_N(E_i)\right) = \sum_i F\left(P_1(E_i), \ldots, P_N(E_i)\right).$$

3.3.2 The Power Means

Consider the class of aggregating measures defined by

$$\Psi(y) \propto k(y) = f^{-1}\left(g(f(p_1((y)), \ldots, f(p_N(y))))\right) \tag{3.3.1}$$

where $f : [0,1] \to \mathbb{R}$ is injective and in C^∞ and $g : \mathbb{R}^{|\mathcal{Z}|} \to \mathbb{R}$ in C^∞. A special case of the f-experts is the power mean

$$\Psi(y) \propto k_r(y) = \left(\frac{1}{N}\sum_{i=1}^{N}(p_i(y))^r\right)^{\frac{1}{r}} \tag{3.3.2}$$

for $r \neq 0$.

Proposition 3.3.1. *The aggregating measures proportional to the r^{th} power mean satisfy unanimity. When*

$$\frac{\frac{\partial}{\partial p_i}k_r(y)}{k_r(y)} \neq \frac{\sum_y \frac{\partial}{\partial p_i}k_r(y)}{\sum_y k_r(y)}.$$

they are responsive to changes in the collection $\left\{p_j\right\}_{j=1}^N$ of probability mass functions. Additionally, the aggregating measures are partially bounded in the sense that

$$\Psi(y) \leq \max_j p_j(y)$$

for $r > 1$ *and*

$$\Psi(y) \geq \min_j p_j(y)$$

otherwise.

Proof. To see that $\Psi(y) \propto k_r(y)$ satisfies unanimity suppose $p_i(y) = p(y)$ for all $i \in \{1, \ldots, N\}$. Then

$$k_r(y) = \left(\frac{1}{N} \sum_{i=1}^{N} (p(y))^r \right)^{\frac{1}{r}} = p(y).$$

That this choice of Ψ is responsive follows from the fact that

$$\frac{\partial}{\partial p_i} \Psi(y) = \frac{\frac{\partial}{\partial p_i} k_r(y) \sum_y k_r(y) - k_r(y) \sum_y \frac{\partial}{\partial p_i} k_r(y)}{\left(\sum_y k_r(y) \right)^2}$$

where

$$\frac{\partial}{\partial p_i} k_r(y) = \frac{p_i(y)^{r-1}}{N} \left(\frac{1}{N} \sum_{i=1}^{N} (p_i(y))^r \right)^{\frac{1-r}{r}}.$$

The numerator is nonzero whenever

$$\frac{p_j(y)^{r-1}}{N} \left(\frac{1}{N} \sum_{i=1}^{N} (p_i(y))^r \right)^{1-r} = \frac{\frac{\partial}{\partial p_i} k_r(y)}{k_r(y)} \neq \frac{\sum_y \frac{\partial}{\partial p_i} k_r(y)}{\sum_y k_r(y)}.$$

To establish the inequalities first let $r > 1$. The power mean inequality states for all real numbers r_1 and r_2, $r_1 < r_2$ implies

$$k_{r_1}(y) \leq k_{r_2}(y).$$

Since the egalitarian aggregating measure is the power mean with $r = 1$, for all $r > 1$

$$k_1(y) \leq k_r(y)$$

so that

$$\sum_y k_r(y) \geq 1$$

with equality when $p_i(y) = p(y)$ for all $i \in \{1, \ldots, N\}$ and

$$\max_j p_j(y) \geq \frac{\max_j p_j(y)}{\sum_y k_r(y)} \geq \frac{k_r(y)}{\sum_y k_r(y)}.$$

Thus for $r < 1$,

$$\sum_y k_r(y) \leq 1$$

with equality when $p_i(y) = p_j(y)$ for all $i, j \in \{1, \ldots, N\}$ and

$$\min_j p_j(y) < \frac{\min_j p_j(y)}{\sum_y k_r(y)} \leq \frac{k_r(y)}{\sum_y k_r(y)}.$$

□

Although these aggregating measures are responsive, it is possible for

$$\frac{\partial}{\partial p_j}\Psi(y) < 0$$

so the aggregating measure moves in the opposite direction as mass function p_j. For example, the matrix of probability mass functions

$$M = \begin{bmatrix} 0.300 & 0.500 & 0.050 & 0.150 \\ 0.220 & 0.220 & 0.440 & 0.120 \\ 0.980 & 0.010 & 0.005 & 0.005 \end{bmatrix}$$

gives rise to an aggregating measure proportional to $k_2(y)$ such that the derivative of Ψ with respect to the second probability mass function (the second row of M) at the first state in the outcome space is negative. More exactly,

$$\Psi(1) = \frac{k_2(1)}{\sum_y k_2(y)} \approx .47 > .094 \approx \frac{\frac{\partial}{\partial p_i}k_2(1)}{\sum_y \frac{\partial}{\partial p_i}k_2(y)}.$$

In this case the third row places a large amount of mass on the first state. Hence increases in the second row's assessment of the first state have a small effect relative to the other two rows. Moreover, this demonstrates that aggregating measures proportional to power means are not egalitarian. For a collection of probability mass function defined on a two-element set both sides of the bounding inequality hold.

Proposition 3.3.2. *Suppose $\{P_1, \ldots, P_N\}$ is a set of measures over the two-point set $\{\alpha, \beta\}$. Furthermore, suppose the set is distinct: at least two differ on every set in the sigma algebra of Ω. Let the density functions be positive throughout Ω. Then for any $k \in \Omega$ the power mean of $\{P_1(k), \ldots, P_N(k)\}$ is a bounded aggregating measure.*

Proof. It is first shown that the proposition holds for the case in which $N = 2$ and the support of measures is the two element set $\{\alpha, \beta\}$ and $r > 1$. Since the measures are distinct, $P_1(\alpha) > P_2(\alpha)$ and $P_1(\beta) < P_2(\beta)$ or $P_1(\alpha) < P_2(\alpha)$ and

$P_1(\beta) > P_2(\beta)$. Without loss of generality let $P_\alpha^{\max}$ be the probability assigned to α by the larger of P_1 and P_2. Define

$$A = \left(\frac{1}{2} \left(P_1(\alpha)^r + P_2(\alpha)^r \right) \right)^{1/r}$$

and

$$B = \left(\frac{1}{2} \left(P_1(\beta)^r + P_2(\beta)^r \right) \right)^{1/r}.$$

It shown that

$$P^{\min}(\alpha) < \frac{A}{A+B} < P^{\max}(\alpha)$$

and

$$P^{\min}(\beta) < \frac{B}{A+B} < P^{\max}(\beta).$$

The power mean inequality states that for any m, n with $m > n$, $k_m(y) \geq k_n(y)$. Since the probability mass functions are distinct, if $r > 1$, then the power mean inequality implies $A + B > 1$ so that

$$P^{\max}(\alpha) \geq A \geq \frac{A}{A+B}$$

and

$$P^{\max}(\beta) \geq B \geq \frac{B}{A+B}.$$

Only the lower bounds remain to be shown. Observe that for the two expert case, the probability mass function giving the maximum mass to β will necessarily give the minimal mass to α. That is,

$$P^{\max}(\beta) \geq \frac{B}{A+B} \implies 1 - P^{\min}(\alpha) \geq \frac{B}{A+B} \implies 1 - \frac{B}{A+B} = \frac{A}{A+B} \geq P^{\min}(\alpha).$$

An analogous argument shows

$$\frac{B}{A+B} \geq P^{\min}(\beta).$$

This argument is extended for any number of experts N when $K = 2$. The addition of extra members of the collection of probability mass functions does not change the logic. The mass function placing maximal weight on α places minimal weight on β. When $r < 1$,

$$A + B < 1$$

implies

$$\frac{A}{A+B} \geq A \geq P^{\min}(\alpha)$$

so that

$$1 - P^{\max}(\beta) \leq \frac{A}{A+B} \implies P^{\max}(\beta) \geq \frac{B}{A+B}.$$

Hence it has been that the power means give rise to an aggregate measure for any collection of probability mass functions over a two-point set. The proof follows precisely because of complementarities between the mass functions.

□

When the set of possible outcomes increases, this argument no longer holds. Take the following stochastic matrix as a counterexample. Let the set of states be $\{\alpha, \beta, \gamma\}$.

$$M = \begin{bmatrix} 0.792 & 0.018 & 0.190 \\ 0.058 & 0.676 & 0.266 \end{bmatrix}$$

Then $\Psi(\gamma) \approx .167 < .19 = \min\{P_1(\gamma), P_2(\gamma)\}$. Taking the limit as r increases to infinity provides conditions about boundedness for the range of values over r. The following proposition derives a necessary condition for boundedness for $r \neq 1$.

Proposition 3.3.3 (**Sufficient Conditions for Boundedness**). *Let $\{P_1, \ldots, P_N\}$ be a collection of distinct probability mass functions over the space $\Omega = \{1, \ldots, K\}$ with $K > 2$. If*

$$\frac{\max_j P_j(k)}{\sum_k \max_j P_j(k)} > \min_j P_j(k)$$

then there exists an $r^* > 1$ such that for all $r > r^*$, $\Psi(y) \propto k_r(y)$ is bounded. Analogously, if

$$\frac{\min_j P_j(k)}{\sum_k \min_j P_j(k)} < \max_j P_j(k)$$

then there exists an $r_* < 1$ such that for all $r < r_*$, $\Psi(y) \propto k_r(y)$ is bounded.

Proof. The value of the power mean converges to the maximum probability placed on state y by any of the probability mass functions as r approaches infinity,

$$\lim_{r\to\infty} k_r(y) = \max_j p_j(y),$$

and it converges to the corresponding minimum as r approaches negative infinity,

$$\lim_{r\to-\infty} k_r(y) = \min_j p_j(y).$$

Hence

$$\lim_{r\to\infty} \Psi(y) = \frac{\lim_{r\to\infty} k_r(y)}{\lim_{r\to\infty} \sum_y k_r(y)} = \frac{\max_j p_j(y)}{\sum_y \max_j p_j(y)}. \tag{3.3.3}$$

Let

$$\epsilon_j(y) = \frac{\max_j P_j(y)}{\sum_k \max_j P_j(y)} - \min_j P_j(k).$$

By 3.3.3, there exists an $r^*(\epsilon_j(y)) > 1$ such that $\left|\Psi_r(y) - \frac{\max_j p_j(y)}{\sum_y \max_j p_j(y)}\right| < \epsilon_j(y)$ for all $r \geq r^*$. Then

$$\Psi_r(y) > \frac{\max_j p_j(y)}{\sum_k \max_j p_j(y)} - \epsilon_j(y) = \min_j p_j(y).$$

Analogously, there exists $r_*(\nu_j(y)) < 1$ such that

$$\Psi_r(y) < \frac{\min_j p_j(y)}{\sum_k \min_j p_j(y)} + \nu_j(y) = \max_j p_j(y).$$

□

This result implies coupled with the observation that for every collection of distinct probability mass functions, there exist at least three possible regions such that $\Psi_r(y)$ is bounded: very small values of r, very large values of r and values of r near 1.

3.3.3 Example: Gaussian Experts

For our Gaussian experts suppose

$$p_j(y) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left\{-\frac{(x-\mu_j)^2}{2\sigma_j^2}\right\}. \tag{3.3.4}$$

The choice of the parameters $\left\{(\mu_j, \sigma_j^2)\right\}$ defines the collection. It can be shown that the entropy of each expert is given by

$$\mathcal{E}(p_j) = \frac{1}{2} \ln\left(2\pi \exp\left\{\sigma_j^2\right\}\right). \tag{3.3.5}$$

Hence, expert i is shaper than expert j if and only if $\sigma_i^2 < \sigma_j^2$. The degree of diversity in the collection is given by the sum

$$\sum_{\{i,j:i\neq j\}} H^2(p_i, p_j) = N(N-1) - \sum_{i,j} \int \sqrt{p_i(u)p_j(u)}du.$$

It can be shown that the product of any two univariate densities $p_i(u)$ and $p_j(u)$ is

$$p_{ij}(y) = p_i(y)p_j(y) = \frac{\Lambda_{ij}}{\sqrt{2\pi}\sigma_{ij}} \exp\left\{-\frac{(y-\mu_{ij})^2}{2\sigma_{ij}^2}\right\}$$

where

- $\mu_{ij} = \frac{\mu_i\sigma_j^2+\mu_j\sigma_i^2}{\sigma_i^2+\sigma_j^2}$,
- $\sigma_{ij} = \sqrt{\frac{\sigma_i^2\sigma_j^2}{\sigma_i^2+\sigma_j^2}}$, and
- $\Lambda_{ij} = \frac{1}{\sqrt{2\pi(\sigma_i^2+\sigma_j^2)}} \exp\left\{-\frac{(\mu_i-\mu_j)^2}{2(\sigma_i^2+\sigma_j^2)}\right\}$

Then

$$\sum H^2(p_i, p_j) = N(N-1) - \sum_{i,j} \sqrt{\Lambda_{ij}} \int \sqrt{N(\mu_{ij}, \sigma_{ij}^2)}dy \leq N(N-1) - \sum_{i,j} \sqrt{\Lambda_{ij}} := UB(\mathbf{p}).$$

As each Λ_{ij} grows larger, the total diversity in the collection decreases. That is,

$$\frac{\partial}{\partial\left|\mu_i-\mu_j\right|} UB = \frac{\sqrt{\Lambda_{ij}}}{2} \frac{\left|\mu_i-\mu_j\right|}{\sigma_i^2+\sigma_j^2} > 0.$$

3.4 Approximating the Egalitarian Measure

3.4.1 Aggregating Measures and Exponential Families

Section 3.2 derives many of the benefits of the egalitarian aggregating measure. But, in general, it does not satisfy conjugacy. That is, if p_i belongs to a family of probability mass functions, then the arithmetic mean of the mass functions need not belong to the same family. Exponential families are generally defined by the parameter θ in the sense that each member can be indexed by θ, and the collection can be written as $\{p_\theta(y)\}_{\theta\in\Theta}$. Differences in the elements of the set are associated with different values of the parameters multiplying the sufficient statistics so that each probability mass function weights the sufficient statistics differently. A probability mass function aggregating the

collection with members given by

$$p_\theta(y) \propto \exp\left\{\theta^\top T(y)\right\}$$

is conjugate to the collection is it can be written as

$$\Psi(y) \propto \exp\left\{F(\theta)^\top T(y)\right\}$$

where $T(y)$ is the sufficient statistic common to each probability mass function. The function $F : \Theta^N \to \Theta$ selects a value amongst the various possibilities. Note the mass functions agree on the value of the sufficient statistics, but not in the parameter values. For a finite set $\Theta_n \subset \Theta$, if the members of the collection are not unanimous so that $\theta_i \neq \theta_j$ for some i and $j \neq i$, then the egalitarian aggregating measure is not conjugate to the exponential family as

$$\begin{aligned}\sum_{\theta\in\Theta_n} p_\theta(y) &= \sum_{i=1}^{n} \exp\left\{\theta_i^\top T(y) - A(\theta_i)\right\} \\ &\neq \exp\left\{F(\theta_1, \ldots, \theta_n)^\top T(y) - G(A(\theta_1), \ldots, A(\theta_n))\right\}\end{aligned}$$

for some $y \in \Omega$.

Proof. Is this true? □

Note, however, since exponential families are closed under multiplication,

$$\prod_{\theta\in\Theta_n} p_\theta(y) = \prod_{i=1}^{n} \exp\left\{\theta_i^\top T(y) - A(\theta_i)\right\} = \exp\left\{\left(\sum_{i=1}^{n} \theta_i\right)^\top T(y) - \sum_{i=1}^{n} A(\theta_i)\right\},$$

and there exist functions F and G such that the product of probability mass functions is contained in the exponential family.

The models described in chapter 1 present a nonstandard exponential family. Consider the family of probability mass functions given by

$$p_i(y; \theta) = \exp\left\{\theta^\top T(y|z_i) - A(\theta)\right\}.$$

Rather than differing the value of the parameter θ, the members of collection differ in the value of the sufficient statistic $T_i(y) = T(y|z_i)$. Variation in the members of the collection originates from differences in the feature associated with the observation process, z, such as aspects of the survey instrument structure, contingencies that must be set to make

measurements. As argued for the exponential family indexed by the parameter θ, this exponential family is not closed under addition. In general, the egalitarian aggregating measure over the $\{p_i(y;\theta)\}$ is not a member of the exponential family. But, the aggregating measure proportional to the product of probability mass functions does belong the family.

3.4.2 Approximation with the Weighted Geometric Mean

The insights from the previous section can be used to define an approximation to the egalitarian aggregating measure. To derive an aggregating measure that satisfies boundedness and is conjugate to the collection of probability mass functions whenever they belong to the same family, the egalitarian measure can be approximated by choosing weights in the geometric weighed mean of the probability mass functions. More formally, suppose $p_i \in \mathcal{F}_z(\Omega)$ for all $j \in \{1,\dots,N\}$ where $\mathcal{F}_z(\Omega)$ is an exponential family of probability measures over Ω indexed by the variable z. Choose a function $F : \mathbb{R}^{KN} \to \mathbb{R}^K$ to minimize the difference

$$\left\| \frac{\exp\{\theta^\top F(g_1(y),\dots,g_N(y))\}}{\sum_y \exp\{\theta^\top F(g_1(y),\dots,g_N(y))\}} - \bar{P} \right\|_2^2 \tag{3.4.1}$$

subject to the boundedness constraints

$$\min_j p_j(y) \le \frac{\exp\{\theta^\top F(g_1(y),\dots,g_N(y))\}}{\sum_y \exp\{\theta^\top F(g_1(y),\dots,g_N(y))\}} \le \max_j p_j(y)$$

for all $y \in \mathcal{Y}$ where $\left\{g_j(y)\right\}_{j=1}^N$ is a collection of sufficient statistics. The function F maps N vectors of length K to a single vector of sufficient statistics of length K. One particular choice of functions is

$$F(g_1(y),\dots,g_N(y)) = \frac{1}{\sum_j w_j} \sum_{j=1}^N w_j g_j(y).$$

Observe that this choice is implied by approximating the egalitarian measure $\bar{P}$ by the class of measures proportional to the weighted geometric mean of the members of the collection as

$$\Psi_w(y) \propto \left(\prod_{j=1}^N \exp\left\{\theta^\top g_j(y)\right\}^{w_j} \right)^{\frac{1}{\sum_{j=1}^N w_j}} = \exp\left\{ \theta^\top \frac{\sum_{j=1}^N w_j g_j(y)}{\sum_{j=1}^N w_j} \right\}$$

for all $y \in \mathcal{Y}$.

Proposition 3.4.1. *Let $N = K$. For every point (P_0, w_0) such that*

$$\Psi_{w_0}(y) = \bar{P}_0(y)$$

for all $y \in \Omega$ with $D_w F(P_0, w_0)$ invertible, there exists an implicit function G such that in a neighborhood around (P_0, w_0)

$$G(\bar{P}(y)) = w(y)$$

for all $y \in \Omega$. Hence the egalitarian measure can be approximated exactly.

Proof. Suppose $N = K$. The implicit function theorem as stated in [33] can be applied to show the result. Define the mapping

$$F(P, \mathbf{w}) = \Psi_w(\mathbf{y}) - \bar{P}(\mathbf{y}).$$

Observe there exist points (P_0, w_0) such that $F(P_0, w_0) = 0$. To see this choose P_0 so that $P_0(i, k) = P_0(j, k)$ for every i and j in $\{1, \ldots, N\}$. If the family of exponentials only differs in the sufficient statistic, then $g_i(y) = g(y)$ for all i. Then since the weighted geometric and egalitarian aggregating measures satisfy unanimity,

$$\Psi_w(y) = \frac{\exp\{\theta^\top g(y)\}}{\sum_y \exp\{\theta^\top g(y)\}} = \bar{P}(y).$$

If the matrix $D_w F(P_0, w_0)$ is invertible, the result holds. □

For all $N = K$, the probability can be approximated exactly.

The aggregating measures proportional to a power mean parametrized by $r \in \mathbb{R}$ also satisfies many desirable properties. But, they can give rise to probability mass functions that are not bounded. The probabilities assigned to certain states can vary greatly from those given by the egalitarian aggregating measure. Observe that the distribution of the aggregating mean can be influenced by the introduction of weights $\{w_i\}_{i=1}^N$ on the N probability mass functions. The weighted power mean is given by

$$k_r(y; w) = \left(\frac{1}{N} \sum_{i=1}^{N} w_i p_i(y)^r \right)^{\frac{1}{r}}$$

with $\sum_{i=1}^N w_i = 1$. The egalitarian measure can be approximated by choosing weights to minimize

$$\left\| \frac{1}{N} \sum_{i=1}^{N} p_i(y) - \Psi_r(y; w) \right\|_2^2$$

subject to the constraint that $w \in \Delta_{N-1}$.

3.5 Wasserstein Barycenters as Aggregate Measures

This section explores properties of Wasserstein barycenters over a finite collection of probability mass functions. Let $\Omega = \{1, \ldots, K\}$ be a set of outcomes and let P be a matrix with N rows and K columns with $P_{j\cdot} \in \Delta_{K-1}$ where $P_{j\cdot}$ is the j^{th} row of P. Additionally, let D be the square matrix of distances between the elements in Ω. It will be shown that for all K and N, the constraint qualification holds. That is, there exist matrices $\left\{\Pi_j\right\}_{j=1}^{N}$ such that the constraints implied by the definition of the Wasserstein barycenter hold. Hence an optimal choice exists. This is demonstrated for the case where the distance matrix is trivial,

$$D = \mathbf{1}_K - \mathbb{I}_K$$

and for symmetric matrices generally. Anisotropic distance matrices can also be accommodated.

3.5.1 Definition of the Wasserstein Barycenter

The following definition can be found in [25]. Let (Ω, d) be a metric space for which every measure is a Radon measure and let $\mathcal{P}(\omega)$ be the set of measures on Ω with finite p^{th} moment for some element in Ω.

Definition 5. *A Wasserstein barycenter of N measures $\{P_1, \ldots, P_N\}$ is a minimizer of*

$$f(\Psi) = \frac{1}{N} \sum_{i=1}^{N} W_p^p(\Psi, P_i).$$

over the set of all measures over the state space Ω where W_p is the Wasserstein distance given by

$$W_p = \left(\inf_{\pi \in \Gamma(\mu,\nu)} \int d(x,y)^p d\pi(x,y) \right)^{1/p}.$$

The authors of [2] present results for the support of discrete Wasserstein barycenter for finitely supported measures. That is, each measure places mass on a finite number of points in Euclidean space. The authors show the set of support points is not contained in the union of the support of the individual measures. Hence discrete Wasserstein barycenters in Euclidean space are not bounded by the component probability measures. This section focuses on the case such that Ω is a finite set. The Wasserstein barycenter is any probability mass function Ψ that minimizes

$$\sum_{j=1}^{N} \left(\min_{\pi \in \Gamma(p_j, \Psi)} \sum_{x=1}^{K} \sum_{y=1}^{K} d(x,y)^p \pi(x,y) \right).$$

This section considers two cases. In the first case the distance function is the indicator function; i.e., the distance is zero when the components match and one otherwise. In the second case, elements in Ω are taken to be matrices with distances defined by the Hamming distance.

3.5.2 Unordered Finite Outcome Space

Suppose for any two points $x, y \in \Omega$, the distance between x and y is defined to be

$$d(x,y) = \begin{cases} 1 & x \neq y \\ 0 & x = y \end{cases}.$$

Then the problem is to choose matrices $\Pi^1, \ldots, \Pi^N$ to minimize the objective function

$$\begin{aligned} f(\mathbf{\Pi}) &= \sum_{l=1}^{N} \left(\sum_{y=1}^{K} \sum_{y'=1}^{K} d(y,y')^2 \Pi^l_{y,y'} \right) \\ &= \sum_{l=1}^{N} \sum_{y=1}^{K} \sum_{y' \neq y} \Pi^l_{y,y'} \end{aligned}$$

subject to the feasibility conditions for $j = 1, \ldots, N$ and $l, m = 1, \ldots, N$

$$\begin{aligned} \sum_j \Pi^l_{y,j} &= p_l(y) \\ \sum_j \Pi^l_{j,y} &= \sum_j \Pi^m_{j,y} \end{aligned}$$

and the non-negativity constraints

$$\Pi^l_{y,y'} \geq 0$$

for all l, y, and y'. Note that the objective function is the sum of the off-diagonal probabilities,

$$\sum_k \sum_{k' \neq k} \Pi^l(k,k') = 1 - \sum_k \Pi^l(k,k).$$

Hence minimizing the objective function is equivalent to maximizing the probability mass placed on the diagonal entries of the joint probability density matrices $\left\{\Pi^l\right\}$.

There are KN row constraints of the form $\sum_j \Pi^l_{k,j} = p_l(k)$ and $K(N-1)$ column constraints of the form $\sum_j \Pi^l_{j,k} = \sum_j \Pi^m_{j,k}$. The latter statement is true because conditions between consecutive values of l are sufficient to equalize the

column sums of the matrices $\left\{\Pi^l\right\}$. Choosing N matrices of dimension $K \times K$ implies the selection of K^2N probability masses. Let A be the matrix with $K(2N-1)$ rows and K^2N columns representing the constraints placed on feasible solutions for the Wasserstein barycenter. A feasible solution for the problem can always be found by solving

$$A\mathbf{\Pi} = b \implies \mathbf{\Pi} = A^+ b$$

where A^+ is the Moore-Penrose inverse of A. Then $\Pi^l_{k,k'} = p_l(k)/N$ for all k' and l. The column sums satisfy

$$\sum_k \Pi^l_{k,k'} = \sum_{k=1}^N p_l(k)/N = \frac{1}{N}.$$

This feasible point gives rise to the uniform distribution over the states in Ω. Moreover, since the objective function is continuous and the set of feasible points is compact, there exists a solution to the Finite Wasserstein barycenter problem.

The following proposition introduces a method to identify a feasible collection of matrices yielding a lower value of the objective function. This implies any collection of matrices solving the Wasserstein barycenter problem satisfies

$$\Pi^l_{kk} = \min\left\{\sum_j \Pi^l_{jk}, \sum_j \Pi^l_{kj}\right\}$$

so that the *full diagonal* criteria is a necessary condition for a collection of joint probability mass functions to comprise a Wasserstein barycenter.

Proposition 3.5.1. *Let $\Pi^1, \ldots, \Pi^N$ be $K \times K$ matrices representing the joint distributions between a candidate barycenter Ψ and the probability mass function $p_1, \ldots, p_N$ defined on the set of outcomes $\Omega = \{1, \ldots, K\}$. Suppose the matrices are feasible in the sense that for all $l = 1, \ldots, N$ and $k = 1, \ldots, K$,*

$$\sum_j \Pi^l_{k,j} = p_l(k)$$

and for all $m = 1, \ldots, N$

$$\sum_j \Pi^l_{j,k} = \sum_j \Pi^m_{j,k},$$

but for some l and i

$$\Pi^l_{i,i} < \min\left\{\sum_j \Pi^l_{i,j}, \sum_j \Pi^l_{j,i}\right\}. \tag{3.5.1}$$

Then there exists a feasible $\tilde{\Pi}^l \neq \Pi^l$ *such that*

$$f(\tilde{\Pi}^l, \Pi^{-l}) < f(\Pi^l, \Pi^{-l}).$$

Proof. First observe that $\Pi^l_{k,k}$ can never exceed $\min\left\{\sum_j \Pi_{k,j}, \sum_j \Pi_{j,k}\right\}$. If it does, then the set of matrices is not a feasible solution for the barycenter problem. Next suppose $\Pi^l_{i,i} < \min\left\{\sum_j \Pi_{i,j}, \sum_j \Pi_{j,i}\right\}$. Recall the objective function can be written

$$f(\mathbf{\Pi}) = \sum_{l=1}^{N}\sum_{k=1}^{K}(1 - \Pi^l_{k,k}).$$

Increasing the probability mass on $\Pi^l_{k,k}$ reduces the value of the objective function. There are three cases associated with 3.5.1:

$$\Pi^l_{k,k} < \sum_j \Pi^l_{k,j} < \sum_j \Pi^l_{j,k},$$

$$\Pi^l_{k,k} < \sum_j \Pi^l_{j,k} < \sum_j \Pi^l_{k,j}$$

and

$$\Pi^l_{k,k} < \sum_j \Pi^l_{j,k} = \sum_j \Pi^l_{k,j}$$

Suppose the first inequality holds. Then there exists some $k_1 \neq k$ such that $\Pi_{k,k_1} > 0$. Setting

$$\begin{aligned}\tilde{\Pi}^l_{k,k} &= \Pi^l_{k,k} + \Pi^l_{k,k_1}\\ \tilde{\Pi}^l_{k,k_1} &= \Pi^l_{k,k_1} - \Pi^l_{k,k_1} = 0\end{aligned}$$

transfers probability mass from an off-diagonal entry to a diagonal entry in the matrix while preserving the row sums in the matrices. Since $\sum_j \Pi_{k,j} < \sum_j \Pi_{j,k}$ there exists a set of indices $\left\{k_2^i\right\}$ and numbers in $[0,1]$, $\{\epsilon_i\}$, such that $\sum_i \epsilon_i \Pi_{k_2^i,k} = \Pi_{k,k_1}$. Then setting

$$\begin{aligned}\tilde{\Pi}^l_{k_2^i,k} &= \Pi^l_{k_2^i,k} - \epsilon_i \Pi^l_{k_2^i,k}\\ \tilde{\Pi}^l_{k_2^i,k_1} &= \Pi^l_{k_2^i,k_1} + \epsilon_i \Pi^l_{k_2^i,k}.\end{aligned}$$

preserves the column sums. Hence $\tilde{\Pi}^l$ satisfies the row and column constraints. Since Π_{k,k_1} has been transferred from an off diagonal term to a diagonal entry, the objective function decreases by Π_{k,k_1}. □

Since the diagonal criteria holds for any Wasserstein barycenter, the objective function must satisfy

$$\begin{aligned} f(\mathbf{\Pi}) &= \sum_{l=1}^{N}\sum_{y=1}^{K} 1 - \min\left\{p_l(y), \sum_j \Pi^l_{j,y}\right\} \\ &= \sum_{l=1}^{N}\sum_{y=1}^{K} \max\left\{p_l(y), \sum_j \Pi^l_{j,y}\right\} - \min\left\{p_l(y), \sum_j \Pi^l_{j,y}\right\} \\ &= \sum_{l=1}^{N}\sum_{y=1}^{K} \left| p_l(y) - \sum_j \Pi^l_{j,y} \right|. \end{aligned}$$

subject to the feasibility conditions for $l, m = 1, \ldots, N$ and $k = 1, \ldots, K$

$$\begin{aligned} \sum_j \Pi^l_{j,k} &= \sum_j \Pi^m_{j,k} \\ \sum_j \Pi^l_{k,j} &= p_l(k) \end{aligned}$$

and the non-negativity constraints on the entries of $\Pi^1, \ldots, \Pi^N$. Note that these constraints imply that $\sum_{y'}\sum_y \Pi^l_{yy'} = 1$ so that

$$\Psi(y') := \sum_y \Pi^l_{yy'}$$

is a probability mass function.

3.5.3 Wasserstein Barycenters for Networks

The previous section considers objects with the distance function equal to one when elements match and zero otherwise. This section generalizes the discussion to collection of probability mass functions over the space of binary networks using the Hamming distance Define the distance between two networks to be the Hamming distance between their associated adjacency matrices. That is,

$$d(Y^{(1)}, Y^{(2)}) = \sum_{m,n} \left| Y^{(1)}_{m,n} - Y^{(2)}_{m,n} \right|.$$

Set $p = 2$ so that the barycenter is defined with respect to the quadratic Wasserstein distance. Choose Ψ to minimize the objective function

$$\sum_{j=1}^{N}\left(\min_{\pi_j}\sum_{x=1}^{K}\sum_{y\neq x}^{K}\left(\sum_{m,n}\left|x_{m,n}-y_{m,n}\right|\right)^2\pi_j(x,y)\right).$$

subject to the feasibility conditions for $l, m = 1, \ldots, N$ and $k = 1, \ldots, K$

$$\sum_j \Pi^l_{j,k} = \sum_j \Pi^m_{j,k}$$
$$\sum_j \Pi^l_{k,j} = p_l(k)$$

and the non-negativity constraints on the entries of $\Pi^1, \ldots, \Pi^N$. Note that these constraints imply that $\sum_{y'}\sum_y \Pi^l_{yy'} = 1$ so that

$$\Psi(y') := \sum_y \Pi^l_{yy'}$$

is a probability mass function. As in the case for unordered data, the optimal bivariate probability mass functions π_j place as much mass as possible on the diagonal elements.

3.6 Comparison to Copulas

Thus far the examination of aggregating has been limited to probability mass functions over discrete spaces. Assume the density functions are absolutely continuous so that for any set $B \in \sigma(\mathcal{Y})$,

$$P(y \in B) = \int_B p_l(y)dy.$$

The notions of unanimity and boundedness extend to the continuous case. A continuous aggregating measure over the collection $\{p_l\}_{l=1}^N$ satisfies unanimity if

$$\Psi(B) = P(B)$$

whenever

$$P_l(B) = P_m(B) = P(B)$$

for all $l, m \in \{1, \ldots, N\}$. Additionally, the aggregating measure is bounded if for all $B \in \sigma(\mathcal{Y})$,

$$\min_l P_l(B) \leq \Psi(B) \leq \max_l P_l(B).$$

This section explores the relations between copulas and aggregating measures of continuous probability functions. [76] develops the theory of copulas; from it, the following definition is taken.

Definition 6. *Let $S_1, \ldots, S_{|\mathcal{Z}|}$ be nonempty subsets of $\bar{\mathbb{R}}$, and let H be an $|\mathcal{Z}|$ real function such that the domain of H is the Cartesian product of $\{S_i\}$. Let $B = [\mathbf{a}, \mathbf{b}]$ be an $|\mathcal{Z}|$-box all of whose vertices are in the domain of H. Then the H-volume of B is given by*

$$V_H(B) = \sum_c sign(c) H(c), \tag{3.6.1}$$

or the $|\mathcal{Z}|^{th}$-order difference of H on B

$$V_H(B) = \mathbf{\Delta}_{\mathbf{a}}^{\mathbf{b}} H(\mathbf{t}) = \Delta_{a_{|\mathcal{Z}|}}^{b_{|\mathcal{Z}|}} \cdots \Delta_{a_1}^{b_1} H(\mathbf{t}) \tag{3.6.2}$$

where

$$\Delta_{a_k}^{b_k} H(\mathbf{t}) = H(t_1, \ldots, t_{k-1}, b_k, t_{k+1}, \ldots, t_{|\mathcal{Z}|}) - H(t_1, \ldots, t_{k-1}, a_k, t_{k+1}, \ldots, t_{|\mathcal{Z}|}). \tag{3.6.3}$$

An n-place real function H is $n-$ increasing if $V_H(B) \geq 0$ for all n-boxes B whose vertices lie in Dom(H).

This last definition enables the following definition of a n-dimensional copula.

Definition 7. *A function $C : [0, 1]^{|\mathcal{Z}|} \to [0, 1]$ is copula if for all i in $1, \ldots, \mathcal{Z}$,*

- $C(u_1, \ldots, u_{i-1}, 0, u_{i+1}, \ldots, u_{|\mathcal{Z}|}) = 0$ *and*
- $C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i$*; and*
- *For every* $\mathbf{a}$ *and* $\mathbf{b}$ *in* $\mathbf{I}^N$ *such that $a_l \leq b_l$ for all $l = 1, \ldots, N$,*

$$V_C([\mathbf{a}, \mathbf{b}]) \geq 0.$$

A result, known as the n-dimensional Fréchet-Hoeffding bounds, implies there is exactly one bounded aggregating measure in the set of copulas.

Proposition 3.6.1. *This set of continuous aggregating measures that are also a copula is nonempty. In particular, the copula*

$$C(u_1, \ldots, u_N) = \min_l \left(\int_{-\infty}^{y} p_l(u) du \right)$$

has associated aggregating measure

$$\Psi(B) = P^*(B)$$

where

$$P^*(B) = \min_l \left(\int_B p_l(u)\, du \right).$$

Proof. By construction $\Psi(B)$ is the copula that achieves the Fréchet-Hoeffding upper bound. To see that Ψ is bounded, observe that for all B,

$$\min_l P_l(B) = \Psi(B) \leq \max_l P_l(B).$$

Since this choice satisfies boundedness, it satisfies unanimity. □

Remark 2. *When absolutely continuous an adjusted measure, ϕ, is a density and has an associated CDF given by*

$$F(y) = \int_0^y \phi(y) dy$$

An adjusted distribution's CDF is a integral of a function of densities while the copula is a function of the individual integrals.

Remark 3. *Copulas are defined for multivariate distributions. Note that adjusted distributions are defined by multiple, univariate distributions. That is, evaluate the probability a single realization of a single random variable with a pool of experts.*

Example 1. *Adjusted measures with $|\mathcal{Z}|$-increasing, grounded CDF. Two adjusted measures exist with a CDF that is a near-copula. Observe that both*

$$\phi_1(y) = \frac{\sum_j \left(\prod_{i \neq j} \int_{-\infty}^{y} p(u|z_i) du \right) p(y|z_j)}{\int_{-\infty}^{\infty} \sum_j \left(\prod_{i \neq j} \int_{-\infty}^{y} p(u|z_i) du \right) p(v|z_j) dv}$$

and

$$\phi_2(y) = \frac{\sum_j \left(\prod_{i\neq j} \int_{-\infty}^{y} p(u|z_i)du\right) p(y|z_j)}{\sum_j \left(\prod_{i\neq j} \int_{-\infty}^{y} p(u|z_i)du\right)}$$

satisfy boundedness, although the demonstration of the former is more complicated. Both functions have a grounded and $|\mathcal{Z}|$-increasing CDF, but when all but of the component integrals are equal to one, the copula does not equal the marginal CDF. This means that if all experts have placed all their mass save one, this last expert does not solely determine the CDF. But, if there is an expert who places no mass in the region of study, she becomes the effective dictator. Note that neither function is defined when more than one expert places no mass on the region. Finally, note that

$$C(u_1, \ldots, u_n) = \int_{-\infty}^{y} \sum_j \left(\prod_{i\neq j} \int_{-\infty}^{y} p(u|z_i)du\right) p(v|z_j)dv \tag{3.6.4}$$

is a multiplicative copula.

3.7 A Probabilistic Approach

Rather than define an aggregate measure as an integral of a normalized generalized mean, let the aggregate measure be the probability $y \in B$ conditioned on the set of component measures. That is,

$$\Psi(B) = \mathbb{E}\left[\, Y \in B \,|\, \{P_i(B)\} \,\right]. \tag{3.7.1}$$

For countable additivity to hold, for all partitions of B given by $\left\{E_j\right\}_j$,

$$\mathbb{E}\left[\, Y \in \cup_{j=1}^{M} E_j \,\middle|\, \left\{P_i(\cup_{j=1}^{M} E_j)\right\} \right] = \sum_{j=1}^{M} \mathbb{E}\left[\, Y \in E_j \,\middle|\, \left\{P_i(E_j)\right\} \right]. \tag{3.7.2}$$

Hence the expectation operator chosen must be linear. Note that unlike the kernel method, these models are undefined whenever all measures agree. This can be corrected by setting $\Psi(B)$ equal to the consensus probability. Using a Beta distribution to model $P(B)$ results in the egalitarian aggregate measure.

Example 2. *The Beta distribution is a natural choice to model probabilities. Conceptually, the probability that $\Psi(B)$ is contained in some set S in the Borel set of $[0, 1]$ is modelled as a beta distribution. That is,*

$$\Pr\left\{\Psi(B) \in S \,|\, \{P_i(B)\}\right\} = \int_S \frac{1}{B(\alpha, \beta)} u^{\alpha-1}(1-u)^{\beta-1}\, du.$$

Under the assumption that the measures are independent of one another the maximum likelihood estimates gives

$$\mathbb{E}\left[\Psi(B)\right] = \frac{1}{N}\sum_{i=1}^{N} P_i(B),$$

and

$$\mathrm{Var}\left[\Psi(B)\right] = \frac{1}{N-1}\sum_{i=1}^{N}(P_i(B) - \mathbb{E}\left[\Psi(B)\right])^2.$$

Hence modeling the uncertainty associated with the aggregation of measures via a Beta distribution implies the egalitarian measure.

Chapter 4

A Joint Model for Response and Network Formation

4.1 The Observation Process for Social Networks

Chapter 1 describes a general model for the observation and data-generating processes associated with surveys. In this chapter this line of thought is applied to surveys of social networks. [47] and [48] introduce a theoretical and computational framework for inference in networks surveys. They distinguish between the *design mechanism* and the *out-of-design mechanism*. The former comprises the "part of the observation process under the control of the surveyor," such as the process used to sample dyads and the survey instrument structure. The authors focus on the method used to sampled nodes; in this chapter focuses on decision of sampled dyads to provide responses. A model suggested in [47] accounts for the reliability of reports by supposing respondents erroneously report non-links as links, and vice versa, with probabilities α_1 and α_0. Under these assumptions the sampling mechanism is non-ignorable That is, the joint likelihood of the network and sampling mechanisms is not proportional to the product of their marginal likelihoods.

The present model extends to cases in which the sampling D and response processes R need not give rise to samples satisfying the missing at random condition. For example, a sampled dyad's response behavior may depend on the response decisions of other dyads. In this case the network survey is relational: the sampling process or response process depends on the underlying network structure. Consider two more common examples. In response driven sampling (RDS) (see [41]) the sampling process depends upon the local network of the seeds. Whether a node's edges are sampled depends on the sampling status of alters in some network Y. In egocentric survey sampling in which it

is possible for sampled nodes to not respond, the response decisions may depend on the response decisions of other nodes, in particular, the nodes in local networks.

The extended network survey model can be subsumed into the observation and data-generating processes. The observation process proceeds as follows. A network is represented as a collection of dyadic relations; this conception is represented as an adjacency matrix; the adjacency matrix is sampled by a survey with a specific instrument structure; individuals are sampled from the population of interest; and data is recorded for those sampled nodes who respond to the survey. Let the observation process be $O = \{Z, D, R\}$ where the random variable Z represents the instrument structure, D is the sampling matrix and R is the response matrix. The data-generating process $\mathcal{X} = \{Y, X\}$ contains the adjacency matrix Y, and node and dyad attributes.

Whenever an entire network cannot be surveyed, a sample is taken which introduces two additional sources of uncertainty: uncertainty associated with the sampling and response processes. Let sampling and response be denoted by the random variables D and R, respectively. The sampling and response processes determine which dyads are surveyed and which are missing. If a dyad is not sampled or if a sampled dyad does not respond, then it is missing. The nodes were sampled randomly in the survey data analyzed later in the paper, but the response decisions of the sampled nodes were not made randomly. 4.1 depicts three different models of observation and data-generation. The upper left and right panels differ in how they represent the relationship between the network and the response decisions. In the left panel, the response decisions do not depend upon the network Y. In the right, the response decisions depend upon the response decisions of other nodes to whom the nodes are connected. The right panel presents a graphical model with a social response mechanism; survey completion depends on social structure. The third model specifies that the sampling process depends on social structure and other attributes X. This is a social sampling mechanism. Respondent driven sampling provides an example of this. The instrument structure Z and the response decisions R could be connected by an edge if response decisions are made after viewing the survey instrument structure. Sampled nodes may decide to skip a survey if, for example, its structure is especially complex.

Let V be a set of N nodes. The probability a dyad is sampled can depend on the network Y and the sampling status of other dyads. Whether a dyad responds to a survey depends on the response decisions of other other nodes in Y. The joint probability of the relational system can be decomposed as

$$P_\gamma(Y = y, R = r, D = d|x, z) = P_\theta(Y = y|x, z)P_\psi(D = d|y, x)P_\xi(R = r|d, y, x)$$

where

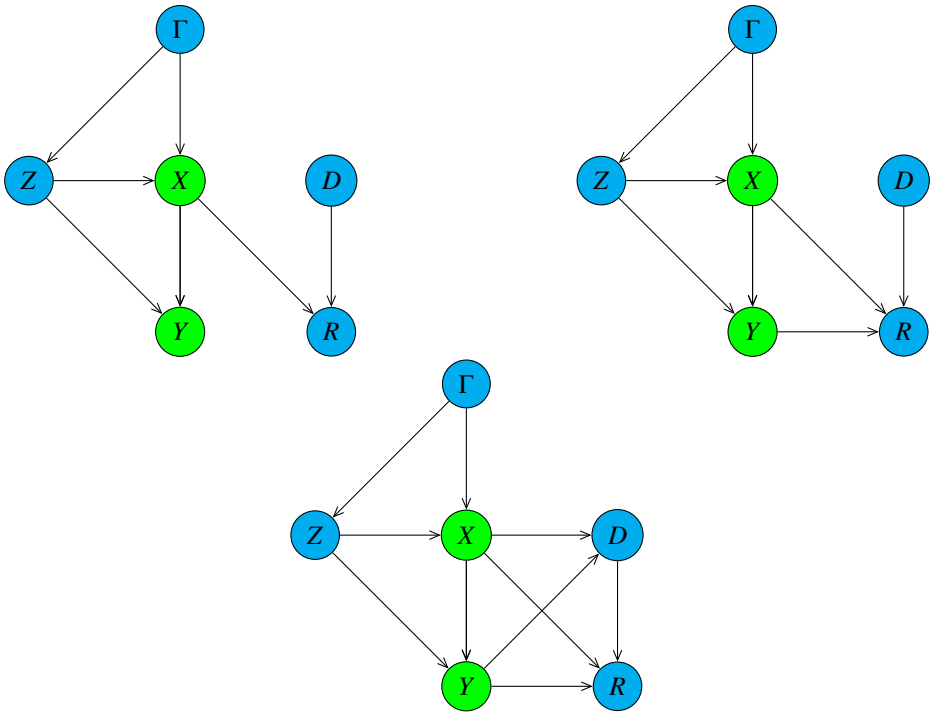

Figure 4.1: Observation (blue nodes) and Data-Generation (green nodes) in a Network Survey.

- D is a binary matrix indicating whether the dyad was sampled,
- R is a binary matrix indicating whether the dyad responded to the survey,
- X is a matrix of dyad attributes possibly derived from node features, and
- $\gamma = (\psi, \xi, \theta)$ is a vector of parameters governing the sampling, response and the network formation process, respectively.

The decomposition reveals the implicit assumption that networks exist prior to the survey process used to elicit them. That is, conditioned on the survey attributes Z and dyad attributes X, the network does not depend on response or sampling. To focus on the social response mechanism, assume ego were randomly sampled so that

$$p(D|Y) = p(D|Y')$$

for all Y, Y' in the space of networks $\mathcal{Y}$. Implications of this assumption are discussed in Section 4.2; generally, sampling parameters can be estimated separately from the response and network formation parameters. A generalization to estimate the three objects simultaneously is discussed in in Section 4.5. Assume the network Y can be represented as an exponential random graph model (ERGM) (see [59] for a good introduction). This framework posits the distribution of the network via sufficient statistics of the demographic and organizational data, the survey instrument structure

and the network itself. If no data is missing, the probability of observing a graph is given by

$$P_\theta(Y = y|x, z) = \frac{\exp\{\theta^\top g(y|x, z)\}}{K(\theta)}$$

where $g(y|x, z)$ is a vector of sufficient statistics and

$$K(\theta) = \sum_{y \in \mathcal{Y}} \exp\left\{\theta^\top g(y|x, z)\right\}$$

is a normalizing constant ensuring the probabilities sum to one. More generally, the survey can obtain information on several relational variables $Y_1, \ldots, Y_K$. In this case more elaborate probability models are required. The appendix provides strong conditions under which the multiple networks can be treated as a single network.

4.1.1 Violation of SUTVA

In an idealized survey setting the stable unit treatment value assumption (SUTVA) would hold (see [84] and [87]). That is, the portion of the network supplied by an individual would not depend upon the assignment of survey instrument structure to other respondents so that for every pair i and j of respondents

$$p(Y_i = y_i|Z_i = z_i, Z_j = z'_j, X) = p(Y_i = y_i|Z_i = z, Z_j = z''_j, X) = p(Y_i = y_i|Z_i = z_i, X) \tag{4.1.1}$$

for all $z_i, z'_j, z''_j \in \mathcal{Z}$ where Y_i is the i^{th} row of the adjacency matrix. But, networks exhibit dependence across dyads. If the treatment alters the ties reported by respondents, then SUTVA cannot hold for an exponential random graph model with arbitrary dependence across dyads. To see this suppose the reports are distributed as an exponential random graph; i.e., let

$$P(Y = y|Z = z) = \frac{\exp\{\theta^\top g(Y|Z, X)\}}{K(\theta)}$$

where $g(Y|Z, X)$ is a vector of sufficient statistics, some of which depend on Z. Then the probability of the i^{th} row of the the adjacency matrix is

$$P(Y_{i*} = y_{i*}|Z = z, X) = \sum_{y_{-i*} \in \mathcal{Y}_{-i*}} \frac{\exp\{\theta^\top g(y_{-i*}, y_{i*}|Z, X)\}}{K(\theta)}.$$

It is clear that $P(Y = y|Z_{-j} = z_{-j}, Z_j = z_j, X) \neq P(Y = y|Z_{-j} = z_{-j}, Z_j = z'_j, X)$ for all j, z_{-j}, z_j and z'_j whenever the coefficients associated with terms in Z are nonzero. Then for all $i \neq j$ in models with dyadic dependence between

dyads incident on (i, j),

$$P(Y_{i*} = y_{i*}|Z_{-j} = z_{-j}, Z_j = z_j) \neq P(Y_{i*} = y_{i*}|Z_{-j} = z_{-j}, Z_j = z'_j)$$

Hence a network survey modeled as an exponential random graph with dyadic dependence cannot satisfy SUTVA. Intuitively, suppose j receives z'_j rather than z_j so that she no longer nominates i. Then under a model with mutuality or transitivity terms, the probability that i would nominate j changes based upon values the corresponding parameters.

4.2 Modeling Social Survey Response

This section presents the sampling and response mechanisms. In network surveys it is possible that D and R depend on the network structure they measure. That is,

$$p(D, R|X, Y) \neq p(D, R|X, Y')$$

for some $Y, Y' \in \mathcal{Y}$. The following subsections explain the model choices for D and R with regards to the network Y.

4.2.1 The Sampling Process

The sampling status D and the response R are random variables. In the most general setting, D and R are random matrices with components taking values in $\{0, 1\}$ so that $D_{ij} = 1$ indicates the dyad ij is sampled and $R_{ij} = 1$ indicates the sampled dyad responds to the survey. Then the dyads in Y can be partitioned by the values of Y into Y_{obs}, Y_{ns} and Y_{nr} representing the observed, non-sampled and sampled but missing dyads, respectively. For the former the probability a dyad is sampled can depend upon the sampling status of other dyads to which it is connected in the underlying network. This holds for *respondent driven sampling* methods; an edge can only be sampled if it is connected to a seed node. The organizational survey, however, followed an ego-centric design. Subjects were selected with probability η. This design is said to be conventional in the sense that

$$P(D = d|Y = y; \eta) = P(D = d; \eta)$$

for all $y \in \mathcal{Y}$. If $D_i = 1$, then dyads $\{ij\}_{j \neq i}$ are sampled. Two immediate examples demonstrate this. [48] shows the probability of observing d given an ego-centric design with probability η is given by

$$p(D = d|Y, \eta) = \eta^{\mathbf{1}^\top s}(1 - \eta)^{(|V|-\mathbf{1})^\top s}$$

where $s \in \{0,1\}^{|V|}$ is the vector of sampling statuses. Alternatively, suppose a sample of size n is drawn from the set of nodes V. Then since a dyad is sampled whenever i or j is chosen, the probability of observing dyad ij is

$$\begin{aligned} p(D_{ij} = 1|Y = y; n) &= \Pr\left\{S_i = 1, S_j = 0\right\} + \Pr\left\{S_i = 0, S_j = 1\right\} + \Pr\left\{S_i = 1, S_j = 1\right\} \\ &= 2\frac{n(n-1)}{|V|(|V|-2)} + \frac{n(n-1)}{|V|(|V|-1)} \end{aligned}$$

and does not depend on the underlying network. Since the sampling matrix is independent of the network, the joint probability mass function can be factored as

$$p(D, R, Y|X, Z) = p(D; \eta)p(R, Y|X, Z, D; \xi, \theta).$$

Inference about the parameters associated with (R, Y) does not require evaluation of $p(D; \eta)$. The next section discusses models for the response decisions that violate this principle.

4.2.2 Regression with Interactions

In many cases response decisions in surveys cannot be coerced. Sampled units may not agree to participate in the survey and those who do may not answer specific questions. For these reasons response behavior belongs to the out-of-design mechanism of the sampling process. When non-response is present, its patterns may be associated with characteristics of the sampled unit and the decisions of other sampled units. These two approaches are presented in 4.1. The former is a standard procedure in survey statistics. Models for survey response generally assume conditional independence between units. That is, after accounting for observable features of the sampled nodes, decisions to complete (or partially complete) the survey are independent; sampled units are not influenced by the response decisions of others. Inference of the data must account for non-response bias by modeling it explicitly. After computing the probability of response for each the units, they are weighted so that response is no longer correlated with the variables of interest in a study. Under these conditions the joint probability of the response vector and network structure is given by

$$p(R, Y; \xi, \theta) = p(R|X; \xi)p(Y|X, Z; \theta).$$

Inference for ξ and θ can be completed separately. Logistic regression can be used to estimate the probability of response for each sampled unit. Whenever nodes can communicate prior to completion of a survey or can formulate the expected responses of their peers, then a social response model should be considered. Under a social response model the probability that unit i responds depends on R_{-i}, the response decisions of the other sampled units. In

complex social structures the dependence structures between sampled units is unclear. Fortunately, when network information is available, it can be integrated into the model.

Exponential Family Model with Neighborhood Structure

Ising and Curie-Weiss models either assume all units depend on others or assume a neighborhood structure based on physical proximity. But, for social networks one's peer group can depend on a confluence of factors. The network structure elicited by the survey serves as a reasonable approximation to the peer group. Response can therefore depend on the sampling statuses and response decisions of nodes to whom they are connected in network Y. Suppose Y is a directed network for the nodes in V. Then the response vector can be modeled as

$$p(R = r|Y = y) \propto \exp\left\{\xi_1 \sum R_i + \xi_2 \sum_{i<j} R_i R_j \max\left\{Y_{ij}, Y_{ji}\right\}\right\}.$$

Response decisions depend on number of total responses and the number of mutual responses between those connected in the directed network. Sufficient statistics derived from the set of features of the sampled units can be added to the model. Then the joint probability of the response vector and network can be decomposed as

$$p(R, Y; \xi, \theta) = p(R|Y, X; \xi)p(Y|X, Z; \theta).$$

consider the model

$$p(R = r|D = d, Y = y; \xi) \propto \exp\left\{\xi^\top h(r|d, y, x)\right\}.$$

If the vector of sufficient statistics depends on other response values, then R is an exponential random graph. Otherwise, response decisions can be modeled as a logistic regression. Let $N_Y(i, k)$ be the set of nodes that are k edges away

from i and define the functions

$$h_1(R|Y,X,Z) = h_1(r|y) = \left(\prod_{i:D_i=1} \sum_{j \in N(i,1):D_j=1} \frac{R_j}{\left| \left\{ j \in N(i,1) : D_j = 1 \right\} \right|} \right)^{\frac{1}{|\{i:R_i=1\}|}}$$

$$h_2(R|Y,X,Z) = h_2(R|y) = \left(\prod_{i:D_i=1} \sum_{j \in N(i,2):D_j=1} \frac{R_j}{\left| \left\{ j \in N(i,2) : D_j = 1 \right\} \right|} \right)^{\frac{1}{|\{i:R_i=1\}|}}$$

$$h_3(R|Y,X,Z) = h_3(R|Y) = \sum_{i:D_i=1} \sum_{j:D_j=1, j \neq i} \max(Y_{ij}, Y_{ji}) f(X_i, X_j) R_i$$

$$h_4(R|Y,X,Z) = h_4(R|X) = \sum_{i:D_i=1} X_i R_i$$

$$h_5(R|Y,X,Z) = h_5(R|Z) = \sum_{i:D_i=1} f(Z)_i R_i$$

to be the vector of sufficient statistics for the response model. The function h_1 is simply the geometric mean of the response rates within each node's first-degree neighborhood. The function h_2 measures the response rate of neighbor's neighbors. Association between node response and covariate values of the dyads with defined by Y is captured by h_3. Functions h_4 and h_5 measure the association between response and covariates without the influence of Y. This is the information available to estimate survey response models in typical cases. Functions h_1 and h_2 induce dependence between the response behavior of sampled nodes; whether i responds to the survey depends on the response of alters in i's neighborhood. Since response depends on Y, it is not ignorable. This can be seen by writing the joint likelihood as

$$\begin{aligned} L(\theta, \xi, n|Y, D, R, X, Z) &\propto p(Y, D, R|X, Z; \theta, \xi, n) \\ &= p(D = d|Y, X, Z; n) p(R = r|D = d, Y, X, Z; \xi) p(Y = y|X, Z; \theta) \\ &= p(D = d; n) p(R = r|d, y, x, z; \xi) p(Y = y|x, z; \theta) \\ &= L(n|D) L(\theta, \xi|y, d, r, x, z). \end{aligned}$$

Correlated Response Model

While the exponential family model has intriguing properties and allows for flexible terms, estimated distributions can exhibit degeneracy. An alternative is to model the response decisions using the social network as a dependence

structure for the unsystematic variation in R. Let

$$R_i^* = \xi^\top X_i + \epsilon_i$$

where ϵ is distributed as a multivariate Normal random variable with mean zero and variance matrix

$$\Sigma(Y)_{ij} = \begin{cases} \sigma^2 & \text{if } i = j \\ \rho_0 & \text{if } i \neq j, Y_{ij} + Y_{ji} = 1 \\ \rho_1 & \text{if } i \neq j, Y_{ij} Y_{ji} = 1 \\ 0 & \text{else.} \end{cases} \tag{4.2.1}$$

In this construction, R^* is a vector of latent variables with

$$p(R_i = 1|X, Y) = p(R_i^* > 0|X, Y)$$

If i and j are related, then their response decisions have nonzero correlation. The model assumes mutual and asymmetric relationships can have relations of different strengths. More elaborate structures are possible. Given a community structure $\{V_1, \ldots, V_M\}$, nodes belonging to the same cluster V_m have correlation ρ_m while relations between members of different groups are ρ_0. The conditional densities are then

$$p(R_i^*|R_{-i}^*, X, \mu, \Sigma(Y)) = N\left(\mu_i + \Sigma(Y)_{i,-i}\Sigma(Y)_{-i,-i}^{-1}(r_{-i} - \mu_{-i}), \Sigma(Y)_{i,i} - \Sigma(Y)_{i,-i}\Sigma(Y)_{-i,-i}^{-1}\Sigma(Y)_{-i,i}\right).$$

Writing the joint density as an exponential family,

$$p(R^*|X, \gamma(\mu, \Sigma(Y)) = \exp\left\{-\frac{1}{2}\left[(R^*)^\top \Sigma(Y)^{-1} R^* - 2\mu^\top \Sigma(Y)^{-1} R^* + \mu^\top \Sigma(Y)^{-1}\mu + \log\left(2\pi \det \Sigma(Y)\right)\right]\right\}$$

implies parameters

$$\gamma = \left(\Sigma(Y)^{-1}\mu, \Sigma(Y)^{-1}\right)$$

for the sufficient statistics R^* and $R^*(R^*)^\top$. Then the joint distribution of (Y, R^*) is also an exponential family.

4.3 Estimation and Missing Data

Joint estimation exploits the Markov chain Monte Carlo maximum likelihood method developed in [39] and [40] and applied to networks in [59]. [48] applies estimation to missing data in which the data is missing at random. The present case extends this estimation methodology to cases in which data is missing not at random.

4.3.1 Complete Data Likelihood

Before analyzing estimation of the relational system for partially observed data, consider the complete data case. Recall that the joint distribution of response behavior and network structure is an exponential random graph with probability mass function

$$P_{(\xi,\theta)}(R = r, Y = y|x, z) = \frac{\exp\{\xi^\top h(r|y, x) + \theta^\top g(y|x, z)\}}{K(\xi, \theta)}$$

where

$$K(\xi, \theta) = \sum_{y\in\mathcal{Y}} \sum_{r\in\mathcal{R}} P_{(\xi,\theta)}(R = r, Y = y|x, z).$$

To simplify notation let $\beta = (\xi, \theta)$ and $f(r, y|x, z) = (h(r|y, x), g(y|x, z))$. Then the ratio of normalizing constants can be written as

$$\begin{aligned}\frac{K(\beta)}{K(\beta_0)} &= \mathbb{E}_{(\beta_0)}\left[\exp\left\{(\beta - \beta_0)^\top f(r, y|x, z)\right\}\right] \\ &= \sum_{y\in\mathcal{Y}} \sum_{r\in\mathcal{R}} \exp\left\{(\beta - \beta_0)^\top f(r, y|x, z)\right\} P_{(\beta_0)}(r, y|x, z).\end{aligned}$$

To approximate this quantity, suppose $\{(r_m, y_m)\}_{m=1}^M$ is a sample from $P_{(\beta_0)}(r, y|x, z)$ and compute

$$\frac{1}{M}\sum_{m=1}^{M} \exp\left\{(\beta - \beta_0)^\top f(r_m, y_m|x, z)\right\}.$$

This term converges to the expectation as $M \to \infty$. Then the ratio of the probability of observing the data $(r_{\text{obs}}, y_{\text{obs}})$ for parameter (ξ, θ) and reference point (ξ_0, θ_0) is

$$\begin{aligned}\log\left(\frac{P_{(\beta)}(r_{\text{obs}}, y_{\text{obs}})}{P_{(\beta_0)}(r_{\text{obs}}, y_{\text{obs}})}\right) &= (\beta - \beta_0)^\top f(r_{\text{obs}}, y_{\text{obs}}|x, z) - \log\left(\frac{K(\beta)}{K(\beta_0)}\right) \\ &\approx (\beta - \beta_0)^\top f(r_{\text{obs}}, y_{\text{obs}}|x, z) - \log\left(\frac{1}{M}\sum_{m=1}^{M} \exp\left\{(\beta - \beta_0)^\top f(r_m, y_m|x, z)\right\}\right).\end{aligned}$$

4.3.2 Partially Observed Data

This section derives an expression that can be maximized to obtain estimates of β based upon the observed likelihood. Edges in the surveyed network are missing if $D_i = 0$ or $R_i = 0$. Then Y can be decomposed into observed dyads, Y_{obs}, dyads associated with nodes who did not respond,

$$\mathcal{Y}_{\text{nr}} = \left\{ Y_{ij} : R_i = 0, D_i = 1 \right\},$$

and dyads associated with unsampled nodes,

$$\mathcal{Y}_{\text{ns}} = \left\{ Y_{ij} : D_i = 0 \right\}$$

so that $Y = Y_{\text{obs}} + Y_{\text{ns}} + Y_{\text{nr}}$. The pair (v, w) is concordant with the observed dyads y_{obs} if $y_{\text{obs}} + v + w \in \mathcal{Y}$. Let

$$\mathcal{Y}(y_{\text{obs}}) = \{u = v + w : y_{\text{obs}} + u \in \mathcal{Y}, v \in \mathcal{Y}_{\text{ns}}, w \in \mathcal{Y}_{\text{nr}}\}$$

be the set of all subnetworks concordant with y_{obs}. Define the conditional distribution of Y given the observed responses r_{obs} and the observed dyads y_{obs}

$$P_\beta(Y_{\text{nr}} = v, Y_{\text{ns}} = w | R = r_{\text{obs}}, Y_{\text{obs}} = y_{\text{obs}}, x, z) = \frac{\exp\{\beta^\top f(r_{\text{obs}}, y_{\text{obs}} + u | x, z)\}}{K(\beta | r_{\text{obs}}, y_{\text{obs}})}$$

for $v + w = u \in \mathcal{Y}(y_{\text{obs}})$ where

$$K(\beta | r_{\text{obs}}, y_{\text{obs}}) = \sum_{u \in \mathcal{Y}(\text{obs})} \exp\left\{\beta^\top f(r, y_{\text{obs}} + u | x, z)\right\}$$

is the conditional normalizing constant. Then reasoning as in Section 4.1 of [48] conclude

$$\begin{aligned} L(\beta | Y_{\text{obs}} = y_{\text{obs}}, R = r_{\text{obs}}) &\propto \sum_{u \in \mathcal{Y}(y_{\text{obs}})} P_\beta(R = r_{\text{obs}}, Y = y_{\text{obs}} + u | x, z) \\ &= \sum_{u \in \mathcal{Y}(y_{\text{obs}})} \frac{\exp\{\beta^\top f(r_{\text{obs}}, y_{\text{obs}} + u | x, z)\}}{K(\beta)} \\ &= \frac{K(\beta | r_{\text{obs}}, y_{\text{obs}})}{K(\beta)}. \end{aligned}$$

Then the likelihood ratio satisfies

$$\frac{L(\beta | r_{\text{obs}}, y_{\text{obs}})}{L(\beta_0 | r_{\text{obs}}, y_{\text{obs}})} \propto \frac{K(\beta | r_{\text{obs}}, y_{\text{obs}}) K(\beta_0)}{K(\beta_0 | r_{\text{obs}}, y_{\text{obs}}) K(\beta)}$$

so that

$$\begin{aligned}
\ell(\beta|r_{\text{obs}}, y_{\text{obs}}) - \ell(\beta_0|r_{\text{obs}}, y_{\text{obs}}) &= \log\left(\frac{K(\beta|r_{\text{obs}}, y_{\text{obs}})}{K(\beta_0|r_{\text{obs}}, y_{\text{obs}})}\right) - \log\left(\frac{K(\beta)}{K(\beta_0)}\right) \\
&= \log\left(\mathbb{E}_{\beta_0}\left[\exp\left\{(\beta-\beta_0)^\top f(r, y|x, z)\right\}|r_{\text{obs}}, y_{\text{obs}}\right]\right) \\
&\quad - \log\left(\mathbb{E}_{\beta_0}\left[\exp\left\{(\beta-\beta_0)^\top f(r, y|x, z)\right\}\right]\right) \\
&\approx \log\left(\frac{1}{M}\sum_{m=1}^{M}\exp\left\{(\beta-\beta_0)^\top f(r'_m, y'_m|x, z)\right\}\right) \\
&\quad - \log\left(\frac{1}{M}\sum_{m=1}^{M}\exp\left\{(\beta-\beta_0)^\top f(r_m, y_m|x, z)\right\}\right)
\end{aligned} \tag{4.3.1}$$

where $\{(r_m, y_m)\}_{m=1}^M$ is a sample from $P_{(\beta_0)}(r, y|x, z)$ and $\{(r'_m, y'_m)\}_{m=1}^M$ is a sample from $P_{(\beta_0)}(Y_{\text{nr}}, Y_{\text{ns}}|r_{\text{obs}}, y_{\text{obs}}, x, z)$. Note that since there is no missing data in the response vector, $\{(r'_m, y'_m)\}_{m=1}^M$ does not contain sampled for r; the observed values are used.

4.3.3 Methods to Obtain Initial Values

As indicated in equation 4.3.1, a value for β_0 is needed. The closer the initial value is to the actual MCMLE values, the better. This subsection discusses two methods to do so: contrastive divergence and maximum pseudolikelihood. The use of these methods in network data is not new; see [59] or (other citation) for a review. But, the use of these methods to solve for initial values for the parameters associated with the response process is not as well studied. In particular, there is an important complication. The dyads in the network are only observed if an individual was sampled and completed the survey. Hence

$$R_i = 0 \vee D_i = 0 \implies Y_{ij} = \text{NA}$$

for all $j \neq i$. When $D_i = 1$ and $R_i = 0$ the dyads $\left\{Y_{ji}\right\}_{j:R_j=1}$ are observed. Hence node i can have no observed mutual ties. If only observed data is used to estimate ξ_0, then the only sufficient statistic available to us are the number of respondents in node i's in-neighborhood,

$$\sum_{j\neq i:R_j=1} Y_{ji}R_j. \tag{4.3.2}$$

In these cases it is unknown whether those who nominated j are known to i. For this reason the response decisions of mutual acquaintances may better capture the data-generating process. If a sampled individual i did not respond to the survey, then their neighborhood of mutual ties (i.e., the number of alters an individual nominates who also nominates the individual in return) is missing. Hence ξ_0 cannot be estimated with the observed data. To side-step this issue, values are imputed for the missing dyads in the network. In the following subsections we apply maximum pseudo

likelihood and contrastive divergence to exponential family models with these two sufficient statistics. For the latter a sample of imputed networks are drawn and then estimates are compute on the samples.

Maximum Pseudolikelihood Estimates

Let R be the response vector of sampled individuals in the population and let $N(i, g(R|Y))$ be the set of responses on which R_i depends conditioned on the response decisions of all other nodes as specified by the sufficient statistics $g(R|Y)$. More precisely define the set to be

$$N(i, g(R|Y)) = \left\{ j \neq i : R_i \not\perp\!\!\!\perp R_j | R_k, k \notin \{j, i\}, g(R|Y) \right\}.$$

Then the pseudolikelihood of the vector of responses is given by

$$\mathcal{L}_{\text{pseudo}}(\xi|R, Y, X) = p(R = r) \prod_i p\left(R_i = 1 | R_j = r_j, \forall j \in N(i, g(R|Y))\right).$$

The pseudolikelihood approximates the true likelihood. In the present case the response models takes the form

$$p_\xi(R = r|Y, X) \propto \exp\left\{\xi^\top g(r|Y, X)\right\}.$$

The pseudolikelihood becomes

$$p_\xi(R = r|Y, X) = \prod_i \frac{\exp\left\{\xi^\top \Delta(g(r|Y, X))_i\right\}}{1 + \exp\left\{\xi^\top \Delta(g(r|Y, X))_i\right\}}$$

where

$$\Delta(g(r|Y, X))_i = g(R_i = 1, R_{-i} = r_{-i}|Y, X) - g(R_i = 0, R_{-i} = r_{-i}|Y, X)$$

is the vector of change statistics. Then

$$\hat{\xi} = \text{argmax}\, \mathcal{L}_{\text{pseudo}}(\xi|R, Y, X)$$

are the maximum pseudolikelihood estimates.

If the only sufficient statistic involving the edges of the network Y are given by 4.3.2, then maximum pseudolike-

lihood estimates can be computed without imputing the missing edges in the network. The change statistic is simply

$$\Delta(g(r|Y,X))_i = (\Delta(g(r|X))_i, Y^{top}_{-i,i} R_{-i})$$

and the estimate is approximately 0.0601.

To use the mutual response within mutual ties, an association social theory suggests should be larger, values for the missing dyads must be imputed. The observed degree distribution of respondents within each department is used to sample a degree for all alters who were not sampled or who did not respond to the survey. This methods assumes the outdegree of nodes that were not sampled or did not respond are distributed according to the random sample of respondents from the same department in the organization. The mean of these estimates can be used as an initial value for ξ_0.

Contrastive Divergence Estimates

Estimation of the model parameters θ and ξ can be performed by maximizing the observed likelihood derived in the previous subsection. Optimization depends on the reference parameter $\beta_0 = (\xi_0, \theta_0)$. Choosing β_0 to be a value near the ML estimator aids convergence. To find such a value I use contrastive divergence learning. Extending the work in [47] and [48], I use a run of contrastive divergence to find biased parameter estimates quickly, $\beta_{\text{CD}} = (\xi_{\text{CD}}, \theta_{\text{CD}})$. Gradient descent of the negative log-likelihood proceeds by updating the parameter to

$$\beta^{(t+1)} = \beta^{(t)} - \eta \left. \frac{\partial}{\partial \beta} \ell(\beta|r,y,x,z) \right|_{\beta=\beta^{(t)}}$$

at each iteration $t = 1, 2, \ldots$ where η is the learning rate and

$$\ell(\beta|r,y,x,z) = \beta^\top f(r,y|x,z) - \log(K(\beta)).$$

Learning model parameters in this way requires computation of the (intractable) normalizing constant, $K(\beta)$. Contrastive divergence ([54] and [18]) provides a method to approximate the gradient without evaluating the normalizing constant. Rather than minimizing the negative log-likelihood, the likelihood is approximated by the contrastive diver-

gence after n steps

$$\begin{aligned}\mathrm{CD}_n &= \mathrm{KL}\left(p_0 \| p_\infty\right) - \mathrm{KL}\left(p_n \| p_\infty\right) \\ &= \sum_{r,y} P_{\beta^0}(r, y|x, z) \log\left(\frac{P_{\beta^0}(r, y|x, z)}{P_{\beta^\infty}(r, y|x, z)}\right) - \sum_{r,y} P_{\beta^n}(r, y|x, z) \log\left(\frac{P_{\beta^n}(r, y|x, z)}{P_{\beta^\infty}(r, y|x, z)}\right)\end{aligned}$$

where $r^{(n)}$ is the realization of the Markov chain after n iterations given the current parameter value ξ. The estimates are biased, but if the bias is small, the model can be used. Variances can be estimated by computing the approximate information. Estimate β_{CD} by iterating the following steps until convergence.

- Draw y' in a Gibbs run.
- Draw $r'|y'$ in a Gibbs run.
- Since $\frac{\partial}{\partial \beta_j} E(r, y; \beta) = -f_j(r, y|x, z)$ where $f_j(\cdot)$ denotes the j^{th} sufficient statistic. Set

$$\beta^{(t+1)} = \beta^{(t)} + \eta\left(f\left(r^0, y^0|x, z\right) - f\left(r^{(n)}, y^{(n)}|x, z\right)\right).$$

4.3.4 MCMLE

If the bias is suspected to be large, a run of MCMC removes it. Then given our contrastive divergence estimates β_0, repeat the following steps until a criteria is met.

1. Draw $\{(r_m, y_m)\}_{m=1}^M$ from $P_{(\beta_0)}(r, y|x, z)$ and $\{(r'_m, y'_m)\}_{m=1}^M$ from $P_{(\beta_0)}(Y_{\mathrm{nr}}, Y_{\mathrm{ns}}|r_{\mathrm{obs}}, y_{\mathrm{obs}}, x, z)$.
2. Update the model parameters to

$$(\xi', \theta') = \underset{\xi,\theta}{\operatorname{argmax}} \left\{\ell(\xi, \theta) - \ell(\hat{\xi}_{CD}, \hat{\theta}_{CD})\right\}$$

   where the approximation to $\ell(\xi, \theta) - \ell(\hat{\xi}_{CD}, \hat{\theta}_{CD})$ is defined by 4.3.1.

The first step is accomplished by MCMC. The complete data and conditional data distributions can both be decomposed so that the network and the response can be sampled serially. Given the current value of $\beta^{(t)} = (\xi^{(t)}, \theta^{(t)})$, sample a network y' from $P_{\theta_0}(y|x, z)$. Then sample a response vector r' from $P_{\xi_0}(r|y', x, z)$. Generally, the Gibbs procedures for each are straightforward for both cases and are given by

$$P_\xi(R_i = 1|R_{-i} = r_{-i}, y, x) = \frac{\exp\left\{\xi^\top \Delta(h(r))_i\right\}}{1 + \exp\left\{\xi^\top \Delta(h(r))_i\right\}}$$

and

$$P_\theta(Y_{ij} = 1|Y_{-ij} = y_{-ij}, x, z) = \frac{\exp\left\{\theta^\top \Delta(g(y))_{ij}\right\}}{1 + \exp\left\{\theta^\top \Delta(g(y))_{ij}\right\}}.$$

4.4 Sampling from the aggregate distribution

Given the estimate of θ from the prior section, simulation of random samples is possible. Divide θ into components not associated with Z, θ_1, and those associated with Z, θ_2. Then the survey instrument free distribution is given by

$$\begin{aligned}\Psi_\theta(Y = y) = \frac{1}{|\mathcal{Z}|}\sum_{z\in\mathcal{Z}} p(y|z) &= \frac{\exp\left\{\theta_1^\top g_1(y|x)\right\}}{|\mathcal{Z}|K(\theta)}\sum_{z\in\mathcal{Z}}\exp\left\{\theta_2^\top g_2(y|z,x)\right\} \\ &= \frac{\exp\left\{\theta_1^\top g_1(y|x) + \log\frac{1}{|\mathcal{Z}|}\sum_{z\in\mathcal{Z}}\exp\left\{\theta_2^\top g_2(y|z)\right\}\right\}}{K(\theta)}.\end{aligned}$$

Under Ψ, Y is no longer an exponential random graph. To simulate a random sample free from the influence of the survey instrument structure, an MCMC run through the network produces a sample. Thus to produce a random network survey sample with respondents given in R_{obs} use the conditional probabilities

$$\Psi_\theta(Y_{ij} = 1|Y_{-ij} = y_{-ij}, x) = \frac{\exp\left\{\theta_1^\top \Delta(g_1(y))_{ij}\right\}\sum_{z\in\mathcal{Z}}\exp\left\{\theta_2^\top g_2(y|z,x)\right\}}{1 + \exp\left\{\theta^\top \Delta(g_1(y))_{ij}\right\}\sum_{z\in\mathcal{Z}}\exp\left\{\theta_2^\top g_2(y|z,x)\right\}}$$

for all i, j, to accept Metropolis steps with probability

$$\min\left\{1, \exp\left\{\theta_1^\top \Delta(g_1(y))_{ij}\right\}\frac{\sum_{z\in\mathcal{Z}}\exp\left\{\theta_2^\top g_2(y_{ij}^+|z,x)\right\}}{\sum_{z\in\mathcal{Z}}\exp\left\{\theta_2^\top g_2(y_{ij}^-|z,x)\right\}}\right\}.$$

Unlike an exponential random graph model, Ψ, requires computation of $\exp\left\{\theta_2^\top g_2(y|z,x\right\}$ for every z increasing the computational burden. Chapter two explores alternatives to the arithmetic mean of probabilities. The geometric mean is a special case of the power means. Under this alternative the probability of a network becomes

$$\begin{aligned}\tilde{\Psi}(y) &\propto \left(\prod_{z\in\mathcal{Z}}\exp\left\{\theta^\top g(y|x,z)\right\}\right)^{1/|\mathcal{Z}|} \\ &\propto \exp\left\{\theta_1^\top g_1(y|x) + \theta_2^\top\frac{1}{|\mathcal{Z}|}\sum_{z\in\mathcal{Z}} g_2(y|x,z)\right\},\end{aligned}$$

which is an exponential random graph. Note the interpretation of $\tilde{\Psi}$, it is the probability all $|\mathcal{Z}|$ models agree normalized by a constant. The metropolis acceptance probability for a single edge is

$$\min\left\{1, \exp\left\{\theta_1^\top \Delta(g_1(y|x))_{ij} + \theta_2^\top \Delta\left(\frac{1}{|\mathcal{Z}|}\sum_{z\in\mathcal{Z}} g_2(y|x,z)\right)_{ij}\right\}\right\}.$$

The geometric mean confers computational benefits in that one only must compute the difference between sufficient statistics, and not the statistics themselves. Any results that hold for exponential random graphs hold for Y under $\tilde{\Psi}$. But, as discussed in chapter two, the probabilities $\tilde{\Psi}$ assigns to a given network may be greater than $\max_z p(y|z,x)$ or less than $\min_z p(y|z,x)$ due to unequal weighting of the models.

Preserving the observed response decisions in the sampling implies the simulations control for Z given R. By the letting the response vector differ from the observed so that there is some i such that Y_{ij} is missing for all $j \neq i$, the samples are from the joint distribution of R and Y. Rather than sampling, the model imputes the values of $\left\{Y_{ij}\right\}_{j\neq i}$. If the number of respondents is held fixed, then the original sample and the imputation will have the same number of sampled dyads. The observed sufficient statistics may be a reasonable target for imputations. Doing so gives the researcher insight into new regions of the network. Given D, X, and Z, for all $t = \{1, \ldots, T\}$, response and the network can be sampled serially.

1. Use MCMC to sample $Y_{\text{obs}}^{(t)}$ and $Y_{\text{nr}}^{(t)}$ with Ψ or $\tilde{\Psi}$. Set $Y_{\text{ns}}^{(t)} = 0$.

2. Use MCMC to sample $R_{\text{obs}}^{(t)}$ and set $R_{\text{mis}}^{(t)} = 0$.

4.4.1 Statistical inference of the counterfactual distribution

If the sampling distribution of the parameters of the estimation routine converges to a normal distribution, then the Delta-method provides a method to estimate the uncertainty in the aggregate probability mass function. Asymptotically, the aggregate probability mass placed at $Y = y$ is normally distributed with variance

$$\text{Var}\left(p^*\left(\widehat{Y = y}|x;\hat{\theta}\right)\right) \approx (\nabla_\theta p^*\left(Y = y|x;\hat{\theta}\right))^\top \hat{\Sigma}(\hat{\theta}) \nabla_\theta p^*\left(Y = y|x;\hat{\theta}\right),$$

In the case of independence between dyads, the marginal distribution of edge ij is given by

$$\delta_{ij}^* = p^*(y_{ij} = 1|\theta, x) = \frac{1}{|\mathcal{Z}|}\sum_{z\in\mathcal{Z}} \frac{\exp\left\{\theta^T \Delta_{ij}(z)\right\}}{1 + \exp\left\{\theta^T \Delta_{ij}(z)\right\}}$$

where $\Delta_{ij}(z) = g(y_{ij}^{+}, x, z) - g(y_{ij}^{-}, x, z)$. Otherwise the marginal distribution is computed by summing the joint distribution over all edges not equal to ij:

$$\delta_{ij}^{*} = p^{*}(y_{ij} = 1|\theta, x) = \frac{1}{|\mathcal{Z}|} \sum_{z \in \mathcal{Z}} \sum_{y_{ij}^{c}} p(Y_{ij} = 1, Y_{ij}^{c} = y_{ij}^{c}|\theta, x, z).$$

The Δ-method provides an approximation to the variance of the edge aggregate probability

$$\text{Var}\left(p^{*}\left(\widehat{y_{ij} = 1}|x; \hat{\theta}\right)\right) \approx \frac{1}{|\mathcal{Z}|^{2}} \left(\sum_{z \in \mathcal{Z}} \frac{\Delta_{ij}(z) \exp\left\{\theta^{T} \Delta_{ij}(z)\right\}}{\left(1 + \exp\left\{\theta^{T} \Delta_{ij}(z)\right\}\right)^{2}} \right)^{\top} \hat{\Sigma}(\hat{\theta}) \sum_{z \in \mathcal{Z}} \frac{\Delta_{ij}(z) \exp\left\{\theta^{T} \Delta_{ij}(z)\right\}}{\left(1 + \exp\left\{\theta^{T} \Delta_{ij}(z)\right\}\right)^{2}}.$$

Compute the sample analogs

$$p^{*}(Y|x) = \frac{1}{S} \sum_{s=1}^{S} p(Y|z_{s}, x) \tag{4.4.1}$$

and

$$\text{Var}\left(p^{*}\left(\widehat{y_{ij} = 1}|x; \hat{\theta}\right)\right) \approx \frac{1}{S^{2}} \sum_{s=1}^{S} \left[\hat{\delta}_{ij}(z_{s})(1 - \hat{\delta}_{ij}(z_{s}))\right] \Delta_{ij}^{T}(z_{s}) \hat{\Sigma} \sum_{s=1}^{S} \left[\hat{\delta}_{ij}(z_{s})(1 - \hat{\delta}_{ij}(z_{s}))\right] \Delta_{ij}(z_{s}). \tag{4.4.2}$$

The estimate of the population variance for edge ij is then

$$\left(\frac{|\mathcal{Z}| - S}{|\mathcal{Z}|}\right) \text{Var}\left(p^{*}\left(\widehat{y_{ij} = 1}|x; \hat{\theta}\right)\right). \tag{4.4.3}$$

Computation of 4.4.2 can be costly if our network or S is large. To do this observe that we can compute it serially by constructing the difference

$$\begin{aligned} \text{Var}\left(p_{(r+1)}^{*}(y_{ij} = 1|x; \hat{\theta})\right) - \text{Var}\left(p_{(r)}^{*}(y_{ij} = 1|x; \hat{\theta})\right) &\approx \left[\hat{\delta}_{ij}(z_{(r+1)})(1 - \hat{\delta}_{ij}(z_{(r+1)}))\right]^{2} \Delta_{ij}^{T}(z_{(r+1)}) \hat{\Sigma} \Delta_{ij}(z_{(r+1)}) \\ &+ 2 \sum_{s=1}^{r} \hat{\delta}_{ij}(z_{s})(1 - \hat{\delta}_{ij}(z_{s})) \Delta_{ij}^{T}(z_{s}) \hat{\Sigma} \hat{\delta}_{ij}(z_{(r+1)})(1 - \hat{\delta}_{ij}(z_{(r+1)})) \Delta_{ij}(z_{(r+1)}). \end{aligned}$$

4.5 Discussion

This chapter presents a method to estimate the network formation and survey response process simultaneously. Sampled individuals' decisions to complete the network survey depend on the completion of those to whom they are connected in the social network. This accomplish by modeling the network as an exponential random graph and the vector indicating survey completion as an exponential family. But, if the sampling process depends on attributes of the pool of possible respondents and the network which connects, then specify the sampling matrix D is distributed

according to an exponential family model,

$$p_\psi(D = d|Y, X) \propto \exp\left\{\psi^\top f(d|Y)\right\}.$$

Then the relational system is distributed as

$$p_\gamma(Y = y, R = r, D = d|X, Z) \propto \exp\left\{\theta^\top g(y|X, Z) + \xi^\top h(r|Y, D, X, Z) + \psi^\top f(d|Y, X)\right\}.$$

4.3.1 can be extended to include ψ. As is the case for ξ and θ, a good value for ψ_0 is required. This can be estimated using maximum pseudolikelihood or contrastive divergence. Note this methodology can be extended to any series of random variables that depend on an underlying network. Network surveys provide an interesting case of the types of problems that might be modeled in this way.

Chapter 5

Lifting the Fog

Chapter 4 develops a general framework for network surveys with non-ignorable dependence between random variables in the observation process and random variables in the data-generating process. This chapter applies to these ideas to a specific network survey with a complex instrument structure. The structure did not vary across respondents. Because sufficient statistics can be created using the order between names - and this feature does vary across dyads - it is possible to estimate the effect of instrument structure. Then sufficient statistics are created to model the response process. The model estimates are used to compute counterfactual quantities of interest as well as compare the actual instrument design to other possibilities.

5.1 Operationalizing Instrument Structure in a Network Survey

Chapter 4 introduces the distinction between the *design mechanism* and the *out-of-design mechanism*. When designing experiments and surveys, researchers choose a structure for data collection instrument. Stimuli must be presented in some sequence; when choosing from a list, the feasible options must be presented in some order. As the complexity of a survey or experiment increases, so does the number of choices facing the researcher. The more complicated the structure, the greater the opportunity for artifacts of the design to interact with the subject's cognitive processes. If changes in the structure lead to changes in the subject's brain in a way that systematically alters the response process, then the structure acts like a lens, influencing the observed information. A sample becomes more dissimilar from the thing the researcher wants to measure as the association between cognitive processes and the instrument structure grows stronger. This section describes the Lifting the Fog (*LTF*) survey.

The organizational survey motivating the methods developed in this chapter was given to a 20% random sample of employees. Sampled individuals were asked to select known others from a roster. After selecting no more than

56 alters, they were asked questions about the nature of their relationships with the identified nodes. Seventy-seven percent of sampled employees completed the survey. See [65] and [66] for additional details. The survey roster and the form onto which the respondents record their responses are components of the survey instrument structure. The organization was divided into divisions. Each division contained a set of departments. The roster preserved this nested structure as it listed divisions numerically. Within divisions, departments were listed. Finally, names of the nodes appeared in alphabetic order within the department blocks. Hence each node has an overall, intra-division and intra-department order. Figure 5.1 illustrates a simple example of this nested structure. The survey instrument structure

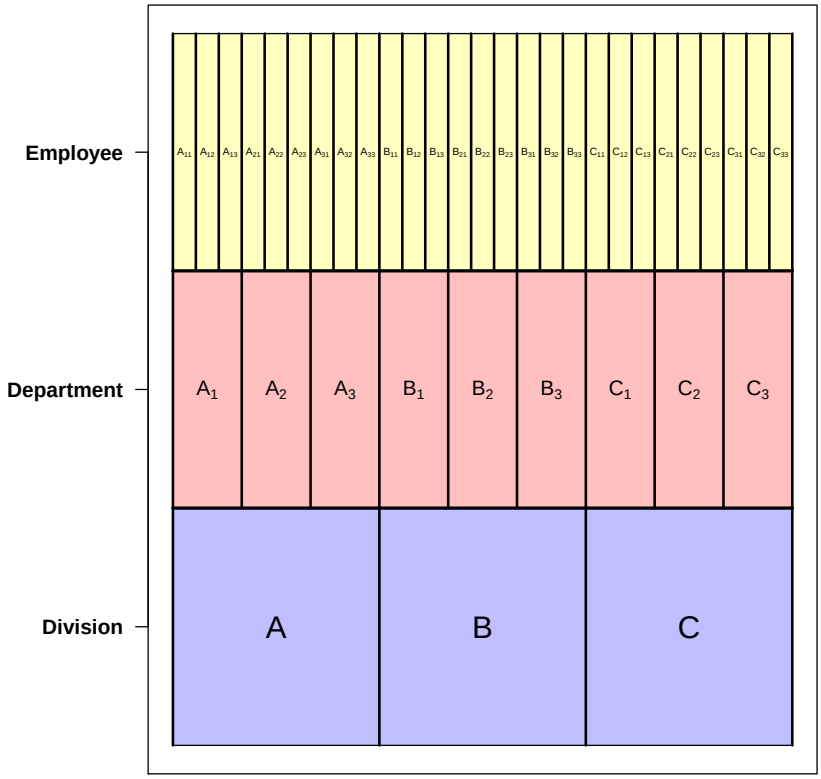

Figure 5.1: Simplified Version of Instrument Structure.

provides several natural hypotheses.

1. The number of response options may have distorted the neighborhood of a respondent. Some may have enumerated more or fewer than they would have given a different instrument structure.

2. Respondents might have used the order in which names appeared on the roster to complete the survey.

The first hypothesis cannot be operationalized. The number of spaces provided to enumerate names did not vary across respondents. But, the second hypothesis can be investigated because sufficient statistics constructed from the surveys

vary across respondents.

5.1.1 Permutation Tests for Survey Instrument Structure

This section presents a series of diagnostic tests to determine whether the survey responses depend on the instrument structure. Permutation tests, [42], can assess whether the observed network data depends on the survey instrument structure z. Let the null hypothesis be that the sampled network does not depend on z. That is, under the null hypothesis,

$$P(Y|z) = P(Y|z')$$

for all $z, z' \in \mathcal{Z}$. If this holds, then any function of sufficient statistic, $F(Y)$, of the network is independent of z as well. That is, for all $z, z' \in \mathcal{Z}$ and $F : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$,

$$P(F(Y)|z) = P(F(Y)|z').$$

Rejecting the null hypothesis that a function of Y is independent of the instrument structure implies that z is not independent of Y. And, therefore, the observation and data-generating processes depend upon one another.

The sufficient statistics interact attributes of the network with the survey instrument structure. These functions are guided by the following substantive hypotheses.

1. Names appearing towards the beginning of the roster are more likely to be nominated by respondents.

2. Names appearing closer to a the respondent's name might have been more likely to be nominated than names far from the respondent's name.

3. Respondents might have been more likely to nominate groups of names close to one another rather than names appearing farther apart from one another.

4. Respondents might be more likely to nominate names which appear towards the beginning of division blocks.

Distance Measures of the Survey Instrument Structure

Respondents may use salient names as reference points in the roster - alters they know well including themselves. Alters may be clustered closer to these reference points.

- The total distance between egos and their associated alters:

$$\sum_i \sum_{j \in N(i,1)} \left| \text{absorder}_i - \text{absorder}_j \right|. \tag{5.1.1}$$

- The total distance between an ego's nominated alters:

$$\sum_i \sum_{j \in N(i,1)} \sum_{k \in N(i,1), k<j} \left| \text{absorder}_j - \text{absorder}_k \right| \tag{5.1.2}$$

- The total absolute order in which a nominated name appears in the roster:

$$\sum_{i,j<i} Y_{ij} \text{absorder}_j. \tag{5.1.3}$$

- The total order in which a name appears in its corresponding division block:

$$\sum_{i,j<i} Y_{ij} \text{divorder}_j. \tag{5.1.4}$$

Each of the sufficient statistics are evaluated after sampling from the set of all possible orderings such that the order in which divisions appear, the order in which departments appear in the division, and the order of names within departments are permuted. Figure 5.2 presents histograms of samples of size $50,000$ - this sample represents just $5.74 \times 10^{-8}\%$ of the number of ways to permute the order in which division blocks appear. Only 5.1.1 does not present strong evidence that the observed data is not independent of the survey instrument structure.

Reciprocity

Reciprocity measures the extent to which $Y_{ij} = 1$ implies $Y_{ji} = 1$ for nodes i and j in $\{v \in V : R_v = 1\}$ and $i \neq j$. If the network and survey instrument structure are independent, then the reciprocity would be equal under different survey instrument structures. In particular, the reciprocity for a given node should not depend on its location in a list. Under the assumption that the network is independent from the survey instrument structure, it must be the case that

$$\Pr\left\{Y_{ji} = 1 | Y_{ij} = 1, z\right\} = \Pr\left\{Y_{ji} = 1 | Y_{ij} = 1, z'\right\}$$

for all $z, z' \in \mathcal{Z}$. Let the set of nodes located in the x^{th} quantile be given by

$$S(x) = \{i \in V : F(z_i) \leq x\}.$$

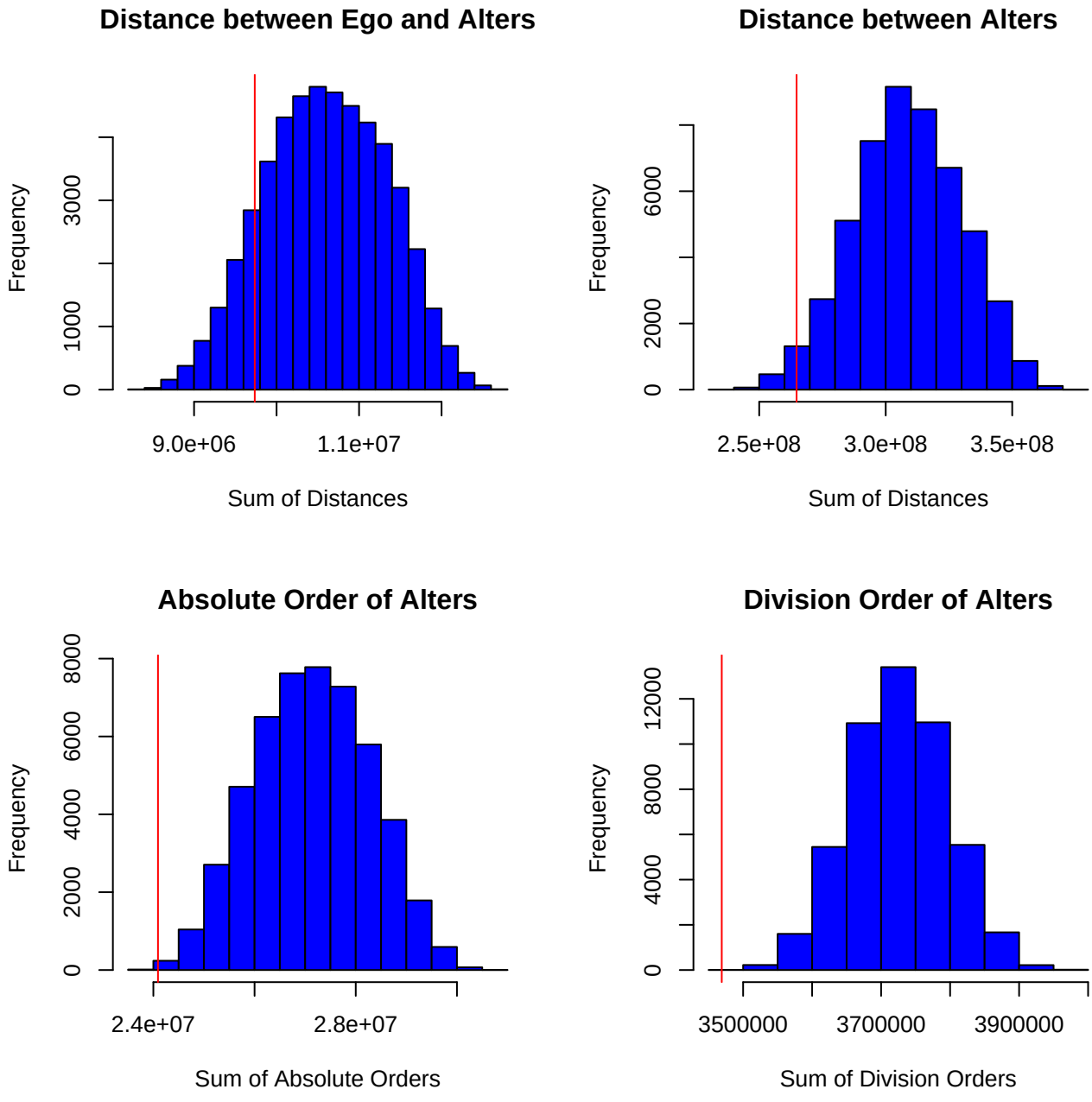

Figure 5.2: Sufficient Statistics for the Roster of Names

Then the rate of reciprocity is equal under different z and, therefore, $S(x)$.

$$F_S(Y) = \frac{\sum_{i:z_i \in S} Y_{ij} Y_{ji}}{\sum_{i:z_i \in S} Y_{ij}}$$

computes the rate of reciprocity for all egos with list locations in the set S and

$$F_S(Y) = \frac{\sum_{i:z_i \in S} Y_{ji}}{|S|}$$

computes the mean in-degree for all egos with list locations in S. When the number of sampled nodes is low, there will be fewer reciprocal ties. In the case of the *Lifting the Fog* survey, implementation of this method is not suggested as the number of reciprocal ties is very low. Reciprocity can be incorporated in ERGM models by introducing a sufficient statistic equal to the number of reciprocal relationships.

Proximity in the Survey Instrument Structure

Intuition suggests that given a survey respondent finds her name in the survey roster, she is more likely to see names near her own. The total distance between nodes that are connected by at least one edge is

$$F(Y, z) = \sum_{i,j:Y_{ij}+Y_{ji}=1} d(z_i, z_j)$$

and the total distance for mutual ties is given by

$$F(Y, z) = \sum_{i,j:Y_{ij}Y_{ji}=1} d(z_i, z_j).$$

To test this compute the total distance between the nodes comprising each dyad for which an edge is present under different survey instrument structures. Tests for three sufficient statistics are implemented: the sum of total distance to alters, the sum of the total distance between alters in the same department and the total distance between alters from different divisions. Figure 5.3 displays the results. In each plot the red line indicates the observed value of the sufficient statistic; the sample p-value is reported above the red line.

Concentration

In chapter 3 a measure for the concentration of a probability measure is discussed. Recall the survey instrument structure is a list of length N. Each node selects M_i for $i = 1, \ldots, N$ alters. Denote this vector $v_i \in \{0, 1\}^N$. Define the

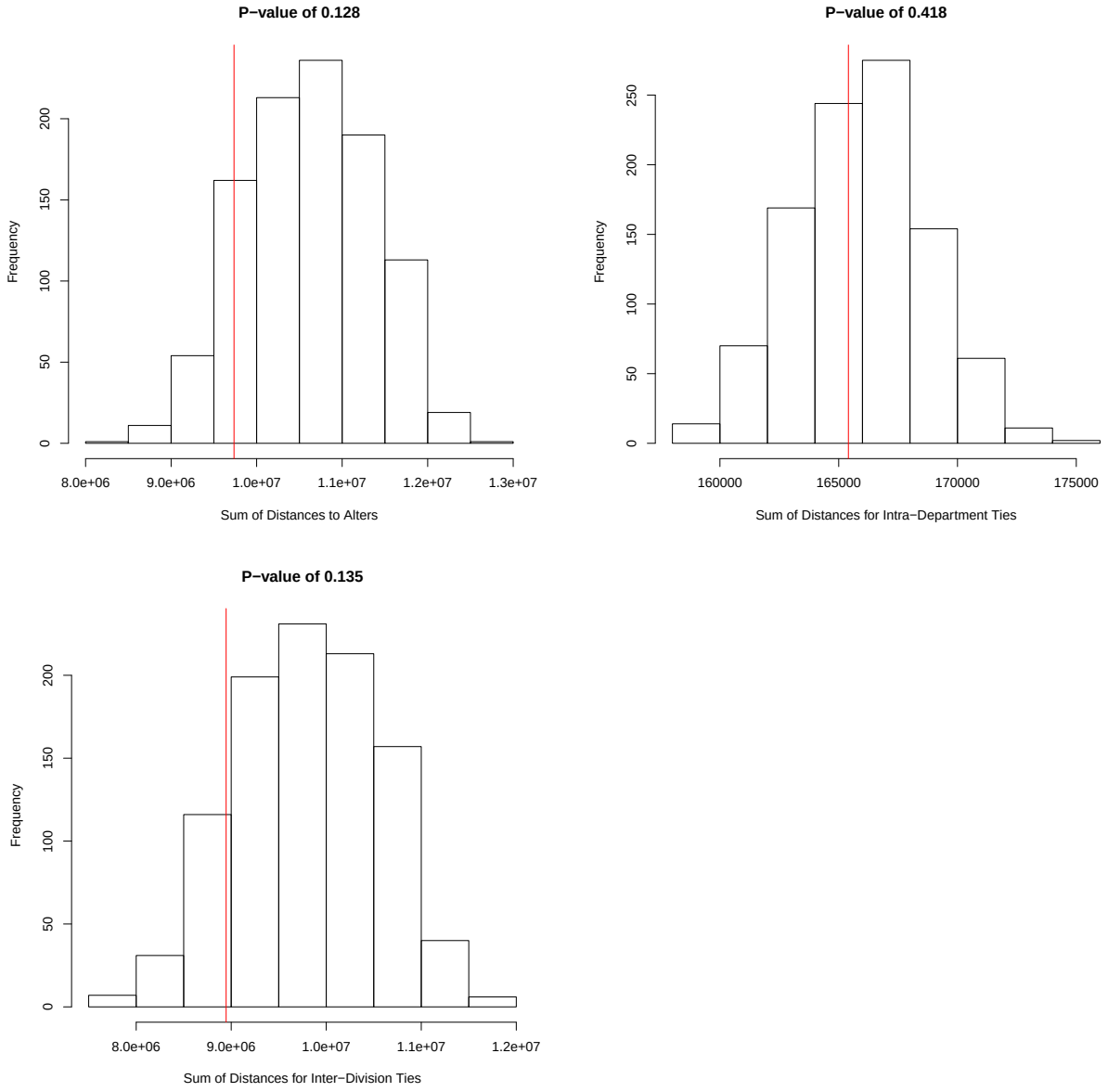

Figure 5.3: Permutation Tests for Dependence of Instrument Structure and Network Structure.

finite concentration to be

$$C_k(v_i) = \sum_x \mathbb{I}\{d(x, v_i) \geq k\}$$

where

$$d(x, v_i) = \min_{y \in A} d(x, y).$$

If the survey process is independent from the instrument structure, then the degree of concentration of v_i does not depend on z. Hence the concentration of v_i satisfies

$$\Pr\{C_k(v_i) \,|\, z\} = \Pr\left\{C_k(v_i) \,\middle|\, z'\right\}$$

for $z, z' \in \mathcal{Z}$. The nested structure of the survey instrument structure can be exploited to test within and between departments and divisions. This must be done as the coincidence of department and division with survey order overstates the effect. The measure of concentration given here is contrasted with the range and standard deviation of v_i. The degree of concentration does not suggest a particularly strong effect.

5.2 The Model

Chapter 4 presents a general model for observation and data-generation in network surveys. This section adapts the general model to the LTF survey data. The assumptions about the Lifting the Fog survey data are encapsulated in figure 5.4. The sets of variables X_1 and X_2 are comprised of node and network attributes such that X_1 contains the data causing the observed survey instrument structure. The data in X_2 is marginally independent of Z. More specifically X_1 contains the department and division of each node. The nesting structure described in section 5.1 determines the order of the options in the survey instrument structure. Individuals within the same department must be adjacent to one another, departments within the same division must be adjacent to one another and divisions were placed in a numerical order. The outcome space of Z given the department and division assignments is all orders that can be constructed by permuting the order of individuals within departments, the order of departments within divisions and the order of divisions amongst themselves. Let $N(z, x)$ be a binary variable indicating whether the z satisfies the nesting structure given in x. Then the survey instrument structure is modeled as

$$p(Z = z | X_1 = x_1) = \frac{|z : N(z, x_1) = 1|}{|\mathcal{Z}|}.$$

The effect of the survey instrument structure on Y is confounded by X_1. Before aggregating over all possible survey instrument structures, the effect of X_1 on Y and the effect of Z on Y must be disentangled. Pearl's so-called backdoor adjustment formula described in [80] can be applied to identify the effect. Conditioning on X_1 blocks the backdoor paths $Z \leftarrow X_1 \rightarrow Y$, $Z \leftarrow X_1 \leftrightarrow X_2 \rightarrow Y$ and $Z \leftarrow X_1 \leftrightarrow X_2 \rightarrow R \leftarrow Y$. The latter of the group contains a collider, R, that is not in the conditioning set. Furthermore, it has no descendants so that none can be contained in the conditioning set.

$$\begin{aligned} p(y|\text{do}(Z = z), X_2 = x_2) &= \sum_{x_1} p(y|Z = z, X_1 = x_1, X_2 = x_2) p(x_1|X_2 = x_2) \\ &= \sum_{x_1} \frac{\exp\{\theta^\top g(Y|X_1, X_2, Z)\}}{K(\theta)} p(x_1|X_2 = x_2). \end{aligned} \tag{5.2.1}$$

If it is believed that Y and X_1 are independent given X_2, then there is no confounding and the effect of the survey instrument structure on the network data is identified, then the only open backdoor path from Z to Y is $Z \leftarrow X_1 \leftrightarrow X_2 \rightarrow Y$. Conditioning on either X_1 or X_2 suffices to block the path and identify the effect of Z on Y. The probability

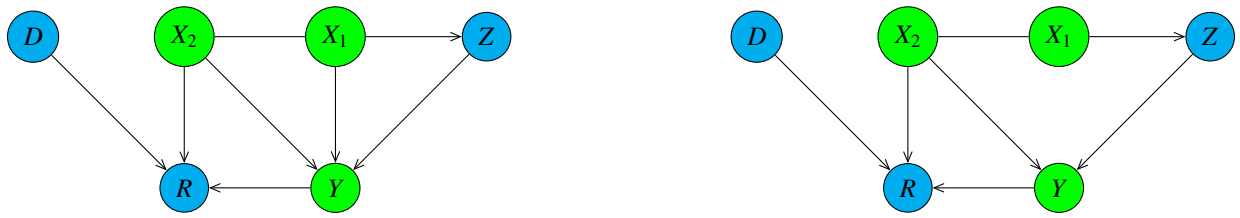

Figure 5.4: Models for Observation (blue nodes) and Data-Generation (green nodes).

measure $p(X_1|X_2)$ models the probability of the assignment of departments to individuals given the information in X_2.

5.2.1 Modeling the Network Data

Exogenous attributes such as age, gender, ethnicity and years of education and attributes related to the individual's role within the institution such as salary, tenure, geographic location, functional area and organization structure are used to explain the variation in the survey nominations. Three terms make estimation of an ergm model with a term based on Z difficult: department, division and mutuality. The previous subsection explains identification of the effect of Z on Y when a covariate X_1 partially determines Z. Practically, if respondents are more likely to nominate names closer to their own, then any variable that is highly correlated with the distance between the respondents and their alters will be correlated with statistics such as $|\text{order}_i - \text{order}_j|$ derived from Z. If i and j belong to the same department $V_{d,g}$, then

the distance between them must satisfy

$$|\text{order}_i - \text{order}_j| \leq \left|V_g\right| - 1$$

and if i and j belong to division $V_{d,\cdot}$ but different departments, then

$$|\text{order}_i - \text{order}_j| \leq \left|V_g\right| - 1$$

Since the order of the departments is determined by the department numbers, these statistics can be furthered bounded. For example if i and j are in adjacent departments, then

$$|\text{order}_i - \text{order}_j| \leq \left|V_g\right| + \left|V_{g+1}\right| - 2.$$

If individuals from the same department are likely to nominate alters from the same department and individuals are more likely to nominate alters listed near their names, then the effect of the order on their nominations is confounded by department. Several terms are possible including

- ordering terms for dyads within large departments,
- ordering terms for dyads between nodes from different departments within large divisions,
- ordering terms for dyads between nodes from different divisions.

Given terms indicating nodes hail from the same department and same division, the ordering terms show a low association with the presence of edges.

A mutuality term in an exponential random graph model can also confound ordering effects. If respondents are more likely to nominate alters listed closer to them on the survey instrument structure, then the graph will display more mutual ties. If ordering effects are asymmetric with respect to the respondent's order on the list or if the ordering effects are not highly mutual the confounding is alleviated. To estimate ordering terms with a mutuality term in the model, one can add the term

$$\sum_i \sum_{j \neq i} Y_{ij} Y_{ji} \left|\text{order}_i - \text{order}_j\right|. \tag{5.2.2}$$

An alternative is to use the dyadcov term in the ergm package. This introduces three terms to the model: the mutuality term in 5.2.2, the lower triangular term

$$\sum_i \sum_{j<i} Y_{ji}(1 - Y_{ij})|\text{order}_i - \text{order}_j|$$

and the upper triangular term

$$\sum_i \sum_{j>i} Y_{ij}(1 - Y_{ji})|\text{order}_i - \text{order}_j|.$$

The lower triangular term captures the interaction of order with receiver and sender behavior. Computation indicates the parameter associated with the mutuality term is positive while the latter two are negative. There are more mutual ties between nodes at a greater distance than expected by random chance. But asymmetric ties are more likely when pairs are near one another. This suggests the following hypothesis: when nominating alters respondents scanned the names farther from their own with purpose, whereas, they nominate alters nearer to their own more haphazardly. Models containing these terms indicate department and division are not highly associated with the presence of edges in the network survey given the values of the other parameters in the model. Two options are possible: omit the terms from the model and interpret the coefficients of the dyadcov terms are the effect of the instrument structure on Y, or include one or both and adjust according to 5.2.1. Small parameter values for department and division term implies small variation between probability mass functions conditioned on different department assignments. Hence the effect on the adjustment may be small in the sense that the absolute difference

$$|p(Y|Z, X_1, X_2) - p(Y|\text{do}(Z), X_2)|$$

is small for many $y \in \mathcal{Y}$.

5.2.2 Model Results

The model for network formation depends on variables capturing organizational structure in the organization, personal attributes of the nodes, organization-related attributes of the nodes such as tenure and salary and the survey instrument structure. (Make a quick plot with confidence intervals for the model terms)

5.3 Aggregating Measure

As argued in section 5.2, the effect of the instrument structure on the data generation process obscures the object of scientific interest. It imbues the resultant data with artifacts from the sampling process. The models examined in the previous subsection demonstrate a clear link between the network and the instrument structure. The data required to produce the term is the matrix of distances between each pair of individuals, D_{obs}. Values for the distance matrix D such that $D \neq D_{\text{obs}}$ give rise to counterfactual conditional probability mass functions. Hence it is possible to estimate the probability distribution for the network survey had it been obscured by a different survey instrument structure.

5.3.1 Exchangeability

Using counterfactual distance matrices requires assumptions about exchangeability. Suppose only a sample of the possible instrument structures were implemented. Then one must assume the data is exchangeable to compute probability distributions for Y conditioned on counterfactual instrument structures. In its simplest version, exchangeability means that the potential outcome Y^z is distributed independently from the value of Z used to observe the data (see [51]).

Dyad ij, if assigned alternate identities kl such that k assumes all of i's information and l does the same for j, would have the same marginal edge probability as dyad kl. This implies there can be no unmeasured variable, such as the actual names of the people on the roster, that affects the probability of nomination. If there were a name effect, then our estimates of the parameters associated with the sufficient statistics containing information from the instrument structure would be conflated with the effect of having a certain name. For example, suppose the names are enumerated in alphabetic order. If names beginning with the letter 'V' are more aesthetically pleasing, then the 'V' effect would be conflated with a greater preponderance to nominate alters listed towards the end of list. Counterfactual survey instrument structures, such as reverse alphabetic order, would contain relatively more nominations from the front of the alphabet. Individuals' names, something not observed in the survey data, do not alter the propensity for nomination in either direction after conditioning on the observed attributes.

Since Z is exchangeable, the effect of Z is causally identified if the positivity assumption holds. That is, there must be a positive probability of receiving each treatment. My assumption that instrument structures receive equal weight assures this is so. Hence Y^z can be interpreted as the counterfactual network whenever $z \neq z_{\text{obs}}$. As before, averaging over the counterfactual probabilities provides an estimate for the expected probability distribution over Y. Then the probability distribution for Y adjusted for the survey instrument structure is given by

$$p^*(Y = y) = \sum_{z \in \mathcal{Z}} p(Y = y | Z = z)\pi(Z = z) = \frac{1}{|\mathcal{Z}|} \sum_{z \in \mathcal{Z}} p(Y = y | Z = z). \tag{5.3.1}$$

In most surveys of social networks, however, the survey instrument structure is not randomly assigned to subjects. Frequently, the survey instrument structure does not vary across respondents. In many settings the distribution of the desired quantity without introducing informed priors or external data sources. Fortunately the relational structure of networks enables the creation of terms that vary across dyads. Hence, even though a single survey instrument structure is observed, a model can account for the influence of the survey instrument structure on the resulting sample. Choosing an unobserved z and transforming it to a vector sufficient statistics makes it possible to simulate counterfactual samples of $Y^{z'}$ for $z' \neq z$.

5.3.2 Types of Exchangeability with Survey Instrument Structure

Sampling from the set of possible orders subject to some set of conditions enables computation of the average distance matrix. Consider the following sampling schemes:

- Sample orders Z by permuting the absolute order in which names are listed subject to no constraints.
- Sample orders Z by permuting the order of division blocks, departments within divisions and names within departments. This preserves the nesting structure, but does not preserve the alphabetic order in which names are listed.
- Sample orders Z by permuting the order of division blocks and departments within the divisions. This method preserves the nesting structure and the order in which names appear within department blocks.

The first method assumes unconditional exchangeability; the second method assumes exchangeability conditioned on the nesting structure and the third method assumes exchangeability conditioned on the nesting structure and the alphabetic order in which names appear in their respective departments. Denote the support of the uniform distribution given each of the assumptions by C_1, C_2 and C_3 respectively. Observe $C_3 \subset C_2 \subset C_1$. Hence there are three probability mass functions defined over the space of all survey instrument structures given by

$$\pi(z|C_t) = \begin{cases} c_t & \text{if } z \in C_t \\ 0 & \text{else} \end{cases}$$

where $c_t = \frac{1}{|C_t|}$. Suppose a prior probability distribution over the three sets of exchangeability assumptions exist. Define the aggregating measure over the three probability mass functions to be

$$\begin{aligned} \pi(z) &= \sum_{t=1}^{3} \pi(z|C_t)\pi(C_t) \\ &= \sum_{t=1}^{3} \frac{\pi(C_1)}{|C_t|} \mathbb{I}\{z \in C_t\}. \end{aligned}$$

To obtain a sample from this distribution, sample from the three pools with probabilities given by the prior.

Given a set of exchangeability assumptions or an aggregating measure over the various sets of exchangeability assumptions, sample $\{z_s\}_{s=1^S}$ and compute the sufficient statistics needed to produce the aggregating measure over networks. A distance matrix is extreme if its components D_{ij} are very large or small relative to the counterfactual population. Let $\bar{D} = \mathbb{E}_z[D]$. Figure 5.5 shows a density plot of the deviations between the observed distance matrix and $\bar{D}$. Then the probability that dyad $D(z)_{ij}$ that depends on the uniformly distributed instrument structure z is less extreme than dyad ij in the observed matrix is given by

$$p\left(\left|D_{ij}^z - \bar{D}_{ij}\right| < \left|D_{ij}^{\text{obs}} - \bar{D}_{ij}\right|\right) \approx \frac{\sum_{z\in\mathcal{Z}} \mathbb{I}\left(\left|D_{ij}^z - \bar{D}_{ij}\right| < \left|D_{ij}^{\text{obs}} - \bar{D}_{ij}\right|\right)}{|\mathcal{Z}|}.$$

If this estimate is large, then dyad has an extreme design relative to the sample. Figure 5.6 shows the probability that for a randomly drawn z, $D_{ij}(z)$ is less extreme than D_{ij}^{obs}. The red bar on the histogram shows all D_{ij}^{obs} such that every z sampled leads to less extreme distances between i and j than in the observed data. These extreme dyads in the observed data are those likely to be affected by artifacts of the instrument structure. Averaging over the dyads provides a measure for the entire matrix

$$\frac{1}{N(N-1)/2} \sum_{i=1}^{N} \sum_{j<i} \frac{\sum_{z\in\mathcal{Z}} \mathbb{I}\left(\left|D_{ij}^z - \bar{D}_{ij}\right| < \left|D_{ij}^{\text{obs}} - \bar{D}_{ij}\right|\right)}{|\mathcal{Z}|}.$$

This provides a measure to compare the total differences between two distance matrices.

5.3.3 Sampling from the Geometric Approximation to the Egalitarian Measure

Given a sample of counterfactual distance matrices, an aggregating measure is defined as

$$\Psi(y|X_2) = \frac{1}{|\mathcal{Z}|} \sum_z p(y|\text{do}(Z = z), X_2 = x_2). \tag{5.3.2}$$

Sampling from this distribution is possible with a Gibbs or metropolized Gibbs algorithm. Since the probability distribution over the network object is no longer an exponential random graph, networks cannot be sampled using standard software. Consider the odds of the presence of an edge

$$\begin{aligned}\frac{\Psi(Y_{ij} = 1, Y_{-(ij)} = y_{-(ij)}|X_2)}{\Psi(Y_{ij} = 0, Y_{-(ij)} = y_{-(ij)}|X_2)} &= \frac{\exp\{\theta_x^\top g(Y_{ij}=1, Y_{-(ij)}=y_{-(ij)}|X_2)\} \sum_{s=1}^S \exp\{\theta_z^\top g(Y_{ij}=1, Y_{-(ij)}=y_{-(ij)}|z_s)\}}{\exp\{\theta_x^\top g(Y_{ij}=0, Y_{-(ij)}=y_{-(ij)}|X_2)\} \sum_{s=1}^S \exp\{\theta_z^\top g(Y_{ij}=0, Y_{-(ij)}=y_{-(ij)}|z_s)\}} \\ &= \exp\left\{\theta_x^\top \Delta(g(y|X_2))_{ij}\right\} \frac{\sum_{s=1}^S \exp\{\theta_z^\top g(Y_{ij}=1, Y_{-(ij)}=y_{-(ij)}|z_s)\}}{\sum_{s=1}^S \exp\{\theta_z^\top g(Y_{ij}=0, Y_{-(ij)}=y_{-(ij)}|z_s)\}}.\end{aligned} \tag{5.3.3}$$

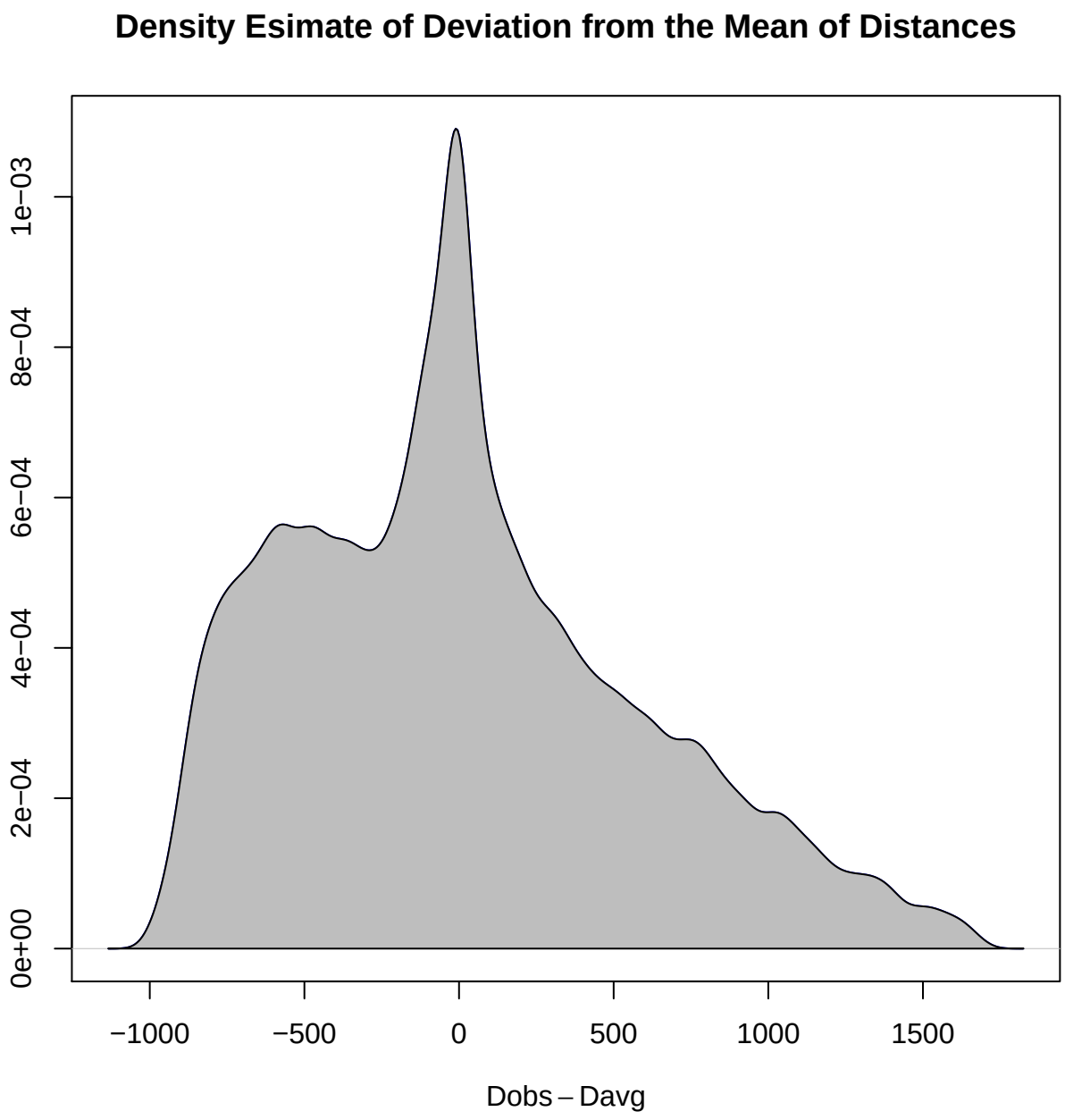

Figure 5.5: Deviations in the Distances between the Observed and Counterfactual Average for all Dyads.

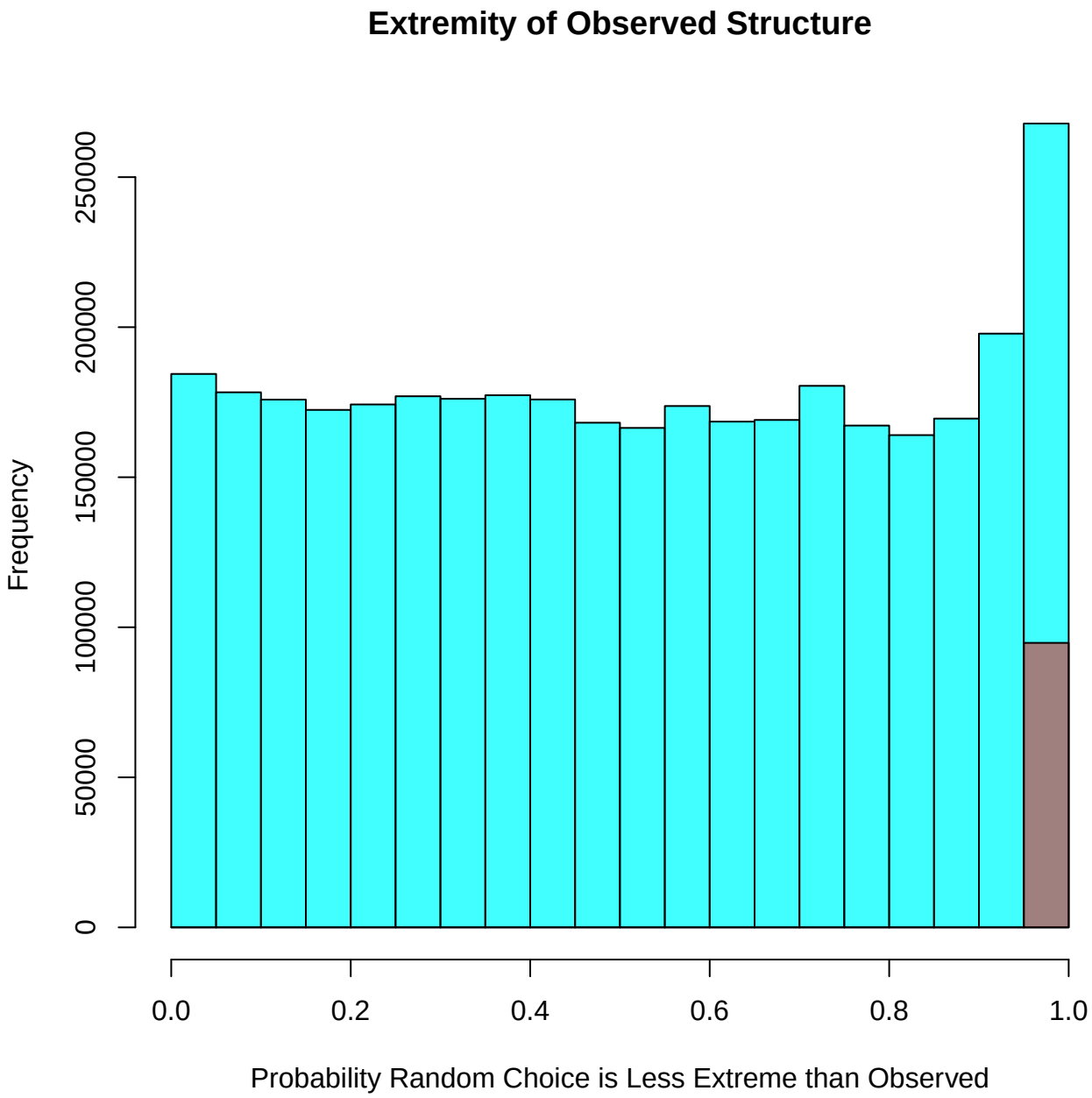

Figure 5.6: Test for Extremity of the Instrument Structure Relative to Counterfactual Population. The red bar indicates all dyads whose distance is most extreme than in any of the randomly sampled instrument structures.

Equation 5.3.3 shows the egalitarian aggregating measure equals an exponential random graph times the ratio of sums of exponential random graphs.

A Gibbs or Metropolis-Hastings sampling routine requires evaluation of $2SK$ sufficient statistics at every step where K is the length of $g(Y|Z)$. An alternative discussed in 3 is use the aggregating measure proportional to the geometric mean of the counterfactual distributions

$$\begin{aligned}\Psi'(y|X_2) &\propto \left(\prod_{s=1}^{S} \exp\left\{\theta_x^\top g(r|X_2) + \theta_z^\top g(r|z_s)\right\}\right)^{1/S} \\ &= \exp\left\{\theta_x^\top g(r|X_2) + \theta_z^\top \frac{1}{S}\sum_{s=1}^{S} g(r|z_s)\right\}.\end{aligned}$$

Unlike the arithmetic mean, the aggregating measure formed by the geometric is an exponential random graph. The sufficient statistics unassociated with the survey instrument structure remain unchanged while the sufficient statistics derived from the survey instrument structure are the mean of the S sampled structures. Hence sampling from the probability mass function is as simple as sampling from any of the counterfactual probability mass functions. Moreover, this choice admits an intuitive interpretation. The aggregating measure proportional to the geometric mean is a product of experts model. The probability is the probability that all S experts agree. Despite its computational advantages, this choice is not guaranteed to be bounded by the probability mass functions that comprise it.

5.3.4 Analyzing the Counterfactual Networks

Networks sampled from the aggregating measured proportional to the geometric mean provide a common reference point for the observed and counterfactual networks. Figure 5.7 displays the population of differences between the observed dyads and the average of draws from the aggregating measure. That, these differences tend to be negative suggests the aggregating mass function places small amounts of mass on dyads without edges in the observed data. The probability that the observed networks is farther from the reference network Y^* that the potential outcome Y^z associated with randomly sampled z is given by

$$p\left(d(Y^z, Y^*) < d(Y^{\text{obs}}, Y^*)\right) \approx \frac{\sum_{z\in\mathcal{Z}} \mathbb{I}\left(\left|D_{ij}^z - \bar{D}_{ij}\right| < \left|D_{ij}^{\text{obs}} - \bar{D}_{ij}\right|\right)}{|\mathcal{Z}|}.$$

Networks sampled from the aggregating measure can be compared to the observed network or networks sampled from the probability mass function conditioned on the observed instrument structure, z_{obs}.

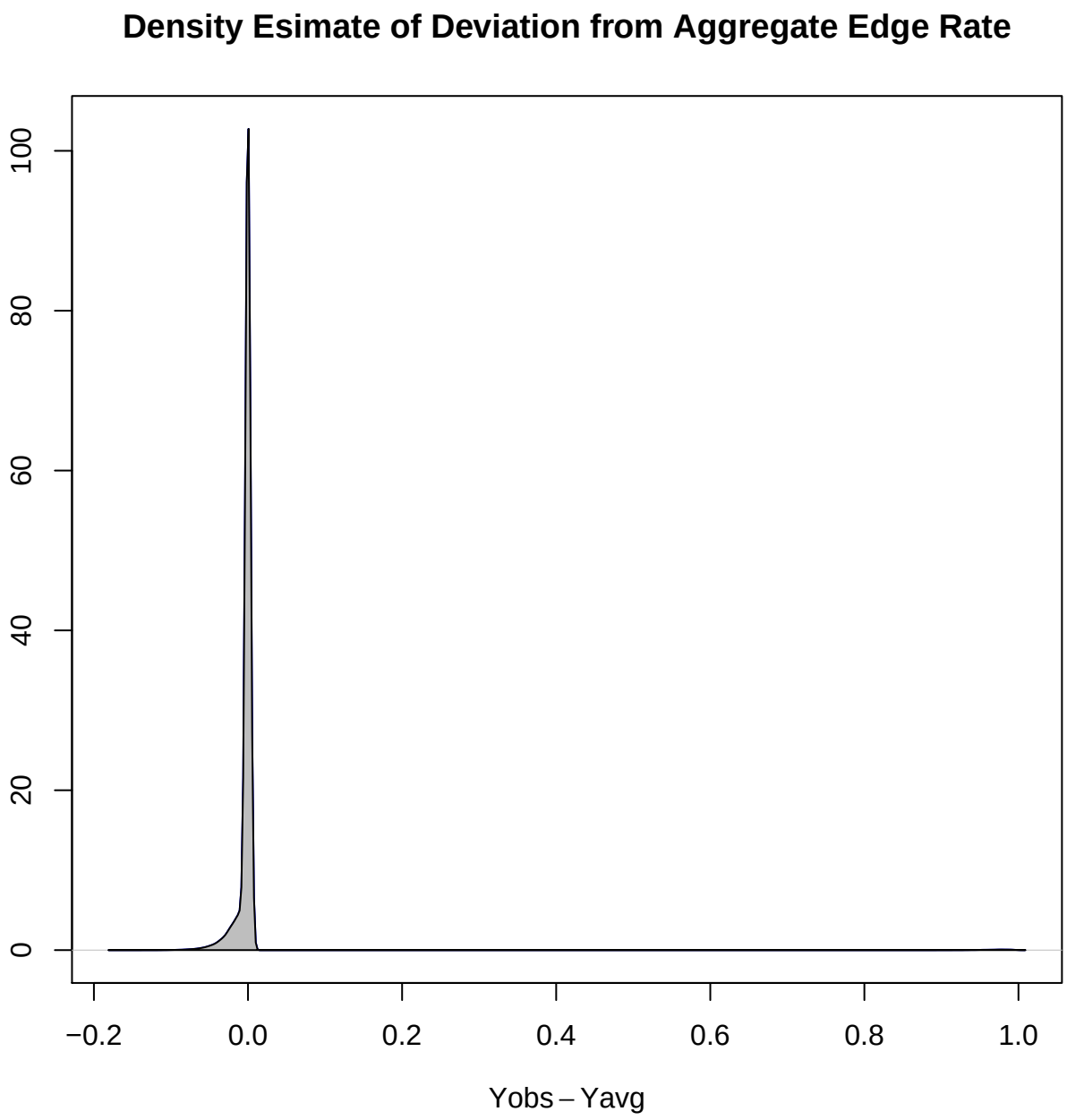

Figure 5.7: Deviations in the Distances between the Observed and Counterfactual Average for all Dyads.

5.4 Appendix

5.4.1 Modeling Multiple Networks

The *LTF* survey motivating this analysis asked respondents whom they knew in the organization. Then respondents answered questions about each of the enumerated relationships such as frequency of communication. The relational system setting permits interdependence between sampled networks: values of dyads in one network depend on the values of dyads in other networks.. For example, the frequency with which i communicates with j can partially determines whether i seeks advice from j. Suppose Y_1 and Y_2 are networks connecting nodes in V. Further suppose there are no dyads connecting Y_1 and Y_2. The networks can be jointly modeled as a single exponential random graph

$$P(Y_1 = y_1, Y_2 = y_2) = K(\theta) \exp\left\{\theta^\top g(Y_1, Y_2)\right\}$$

where $g(Y_1, Y_2)$ is a vector of sufficient statistics derived from the dyads in Y_1 and Y_2. Then the two networks are independent if sufficient statistics are separable. In that case

$$\begin{aligned}
\frac{\exp\{\theta^\top g(Y_1, Y_2)\}}{\sum_{y_1,y_2} \exp\{\theta^\top g(Y_1, Y_2)\}} &= \frac{\exp\left\{\theta_1^\top g(Y_1) + \theta_2^\top g(Y_2)\right\}}{\sum_{y_1,y_2} \exp\left\{\theta_1^\top g(Y_1) + \theta_2^\top g(Y_2)\right\}} \\
&= \frac{\exp\left\{\theta_1^\top g(Y_1)\right\} \exp\left\{\theta_2^\top g(Y_2)\right\}}{\sum_{y_1,y_2} \exp\left\{\theta_1^\top g(Y_1)\right\} \exp\left\{\theta_2^\top g(Y_2)\right\}} \\
&= \frac{\exp\left\{\theta_1^\top g(Y_1)\right\}}{\sum_{y_1} \exp\left\{\theta_1^\top g(Y_1)\right\}} \frac{\exp\left\{\theta_2^\top g(Y_2)\right\}}{\sum_{y_2} \exp\left\{\theta_2^\top g(Y_2)\right\}}
\end{aligned}$$

and the networks can be estimated separately. Suppose now that Y_2 depends on Y_1. The next section provides very conditions under which the networks can not only be estimated separately, but can be condensed to a single network.

5.4.2 Nested Networks

I argue that when certain nesting conditions hold, the set of networks can be simplified to a single object. Suppose $V = \{1, \ldots, N\}$ is a collection of nodes and $Y^{(1)}, Y^{(2)}, \ldots, Y^{(M)}$ are networks comprised of binary edges representing M levels of nested relations. The relation underlying $Y^{(m-1)}$ is a necessary condition for the relation underlying $Y^{(m)}$. That is, for all $m > 1$,

$$Y_{ij}^{(m)} = 1 \implies Y_{ij}^{(m-1)} = 1.$$

Then

$$P(Y^{(m)} = y^m | Y^{(m-1)} = y^{m-1}) = \begin{cases} P\left(Y^{(m)} = y^m\right) & \text{if } Y_{ij}^{(m-1)} - Y_{ij}^{(m)} \geq 0 \forall\, i, j \in V \\ 0 & \text{otherwise} \end{cases}$$

and

$$P\left(Y^{(1)}, Y^{(2)}, \ldots, Y^{(M)}\right) = P(Y^{(1)} = y^1) \prod_{m=2}^{M} P(Y^{(m)} = y^m | Y^{(m-1)} = y^{m-1}).$$

Clearly, the degree can only decrease as m grows. Define the aggregate network

$$Y^0 = \sum_{m=1}^{M} Y^{(m)}$$

as a valued network. Doing so reduces the complexity of the modeling. Observe

$$\sum_{m=1}^{M} Y_{ij}^{(m)} = k \iff Y_{ij}^{(m)} = 1 \,\forall\, m \leq k$$

and

$$\begin{aligned} P\left(\sum_{m=1}^{M} Y^{(m)}\right) &= P\left(Y^{(M)} \middle| \sum_{m=1}^{M-1} Y^{(m)}\right) P\left(\sum_{m=1}^{M-1} Y^{(m)}\right) \\ &= P(Y^{(1)}) \prod_{m=2}^{M} P\left(Y^{(m)} \middle| \sum_{l=1}^{m-1} Y^{(l)}\right) \\ &= P(Y^{(1)}) \prod_{m=2}^{M} P(Y^{(m)} | Y^{(m-1)}) \\ &= P(Y^{(1)}, \ldots, Y^{(M)}). \end{aligned}$$

Hence given full-information, the separate layers can be estimated as a single object. The result also holds if all but the M^{th} level are binary and $Y_{ij}^{(M)} \geq 0$ for all i, j.

Let Y_1 be the binary knowledge network and Y_2 be the non-negative integer-valued communication frequency network. The aggregate network is simply $Y_1 + Y_2$. The assumption $Y_{1,ij} = 0 \implies Y_{2,ij} = 0$ - if i does not know j, then she does not communicate with j - means that the aggregate is functionally equivalent to Y_2 in the sense that the aggregate network is just Y_2 with one added to all components equal to one in the knowledge network. For this reason, a single network will be analyzed throughout the chapter.

5.4.3 Accounting for Unobserved Edges

Now suppose the network is partially observed. Assume that if $Y_{ij}^{(m)}$ is unobserved, then so is $Y_{ij}^{(k)}$ for all $k > m$. Further, suppose that if $Y_{ij}^{(m)} = 0$, then $Y_{ij}^{(k)}$ is not observed for all $k > m$. Since the structure is assumed to be nested, despite the fact the edges are not observed, I set them to zero.

$$\begin{aligned} P(\sum_{m} Y_{\text{obs}}^{(m)}) &= \sum_{\left\{Y_{\text{mis}}^{(m)}\right\}} P(\sum_{m} Y_{\text{obs}}^{(m)}, \sum_{m} Y_{\text{mis}}^{(m)}) \\ &= \sum_{Y_{\text{mis}}^{(1)},\ldots,Y_{\text{mis}}^{(M)}} P\left(Y_{\text{obs}}^{(1)}, Y_{\text{mis}}^{(1)}, \ldots, Y_{\text{obs}}^{(M)}, Y_{\text{mis}}^{(M)}\right) \end{aligned}$$

Hence the equality persists under this pattern of missingness. There are therefore two sources of zeros in each level. A person may know someone, but does not communicate with him or her. Alternatively, a person may not communicate with someone because she or he does not know the alter. These are differentiated in our model as the first and second cases would sum to different values.

Part II: Latent Networks in Group-Level Outcomes

Chapter 6

Review of Model Selection with Hierarchy Constraints

6.1 Introduction

In sports competitions involving teams such as soccer, basketball, lacrosse and hockey, two lineups compete against one another. These sports all feature objects used to score. The period of time in which a team assumes control of the object is defined to be a possession. Each possession can have many outcomes such as length of time, number of teammates who control the object (number of passes) and total movement of players, but the most essential is whether the offensive team scores. Researchers in this area created the adjusted plus-minus model which has become a folk method in sports statistics. Using the possessions as individual observations, the models estimate parameters for each of the players on the court. Penalized versions improve the predictive efficacy of the model while Bayesian models are used to set priors to the model terms, possibly from prior seasons of play.

6.2 Bayesian Model Selection with Hierarchy Constraints

Let V be an arbitrary matrix of variables and G be the space of all models - the set of vectors of length $N = \sum_{k=0}^{|V|} \binom{|V|}{k}$ with each index equal to one or zero. A model, M, is a vector in G. Let $\pi_i(M) = \Pr\{M_i = 1 | M_{-i} = m_{-i}, G\}$ be the probability that the i^{th} term is included in the model. For each term m_i, there exists a family $\mathcal{F}(m_i)$ which contains the lower-order that comprise m_i. Define $\mathcal{P}(m_i) \subset \mathcal{F}(m_i)$ to be the parents of m_i. [89] follows [22] by assuming (1)

conditional independence of terms of the same order and (2) the probability of a term depends on its parents. The probability that term i is included in the model is

$$\pi_i(M) = \Pr\{M_i = 1 | \mathcal{P}(M_i), G\}$$

and the probability of a model M is

$$\pi(M|G) = \prod_i \pi_i(M)^{m_i}(1 - \pi_i(M))^{(1-m_i)}.$$

Under the assumption of strong hierarchy, $m = 0$ for any $m \in \mathcal{P}(m_i)$ implies $m_i = 0$ while weak hierarchy imposes that if $m = 0$ for all $m \in \mathcal{P}(m_i)$, then $m_i = 0$. Several specifications for $\pi(\cdot)$ are possible.

6.2.1 Uniform Prior with a Hierarchy Condition

All terms that are not constrained to be zero by our hierarchy condition share the same probability of inclusion; that is, for all distinct i and j such that m_i and m_j are not prohibited by a hierarchy condition, $\pi_i(M) = \pi_j(M)$. Set $\pi_i(M) \sim \text{Beta}(a, b)$ for all i. Let $N(M|\text{HC}) \leq N$ denote the number of terms not constrained to be zero by the hierarchy condition. Similarly, define $G(\text{HC})$ to be the set of binary vectors which satisfy the specified hierarchy constraint. Then

$$\begin{aligned}
p(M|G(\text{HC})) &= \int_0^1 p(M|\pi, G(\text{HC}))p(\pi|G(\text{HC}))d\pi \\
&= \int_0^1 \left(\prod_{i \in G(\text{HC})} \pi^{m_i}(1-\pi)^{(1-m_i)} \right) \frac{\pi^{a-1}(1-\pi)^{b-1}}{B(a,b)} d\pi \\
&= \frac{1}{B(a,b)} \int_0^1 \pi^{\sum_i m_i + a - 1}(1-\pi)^{|G(\text{HC})| - \sum_i m_i + b - 1} d\pi \\
&= B\left(\sum_{i \in G(\text{HC})} m_i + a, |G(\text{HC})| - \sum_{i \in G(\text{HC})} m_i + b \right) / B(a,b).
\end{aligned}$$

Hence all models with the same number of terms exhibit the same probability.

6.2.2 Independence Prior with a Hierarchy Condition

Suppose the terms are distributed independently with $\pi_i \sim \text{Beta}(a_i, b_i)$. Then the probability of a model becomes

$$\begin{aligned}
p(M|G(\text{HC})) &= \int_{[0,1]^{|G(\text{HC})|}} p(M|\boldsymbol{\pi}, G(\text{HC}))p(\boldsymbol{\pi}|G(\text{HC}))d\boldsymbol{\pi} \\
&= \int_0^1 \cdots \int_0^1 \left(\prod_{i \in G(\text{HC})} \pi_i^{m_i}(1-\pi_i)^{(1-m_i)} \right) \left(\prod_{i \in G(\text{HC})} \frac{\pi_i^{a_i-1}(1-\pi_i)^{b_i-1}}{B(a_i, b_i)} \right) d\pi_1 \ldots d\pi_{|G(\text{HC})|} \\
&= \prod_{i \in G(\text{HC})} \frac{1}{B(a_i, b_i)} \int_0^1 \pi_i^{m_i+a-1}(1-\pi_i)^{1-m_i+b-1} d\pi_i \\
&= \left(\prod_{i \in G(\text{HC}) \cap \{i:m_i=1\}} \frac{B(a_i+1, b_i)}{B(a_i, b_i)} \right) \left(\prod_{i \in G(\text{HC}) \cap \{i:m_i=0\}} \frac{B(a_i, b_i+1)}{B(a_i, b_i)} \right) \\
&= \left(\prod_{i \in G(\text{HC}) \cap \{i:m_i=1\}} \frac{B(a_i, b_i)(a_i/(a_i+b_i))}{B(a_i, b_i)} \right) \left(\prod_{i \in G(\text{HC}) \cap \{i:m_i=0\}} \frac{B(a_i, b_i)(b_i/(a_i+b_i))}{B(a_i, b_i)} \right) \\
&= \left(\prod_{i \in G(\text{HC}) \cap \{i:m_i=1\}} \frac{a_i}{a_i+b_i} \right) \left(\prod_{i \in G(\text{HC}) \cap \{i:m_i=0\}} \frac{b_i}{a_i+b_i} \right).
\end{aligned}$$

6.2.3 Order Prior with a Hierarchy Condition

A more nuanced assumption specifies for all distinct i and j such that $\text{Order}(m_i) = \text{Order}(m_j)$, $\pi_i(M) = \pi_j(M) = \pi_o$. That is, all terms with the same order have equal inclusion probabilities given the hierarchy condition. In particular, assume $\pi_o \sim \text{Beta}(a_o, b_o)$. Let O denote the maximal order of M and $G_o(\text{HC}) = G(\text{HC}) \cap \{i : \text{Order}(m_i) = o\}$. Then

$$\begin{aligned}
p(M|G(\text{HC})) &= \int_{[0,1]^{|G(\text{HC})|}} p(M|\boldsymbol{\pi}, G(\text{HC}))p(\boldsymbol{\pi}|G(\text{HC}))d\boldsymbol{\pi} \\
&= \int_0^1 \cdots \int_0^1 \left(\prod_{i \in G(\text{HC})} \pi_i^{m_i}(1-\pi_i)^{(1-m_i)} \right) \left(\prod_{i \in G(\text{HC})} \frac{\pi_i^{a_i-1}(1-\pi_i)^{b_i-1}}{B(a_i, b_i)} \right) d\pi_1 \ldots d\pi_{|G(\text{HC})|} \\
&= \prod_{o=1}^{O} \int_0^1 \cdots \int_0^1 \left(\prod_{i \in G(\text{HC}) \cap \{i:\text{Order}(m_i)=o\}} \pi_o^{m_i}(1-\pi_o)^{(1-m_i)} \right) \frac{\pi_o^{a_o-1}(1-\pi_o)^{b_o-1}}{B(a_o, b_o)} d\pi_1 \ldots d\pi_O \\
&= \prod_{o=1}^{O} \frac{1}{B(a_o, b_o)} \int_0^1 \pi_o^{\sum_{i \in G_o(\text{HC})} m_i + a_o - 1}(1-\pi_o)^{|G_o(\text{HC})| - \sum_{i \in G_o(\text{HC})} m_i + b_o - 1} d\pi_o \\
&= \prod_{o=1}^{O} B\left(\sum_{i \in G_o(\text{HC})} m_i + a_o, |G_o(\text{HC})| - \sum_{i \in G_o(\text{HC})} m_i + b_o \right) / B(a_o, b_o).
\end{aligned}$$

6.3 Principled Model Averaging: Basketball as a Motivating Example

The model selection literature provides several methods for assigning prior probabilities to possible models from uninformed priors to priors giving differential weights to models which obey various hierarchy conditions. These methods

have a conceptual appeal. Data may be generated according to mechanisms that obey hierarchy conditions; sparsity may increase in the degree of polynomial terms. While useful guidelines, these methods do not use empirical evidence - they do not make use of the data revealing the dependence between variables. This section defines *preferential selection* and *simultaneous selection* for models over the polynomial expansion of a data set V, and introduces a novel prior that maps observed relational data to model probabilities.

[89] consider priors over models with various statistical dependencies between model terms. The graph structure relies on term order, the number of parent terms or the term length. For researchers with substantial prior knowledge about their respective problems, these approaches are insufficient. Suppose a set of networks, $\{Z\}$, is observed with nodes equal to the variables in V. We wish to encode some properties of the networks in the model priors. As an example let $V = [A, B, C]$. We define *preferential selection via Z* by

$$\Pr\{AB = 1 | A = 1, B = 1, Z[A, B] = 1\} \neq \Pr\{AB = 1 | A = 1, B = 1, Z[A, B] = 0\},$$

and *simultaneous selection via Z*

$$\Pr\{AB = 1 | \cdot, Z[A, B] = Z[A, C] = 1, AC = 1\} \neq \Pr\{AB = 1 | \cdot, Z[A, B] = Z[A, C] = 1, AC = 0\}.$$

The first assumption states that interactions between variables related via a prior network have a different propensity for selection than terms which do not. The second assumption states interaction terms which interact in the prior network with at least one of the component terms interacts with another variable such that their interaction is included in the model exhibit differ inclusion propensities than if the latter interaction were not included.

Preferential selection via Z can be implemented in [89]'s scheme. For example, suppose all interactions with an edge in the prior network are group exchangeable and share a common Beta distribution. But, the simultaneity assumption violates conditional independence of equi-ordered terms. The inclusion probabilities can no longer be written in terms of the inclusion of the parents of an interaction term. To address this issue, consider another an alternative approach. For a given model $m \in \mathcal{M}$ let

$$\Pr\left\{m | \theta, \left\{Y_g\right\}\right\} = \exp\left\{\theta^T t\left(m | \left\{Y_g\right\}\right) - A(\theta)\right\} \tag{6.3.1}$$

where Z is a set of relational networks, $t(m, Z)$ is a vector of the sufficient statistics of model m the model, θ measures the association between the sufficient statistics and the model probability and

$$A(\theta) = \log\left(\sum_{m} \exp\left\{\theta' t(m, Z)\right\}\right)$$

is the normalizing constant.

6.3.1 Passes-To Example (Second-Order Interactions with a Binary Network Prior)

Suppose a single, binary network, Z, is observed over three variables $\{A, B, C\}$. The following demonstrates how to encode preferential selection via Z, simultaneous selection via Z with this formulation and a penalty for the hierarchy constraint into the model. Write

$$\begin{aligned}
\boldsymbol{\theta}^T t(m, Z) &= \theta_1(AB + AC + BC) \\
&+ \theta_2(AB * \max(Z_{AB}, Z_{BA}) + AC * \max(Z_{AC}, Z_{CA}) + BC * \max(Z_{BC}, Z_{CB})) \\
&+ \theta_3(AB * \min(Z_{AB}, Z_{BA}) + AC * \min(Z_{AC}, Z_{CA}) + BC * \min(Z_{BC}, Z_{CB})) \\
&+ \theta_4(AB * Z_{AB} * AC * Z_{AC} + AB * Z_{BA} * BC * Z_{BC} + AC * Z_{CA} * BC * Z_{CB}) \\
&+ \alpha * \mathbb{I}_{\{m \notin \mathrm{HC}(\mathcal{M})\}}
\end{aligned}$$

where α is the hierarchy penalty. As $\alpha \to -\infty$ the probability of an offending model goes to zero. Here θ_1 captures the weight given to the total number of two-way interactions present in the model; θ_2 captures the weight given to nodes with that have a relationship in at least one direction; θ_3 captures the weight given to reciprocal relationships and θ_4 captures the weight given to two-stars - instances in which an individual exhibits a network relationship with more than one alter. To extend the binary model to arbitrary dimension write

$$\begin{aligned}
\boldsymbol{\theta}^T t(m, Z) &= \theta_1\left(\sum_{i<j} x_i x_j\right) \\
&+ \theta_2\left(\sum_{i<j} x_i x_j \max(Z_{ij}, Z_{ji})\right) \\
&+ \theta_3\left(\sum_{i<j} x_i x_j \min(Z_{ij}, Z_{ji})\right) \\
&+ \theta_4\left(\sum_{i}\sum_{j\neq i}\sum_{k\notin\{i,j\}} x_i x_j * Z_{ij} * x_i x_k * Z_{ik}\right) \\
&+ \alpha * \mathbb{I}_{\{m \notin \mathrm{HC}(\mathcal{M})\}}.
\end{aligned}$$

6.3.2 Passes-To Example (Second-Order Interactions with a Valued Network Prior)

If the edges of Z contain non-negative counts, then write

$$\begin{aligned}
\boldsymbol{\theta}^T t(m, Z) &= \theta_1(AB + AC + BC) \\
&+ \theta_2(AB * (Z_{AB} + Z_{BA}) + AC * (Z_{AC} + Z_{CA}) + BC * (Z_{BC} + Z_{CB})) \\
&+ \theta_3(AB * \min(Z_{AB} * Z_{BA}) + AC * \min(Z_{AC}, Z_{CA}) + BC * \min(Z_{BC}, Z_{CB})) \\
&+ \theta_4(AB * Z_{AB} * AC * Z_{AC} + AB * Z_{BA} * BC * Z_{BC} + AC * Z_{CA} * BC * Z_{CB}) \\
&+ \alpha * \mathbb{I}_{\{m \notin G(\mathrm{HC})\}}.
\end{aligned}$$

6.3.3 Pick and Roll Example (Third Order Interaction with Multiple Prior Networks)

The previous example only considers models of order two or less, but prior information may exist to model beliefs over the inclusion of higher order terms. Suppose two prior networks are observed: Z^{pass} and Z^{pnr}. Consider the triad formed by a shooter, a ballhandler and a screener. Suppose V contains four individuals $\{A, B, C, D\}$ and specify

$$\begin{aligned}
\boldsymbol{\theta}^T t(m, Z) &= \theta_1(AB + AC + BC) \\
&+ \theta_2(ABC + ABD + ACD + BCD) \\
&+ \theta_3(AB * (\max(Z^{\text{pass}}_{AB}, Z^{\text{pass}}_{BA}) + AC * \max(Z^{\text{pass}}_{AC}, Z^{\text{pass}}_{CA}) + AD * \max(Z^{\text{pass}}_{AD}, Z^{\text{pass}}_{DA}) + BC * \max(Z^{\text{pass}}_{BC}, Z^{\text{pass}}_{CB}) \\
&\quad + BD * \max(Z^{\text{pass}}_{BD}, Z^{\text{pass}}_{DB}) + CD * \max(Z^{\text{pass}}_{CD}, Z^{\text{pass}}_{DC})) \\
&+ \theta_4(AB * (\min(Z^{\text{pass}}_{AB}, Z^{\text{pass}}_{BA}) + AC * \min(Z^{\text{pass}}_{AC}, Z^{\text{pass}}_{CA}) + AD * \min(Z^{\text{pass}}_{AD}, Z^{\text{pass}}_{DA}) + BC * \min(Z^{\text{pass}}_{BC}, Z^{\text{pass}}_{CB}) \\
&\quad + BD * \min(Z^{\text{pass}}_{BD}, Z^{\text{pass}}_{DB}) + CD * \min(Z^{\text{pass}}_{CD}, Z^{\text{pass}}_{DC})) \\
&+ \theta_5(AB * Z^{\text{pass}}_{AB} * AC * Z^{\text{pass}}_{AC} + BA * Z^{\text{pass}}_{BA} * BC * Z^{\text{pass}}_{BC} + CA * Z^{\text{pass}}_{CA} * CB * Z^{\text{pass}}_{CB} \\
&+ \theta_6(AB * (\max(Z^{\text{pnr}}_{AB}, Z^{\text{pnr}}_{BA}) + AC * \max(Z^{\text{pnr}}_{AC}, Z^{\text{pnr}}_{CA}) + AD * \max(Z^{\text{pnr}}_{AD}, Z^{\text{pnr}}_{DA}) + BC * \max(Z^{\text{pnr}}_{BC}, Z^{\text{pnr}}_{CB}) \\
&\quad + BD * \max(Z^{\text{pnr}}_{BD}, Z^{\text{pnr}}_{DB}) + CD * \max(Z^{\text{pnr}}_{CD}, Z^{\text{pnr}}_{DC})) \\
&+ \theta_7(AB * (\min(Z^{\text{pnr}}_{AB}, Z^{\text{pnr}}_{BA}) + AC * \min(Z^{\text{pnr}}_{AC}, Z^{\text{pnr}}_{CA}) + AD * \min(Z^{\text{pnr}}_{AD}, Z^{\text{pnr}}_{DA}) + BC * \min(Z^{\text{pnr}}_{BC}, Z^{\text{pnr}}_{CB}) \\
&\quad + BD * \min(Z^{\text{pnr}}_{BD}, Z^{\text{pnr}}_{DB}) + CD * \min(Z^{\text{pnr}}_{CD}, Z^{\text{pnr}}_{DC})) \\
&+ \theta_8(AB * Z^{\text{pnr}}_{AB} * AC * Z^{\text{pnr}}_{AC} + BA * Z^{\text{pnr}}_{BA} * BC * Z^{\text{pnr}}_{BC} + CA * Z^{\text{pnr}}_{CA} * CB * Z^{\text{pnr}}_{CB} \\
&+ \alpha * \mathbb{I}_{\{m \notin G(\mathrm{HC})\}}.
\end{aligned}$$

Chapter 7

A Latent Network Model for Competitive Interaction

7.1 Introduction

This paper develops methods to learn the value of main and interaction effects when an outcome is observed at a group level. Outcomes are associated with a subset of interacting nodes. Many examples can be subsumed into this paradigm. Interaction between members of a commercial or political organization gives rise to measures of collective efficacy, musicians collaborate to create pieces of music, and teammates coordinate to score points or prevent competitors from scoring. The collective outcome can be attributed to the individuals comprising the group and the ways in which these individuals interact.

Although this methodology applies to many social systems, this work focuses on basketball. This is a particularly good application because relatively few individuals interact to produce group level outcomes repeatedly. Moreover, some of the intra-possession interactions can be observed. Hence we can link relational events to the probability distribution of outcomes. In this application estimation of a latent network model can be understood as a further innovation in the family of the regularized main effect models commonly used in basketball analytics. These models capture a node's impact given the presence of all other nodes on the court. But, they fail to account for the inherently social nature of the sport.

Without doubt, the inclusion of interaction terms captures the data-generating process in a more faithful way. But, even for groups of moderate size, the space of possible models is huge. To avoid overfitting (by estimating far too

many terms), one can perform model selection via a penalized regression method such as the Lasso developed in [91]. While such procedures produce a parsimonious model, they do not provide information as to why certain subsets of nodes are selected and others are not. By placing a structured network prior on the set of feasible interactions terms, the possible models (dyadic and/or triadic interactions), one can estimate the influence of various features on the model probabilities. Typical examples are dyadic relational events such as passes and fouls. Both are examples of directed relational events, but in the former case one teammate throws the ball to the other while in the latter a defensive node fouls an offensive node (or vice versa). Parameters corresponding to sufficient statistics derived from the relational events indicate why certain subsets of players are selected by the model. The network approach provides a method to capture and interpret the extent of a team's connectedness. Bayesian inference learns about the model probabilities from observed relational events, and average over the possible models to make better predictions.

The chapter is organized as follows. Section 7.2 introduces the scientific framework, important concepts and notation. Two variations of the general method are described: *CoordiNet* and *TriadNet*. The primary distinction between the models is the degree of interaction considered in the latent networks. The former considers dyadic interaction while the latter considers dyadic and triadic interaction. As the degree of interaction grows large, hierarchy of interactions are needed to preserve the interpretation of model parameters. Section 7.3 examines possible the estimation strategy for *CoordiNet*. Then alterations needed for *TriadNet* are presented. Section 7.5 applies *CoordiNet* to the Boston-Washington series during the 2016-17 NBA playoffs. Section 7.6 compares the network prior method to competing procedures to select interactions. Model variations and extensions are examined.

7.2 Team Sports as a Relational System

This section presents the conceptual framework and justification for the latent network approach to group interaction. The introduction argues that basketball is not an atomistic sport; people must coordinate their actions to achieve a collective outcome. The ball cannot be passed without a sender and receiver, and the so-called hockey assist (a pass from i to j followed by an assist from j to k) is not possible without three individuals. The triadic interaction of a pick and roll pair with a shooter in the corner requires all three to be present for certain actions to be efficacious. When these interactions, coordinated or not, have a differential effect on the propensity to score, game play is relational. That is, models accounting for interaction provide a more realistic framework than atomistic models of play.

As noted in the introduction, the models developed can be applied to many team sports such as soccer, hockey and lacrosse, but from this point forward focus resides on the model's application to basketball for several reasons. The possession in basketball serves as a natural unit of analysis as it is clearly defined and many possessions occur during

a typical game. Since rosters are small, many combinations of players are observed; to estimate interaction terms, observations with the two nodes separately and together are required. If lineups are changed completely, many players never face or play with one another, limiting the range of interactions.

For each possession many outcomes are possible such as the number of points scored, the amount of time elapsed, the amount of movement during the possession and counts of particular events. The latent network estimated by the models developed in this paper are defined relative to the outcome of interest. Hence, when the number of points scored during a possession serves as the outcome, the network measures the propensity to score as a function of the individuals on the court. Alternatively, if total ball movement per second is the outcome, the latent network measures feet moved per second as a function of the players.

Many events transpire during a possession that lead to the observed outcome of interest. Observed relational events can be understood as networks in-themselves. For example, when every pass is recorded during possessions, one can model the data as a network directly or as a sequence of relational events ([14]). Doing so allows one to model the frequency of interaction, and therefore, make predictions about the flow of intra-possession phenomena. Defining a network relative to an outcome, while using the observed relational data, enables one to make predictions about efficacy of the set of players. Outcomes can be modeled in terms of relational events directly. That is, the number of points scored can be modeled in terms of the number and order of relational events observed during a possession.

A network need not be modeled for play to be relational. Estimating the probability that player i will pass to player j given features describing the possession is a relational account of the play. To continue the example, a passing matrix between nodes on the same team can be estimated as part of this approach. After a pass from i a probability mass function can be estimated for model j's action. But, as models become more complicated, the demands on the data become more extreme. Moreover, the focus shifts from estimating interpretable parameters representing structure to recreating game play.

Highly-detailed intra-possession models require that relevant relational information is present in the data. Given the subtlety and complexity of most possessions, many relational events are missing from the data. For example, play-by-play data sources record a lob thrown to a player for a dunk, but neglect the back screen used to free the dunker from the defense. Modeling outcomes in terms of observed relational events makes interpretation difficult because many crucial relational events are omitted from the model. Omission of such events bias parameter estimates of the relational events that are recorded in the data. Thus, an analyst may identify important relational events, but if they are not recorded in the data, they cannot be used to model intra-possession dynamics. Aside from the lack of important

events, some relational events may not be recognized by human cognition. That is, teammates may coordinate their actions in ways that are difficult to classify or capture with a label. These type of relational events include subconscious positioning relative to others, non-verbal communication such as eye contact or boxing out an opponent so a teammate can rebound a missed shot. So not only will important, known relational events be missing, other important, unknown relational events are missing from the data. True chemistry on the court cannot be fully explained by charting the motions of players. The lobs, screens, spacing and other relational events can instead be thought of as manifestations of team chemistry represented as an underlying or latent network. The latent network approach generalizes the relationships between the nodes in a population.

Two general models for group-level outcomes generated by interacting nodes are depicted in 7.1. In both the right and left panels, the group-level response R is generated by the main effects β, interaction effects Ψ and additional parameters α. The two panels show two types of model selection methods. In the left panel, M is random variable indicating which interactions are to be included in the response model. By choosing to model M as an exponential random graph, it is recognized as a network between nodes. Additionally, this choice allows M to depend on relational data Y between members of the population. Note the prior placed on the network does not directly model the value of nonzero dyads in the interaction adjacency matrix Ψ. Hence the values of θ learned during estimation relate to the probability of inclusion as a function of relational events. The prior provides no information about how relational events are associated with the values.

The model in the right panel does not include M. Rather, the dyads of Ψ are modeled via an additive-multiplicative model. The parameters θ_Ψ do reflect the association between relational events and the value of Ψ directly. By adjusting

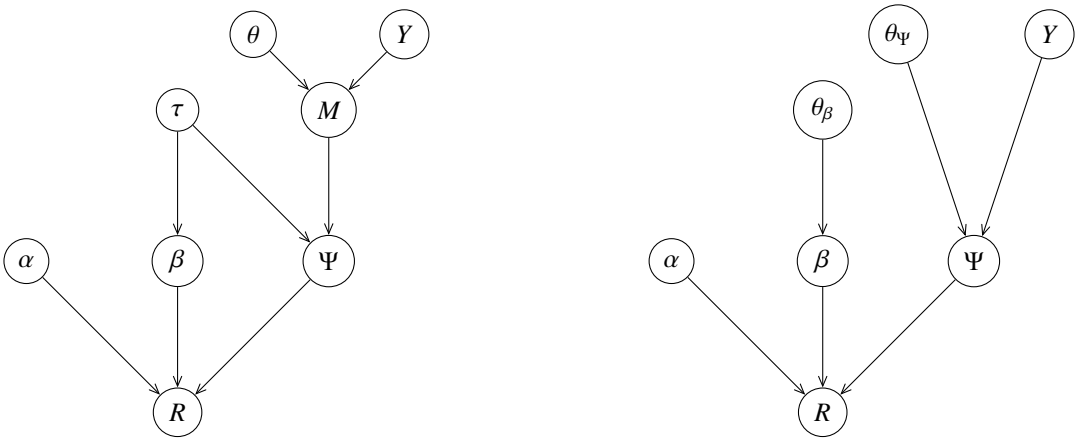

Figure 7.1: Group Outcome Model as a Directed Acyclic Graph

the assumptions associated with the various components in the model, many variations are possible. As discussed in

this section, a network prior is placed over the interaction terms. Although this technique is novel in the study of social networks, network priors have been used in the analysis of brain networks, particularly Bayesian connectomics. [52] and [53] assume connections between regions in the brain are distributed as a network. In the former paper the graph connecting brain regions is distributed as an exponential random graph such that the edge length follows an exponential distribution. This choice prefers networks exhibiting the so-called small-world property. The latter assumes the prior network follows an Erdős-Rényi model. The remainder of this section explores two common variation dubbed *CoordiNet* and *TriadNet*, respectively.

7.2.1 Dyadic Interaction Models: *CoordiNet*

The most basic version of group-level outcome modeling with a network prior, called *CoordiNet*, assumes the possession outcomes are a function of main effects and dyadic interaction effects. Two players exhibit a relation if they share an edge in a given model. The edges are real-valued; pairings associated with a higher propensity for offensive have positive edges, while dyads associated with lower scoring propensities have negative edges. Hence a negative edge between two offensive players indicates a worse outcome for the offensive team. And, a positive edge between an offensive and defensive player indicates a better outcome for the offensive team, and a worse outcome for the defensive team.

An edge is said to be feasible if the dyad appeared on the court at any time during the sample. Since the number of possible models becomes large as the set of the feasible dyads increases, the dyadic model must be able to select good candidates. While we do want to include every feasible edge to clarify interpretation of the model parameters, it would be absurd to consider this model only. The inclusion of too many terms can lead to overfitting, and in some extreme cases, aliasing. Moreover, practical experience dictates that not every dyad has an impact on the offense's propensity to score. Model selection is made tractable by placing an exponential random graph prior on the model space. Observed relational events between dyads guide the model exploration process. Let V be a population of nodes. The model analyzes outcomes generated by subsets of V. Define the following model objects referenced in the two

panels in 7.1 as follows.

$$
\begin{aligned}
R &= \text{a vector of possession-level responses,} \\
\beta &= \text{individual player effects in the underlying model,} \\
\Psi &= \text{interaction effects for the underlying model,} \\
\alpha &= \text{additional parameters for the underlying model,} \\
\left\{Y_g\right\} &= \text{layers of data relating the nodes in } V, \\
\theta &= \text{parameter measuring association between} \\
&\quad \left\{Y_g\right\} \text{ and } M \text{ in the left panel,} \\
\theta_\Psi &= \text{parameter measuring association between} \\
&\quad \left\{Y_g\right\} \text{ and } \Psi \text{ in the right panel,} \\
\tau &= \text{the prior standard deviation for } \Psi \text{ and } \beta \text{ in the left panel,} \\
\gamma &= \text{hyperparameter for prior over } \alpha, \\
\theta_\beta &= \text{parameters for main effects in the right panel.}
\end{aligned}
$$

The relational system contains three pieces of observed data, the possession outcomes R, the lineups present on the court during each possession W and the observed relational events between players $\left\{Y_g\right\}_{g=1}^{G}$. Points are scored as the result of the myriad events which transpire during a possession. Some of these things can be captured in the layers of relational data, $\left\{Y_g\right\}$, such as passes between nodes, assists between nodes, fouls against between two nodes, shots blocked, pick and rolls run, double teams executed and so on. Observable directed actions are manifestations of underlying connections between nodes. Of course, as explored in the discussion, outcomes can be modeled as a function of these actions directly. But, the current interest focus is in the general connections between nodes. Consider that latent ties may exist between players who have not coordinated via one of the observable directed actions $\left\{Y_g\right\}$ available to us. This method includes these instances in the same way that (regularized) adjusted plus-minus models account for individual actions not recorded by boxscores or optimal cameras. Response data is assumed to be conditionally independent of $\left\{Y_g\right\}$ given Ψ so that the outcome is a function of lineup indicators and the parameters α, β and Ψ.

The main effect parameters β represent the vector of main (or individual) effects on the propensity to score. The *CoordiNet* model includes a main effect term for every player in the dataset. The interaction terms in Ψ are selected according to the prior probability distribution over the network. Since every main effect is included, every model explored in the ensuing MCMC sampling satisfies strong hierarchy; i.e., if an interaction between nodes is included

in the model, then both of the main effect terms associated with the dyad are included. The exact form α assumes depends upon the probability distributions chosen to model the response. Since *CoordiNet* takes the response R to be the number of points scored during a possession, an ordinal logistic regression framework is assumed. Then α is the set of cut-point parameters. Precise details follow in Section 7.3.

Since *CoordiNet* assumes only dyadic interaction, Ψ can be represented as an adjacency matrix as in 7.2. Two nodes

$$
\begin{array}{c}
\begin{array}{cccccccc}
o_1 & \dots & o_{|V|} & & x_1 & \dots & x_{|V|} & \\
\end{array} \\
\left[\begin{array}{ccc|ccc|c}
 & & & & & & o_1 \\
 & \Psi^{O-O} & & & \Psi^{O-X} & & \vdots \\
 & & & & & & o_{|V|} \\
\hline
- & - & - & - & - & - & \\
\hline
 & & & & & & x_1 \\
 & \Psi^{X-O} & & & \Psi^{X-X} & & \vdots \\
 & & & & & & x_{|V|}
\end{array}\right]
\end{array}
$$

Figure 7.2: The Adjacency Matrix in *CoordiNet*.

i and j can be tied to one another in four different ways. If i and j play on the same team and appear on the court together while their team is on offense, the offensive edge Ψ_{ij}^{O-O} is feasible. Similarly, the value of defensive coordination between nodes i and j is represented by Ψ_{ij}^{X-X}. If i and j belong to different teams, when i is on defense and j is on offense, their interaction is captured by Ψ_{ij}^{X-O}. Finally, when i's team is is defended by j's, their interaction is captured by Ψ_{ij}^{O-X}. Note that Ψ^{O-O} and Ψ^{X-X} are symmetric matrices. Only dyads connecting pairs of players who have appeared on the court together for at least one possession are feasible and all self-loops are set to zero. The networks Ψ^{X-O} and Ψ^{O-X} are transposes of one another and connect dyads nodes from different teams. Competition between two teams can be represented as interacting subnetworks as in 7.3. The models presented in 7.1 can be decomposed into

$$
\begin{aligned}
p\left(R, \Psi, \beta, \alpha \,\middle|\, W, \left\{Y_g\right\}\right) &= p\left(R \,|\, \Psi, \beta, \alpha, W\right) p\left(\Psi, \beta, \alpha \,\middle|\, \gamma, \tau, \left\{Y_g\right\}\right) \\
&= p\left(R \,|\, \Psi, \beta, \alpha, W\right) p(\alpha|\gamma) p(\beta|\tau) \sum_{m \in \mathcal{M}} p\left(\Psi \,|\, m, \tau\right) p(m|\theta, \left\{Y_g\right\})
\end{aligned}
$$

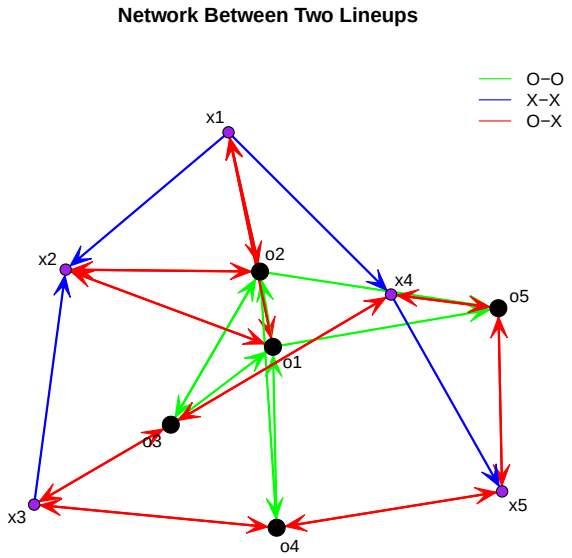

Figure 7.3: Competitive Interaction in *CoordiNet*.

and

$$
\begin{aligned}
p\left(R,\Psi,\beta,\alpha \,\middle|\, W,\left\{Y_g\right\}\right) &= p\left(R \,|\, \Psi,\beta,\alpha,W\right) p\left(\Psi,\beta,\alpha \,\middle|\, \gamma,\theta_\beta,\theta_\Psi,\left\{Y_g\right\}\right) \\
&= p\left(R \,|\, \Psi,\beta,\alpha,W\right) p(\alpha|\gamma) p(\beta|\theta_\beta) p\left(\Psi \,\middle|\, \theta_\Psi,\left\{Y_g\right\}\right),
\end{aligned}
$$

respectively. Here $p(R|\cdot)$ is the probability distribution for the response, and $p(\Psi|m,\tau)$, $p(\beta|\tau)$ and $p(\alpha|\gamma)$ are the prior densities over Ψ, β and α, respectively. These probabilities are specified in Section 7.3.

7.2.2 Triadic Interaction Models: *TriadNet*

Beyond the dyad, the next most fundamental social structure is the triad. *CoordiNet* implicitly assumes that triadic effects are the sum of the main and dyadic effects. But, basketball commonsense suggests this is not the case. Consider three offensive players performing specific roles: a ballhandler, a screener and a spot-up shooter. The ballhandler and screener perform a pick and roll while the shooter waits for any opportunity to shoot or cut to the basket. If all three perform their rolls extremely well, the three-term offensive interaction is positive. Hence the full value of the trio exceeds the sum of the main and dyadic effects. Similarly, particularly young or inexperienced trios may perform

worse than the sum of their main and dyadic effects. Geometrically, the introduction of a third node changes the spatial orientation from a line to a triangle thereby increasing the dimension of the space the defense must traverse to move between the offensive players. More complicated strategies become feasible with three nodes.

The triadic network can be represented as 7.4. Each sublattice represents a different type of triadic interaction. For example the green sublattice (top, left cube closest to the the reader) is the triadic offense cube. Each point represents an offensive interaction between teammates. Since hyperedges are undirected, the symmetry exhibited in the adjacency matrix for *CoordiNet* carries over to the adjacency cube for *TriadNet*. Every two dimensional subspace derived by fixing the value of one coordinate is a symmetric matrix. The red cube represents the defensive sublattice; it exhibits similar symmetry properties. The blue cube represents the defense-offense-defense interaction lattice. Because the edges are undirected every sublattice containing the same number of offense and defense terms contains the same information. In the adjacency lattice the yellow cube contains the same information as the blue cube. Four types of

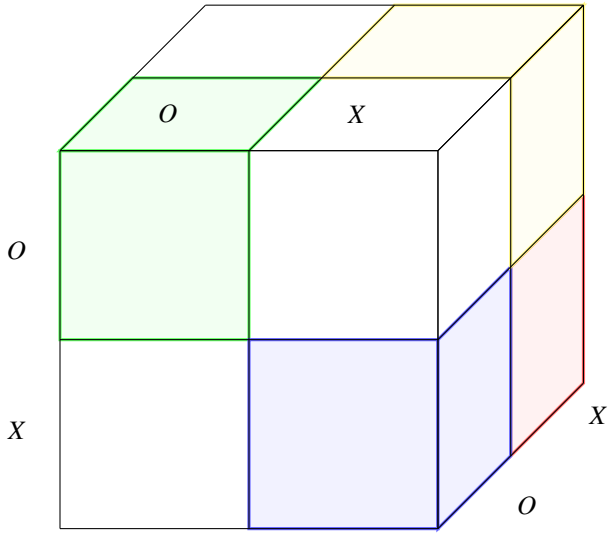

Figure 7.4: The Adjacency Cube in *TriadNet*.

triadic interactions can be included in the model:

- Φ^{O-O-O} - triads of offensive players;

- Φ^{X-X-X} - triads of defensive players;

- Φ^{X-X-O} - triads consisting of two defenders and one offender; and

- Φ^{O-O-X} - triads consisting of two offenders and one defender.

Strong and Weak Hierarchy

Because dyadic effects are chosen probabilistically, interpretability demands some notion of interaction hierarchy. A triadic model, represented by the adjacency matrix and lattice (Ψ, Φ), satisfies strong hierarchy if

1. $\Phi_{ijk}^{O-O-O} \neq 0 \implies \Psi_{ij}^{O-O} \neq 0 \wedge \Psi_{jk}^{O-O} \neq 0 \wedge \Psi_{ik}^{O-O} \neq 0$,
2. $\Phi_{ijk}^{X-X-X} \neq 0 \implies \Psi_{ij}^{X-X} \neq 0 \wedge \Psi_{jk}^{X-X} \neq 0 \wedge \Psi_{ik}^{X-X} \neq 0$,
3. $\Phi_{ijk}^{X-X-O} \neq 0 \implies \Psi_{ij}^{X-X} \neq 0 \wedge \Psi_{jk}^{X-O} \neq 0 \wedge \Psi_{ik}^{X-O} \neq 0$, and
4. $\Phi_{ijk}^{O-O-X} \neq 0 \implies \Psi_{ij}^{O-O} \neq 0 \wedge \Psi_{jk}^{O-X} \neq 0 \wedge \Psi_{ik}^{O-X} \neq 0$.

These conditions constrain the space of feasible models. In order to estimate a triadic effect, three dyadic effects must also be included in the model. If a certain scheme requires three players to implement it effectively, strong hierarchy may miss important coordination between nodes and fail to learn interesting structures. As an alternative consider a version of weak hierarchy in which at least one of the interactions must be present. Formally, the adjacency matrix and lattice (Ψ, Φ), satisfies weak hierarchy, denoted $\mathcal{WH}(\Psi, \Phi) = 1$, if

1. $\Phi_{ijk}^{O-O-O} \neq 0 \implies \Psi_{ij}^{O-O} \neq 0 \vee \Psi_{jk}^{O-O} \neq 0 \vee \Psi_{ik}^{O-O} \neq 0$,
2. $\Phi_{ijk}^{X-X-X} \neq 0 \implies \Psi_{ij}^{X-X} \neq 0 \vee \Psi_{jk}^{X-X} \neq 0 \vee \Psi_{ik}^{X-X} \neq 0$,
3. $\Phi_{ijk}^{X-X-O} \neq 0 \implies \Psi_{ij}^{X-X} \neq 0 \vee \Psi_{jk}^{X-O} \neq 0 \vee \Psi_{ik}^{X-O} \neq 0$, and
4. $\Phi_{ijk}^{O-O-X} \neq 0 \implies \Psi_{ij}^{O-O} \neq 0 \vee \Psi_{jk}^{O-X} \neq 0 \vee \Psi_{ik}^{O-X} \neq 0$.

There are at least two different strategies to induce hierarchy in models. One can sample from triadic networks that satisfy weak hierarchy only. Then every model in the probability distribution over models represented by (Ψ, Φ) satisfies weak hierarchy. This ensures a minimal level of interpretability. However, constraining the model space reduces predictive ability. One cannot minimize mean squared error by restricting the space of possible optima. An alternative approach is to introduce a hierarchy condition as a sufficient statistic in the model probability. Possibilities include the number of violations of weak and/or strong hierarchy in a model or simply whether a model satisfies weak and/or strong hierarchy. Using the number of violations as a sufficient statistics dictates the relative importance of hierarchy in terms of explaining the variation in the possession outcomes.

In addition to hierarchy terms, triadic sufficient statistics can be used in the prior over Φ given Ψ. If triadic relational events are available, they can be used as sufficient statistics. Examples of offensive relational events include

hockey assists in which i passes the ball to j who assists k and the aforementioned ballhandler, screener and shooter triangle. A common $X-X-O$ relational event is the so-called double team. Nodes i and j defend k collectively.

The model grows to admit Φ and m_Φ. As in the dyadic model, the latter object is binary whereas the former is real-valued. The primary change to the model is reflected in the alternative decomposition where the relational data Y and the lineup indicator data W have been suppressed for intelligibility. The probability distribution of the relational system can be written as

$$\begin{aligned} p\,(R,\Phi,\Psi,\beta,\alpha\,) &= p\,(R\,|\,\Phi,\Psi,\beta,\alpha\,)\,p\,(\Phi,\Psi,\beta\,\alpha|\,\gamma,\tau,\,) \\ &= p\,(R\,|\,\Phi,\Psi,\beta,\alpha\,)\,p(\alpha|\gamma)p(\beta|\tau)\sum_{m_\Psi\in\mathcal{M}_\Phi}\sum_{m_\Psi\in\mathcal{M}_\Psi} p\,(\Phi,\Psi\,|\,m_\Phi,m_\Psi,\tau\,)\,p(m_\Phi,m_\Psi|\zeta,\theta) \\ &= p\,(R\,|\,\Phi,\Psi,\beta,\alpha\,)\,p(\alpha|\gamma)p(\beta|\tau)\sum_{m_\Psi\in\mathcal{M}_\Phi}\sum_{m_\Psi\in\mathcal{M}_\Psi} p\,(\Phi\,|\,m_\Phi,m_\Psi,\tau\,)\,p(m_\Phi|m_\Psi,\zeta)p\,(\Psi\,|\,m_\Psi,\tau\,)\,p(m_\Psi|\theta). \end{aligned}$$

7.3 Modeling a Latent Network

This section provides the probability distributions assumed for each random variable in the models depicted in 7.1. The mathematical structure for the response model and cut-point parameters is invariant across the two panels. Recall the left panel performs model selection explicitly by modeling the probability that a subset of dyads is included in the model as an exponential random graph while the right places a model selection prior directly on Ψ. Since the systems contain so many random variables the hyperparameter for the priors over Ψ and β are fixed thereby constraining the strength of the prior. Possible probability models for these parameters are introduced and discussed in 7.6.

Model Averaging versus Model Selection Model selection is helpful whenever the number of features, in this case main effects and interactions of maximum degree two or three, is large and a parsimonious explanation is desired. However, the decision to choose models probabilistically requires a method to interpret the results. One method is sum over the possible values for M - this is model averaging. Then for Ψ,

$$\begin{aligned} \Pr\{\Psi=\psi|R\} &= \sum_{m\in\mathcal{M}} p(\Psi,m|R) \\ &= \sum_{m\in\mathcal{M}} \frac{p(R|m,\Psi)p(\Psi,m)}{p(R)} \\ &\propto \sum_{m\in\mathcal{M}} p(R|m,\Psi)p(\Psi|m)p(m;\theta) \\ &= p(R|\Psi)p(\Psi|m^*)p(m^*;\theta) \end{aligned}$$

where m^* is the model that concords with Ψ given by $\mathbb{I}\{\Psi \neq 0\}$, and

$$p(R|\Psi) = \int\int p(R, \alpha, \beta|\Psi)\, d\alpha d\beta.$$

This method requires that the conditional probability of $p(\Psi|m)$ is estimated. In practice this means running MCMC chains for the model parameters for every model. Note that for a given m, choices of Ψ can be infeasible if a dyad in Ψ is nonzero and the corresponding dyad in M is zero. Hence the probability is then determined by exactly one model.

Alternatively, one can choose some of the highest performing models and compare them. Doing this requires computation of the model probabilities. Generally, the posterior probability of a model is

$$p(m|R) = \frac{p(R|m)p(m)}{\sum_{m' \in \mathcal{M}} p(R|m')p(m')}$$

where

$$\begin{aligned} p(R|m) &= \int\int\int p(R|\Psi, \alpha, \beta, m)p(\alpha, \beta, \Psi|m)\, d\alpha d\beta d\Psi \\ &= \int\int\int p(R|\Psi, \alpha, \beta, m)p(\alpha)p(\beta)p(\Psi|m)\, d\alpha d\beta d\Psi. \end{aligned} \tag{7.3.1}$$

Unless a small group of models are far more probable than the many alternatives, model averaging provides probabilities that each term is included in the model. The MCMC routine reflects this decision. Parameters are sampled serially and individually. Hence model selection can be implemented by introducing a point mass on that probability that a term is zero. Once the sampling routine visits a new model, it must sample the from the parameter space conditioned on the model. The goal of the sampling is to produce samples of the form

$$\left\{\left\{\alpha^{(n)}, \beta^{(n)}, \Psi^{(n)}|R, m\right\}_{n=1}^{N}\right\}_{m \in \mathcal{M}}$$

from the joint posterior distribution with N samples taken for every possible model. In practice only a subset of models will be visited. Before constructing the routine, the next few subsections introduce model assumptions for the various random variables.

7.3.1 The Response Model

The network prior method admits any probabilistic response model at the group level. To make the latent network reflect a node's or dyad's ability to affect the score, the response is chosen to be the number of points scored during

a possession. If parameters are needed for each node and dyad, then the proportional odds logistic regression is a natural choice ([16]). A single set of parameters provide each unit's probability distribution over outcomes. This has its benefits and drawbacks, a higher rate lineup will always have a higher probability to obtain the largest outcome. In Section 7.5 the higher number of points in a possession is four (though possessions of five or more points are possible). Teams without four point plays will be predicted to achieve them with higher probability. Section 7.6 discusses modifications to address this issue. In addition to the vector of responses, the lineups used during play are observed. Denote this matrix as W. Where each column indicates whether a player is on the court during a possession. Let W^O be the columns corresponding to the offensive players and W^X be the columns corresponding to the defensive players. Define the interaction matrices W^{O-O}, W^{X-X} and W^{O-X}. Each contains terms indicating which offense-offense, defense-defense and offense-defense pairs share the court during a possession, respectively. Following [37], for each observation define the latent variable

$$R_i^* = \left\langle w_i^{O-X}, \Psi^{O-X} \right\rangle + \left\langle w_i^{O-O}, \Psi^{O-O} \right\rangle + \left\langle w_i^{X-X}, \Psi^{X-X} \right\rangle + \left\langle w_i^{O}, \beta^{O} \right\rangle + \left\langle w_i^{X}, \beta^{X} \right\rangle + \epsilon_i \tag{7.3.2}$$

$$= s_i + \epsilon_i \tag{7.3.3}$$

where $\langle w_i, \Psi \rangle$ is the inner product of the i^{th} row of the interaction matrix and the vectorization of Ψ, and the ϵ_i are independent and have logistic distribution given by

$$p(\epsilon_i \leq t) = \text{logit}^{-1}(t) = \frac{e^t}{1 + e^t}.$$

This assumption implies

$$p(R_i^* \leq t) = p(s_i + \epsilon_i \leq t) = p(\epsilon_i \leq t - s_i).$$

The quantity s_i is the latent value of the lineup w_i. With outcomes in $\{0, 1, 2, 3, 4\}$ the model requires the estimation of a vector of cut-points α satisfying

$$R_i = \begin{cases} 0 & \text{if } -\infty < R_i^* < \alpha_1 \\ 1 & \text{if } \alpha_1 \leq R_i^* < \alpha_2 \\ 2 & \text{if } \alpha_2 \leq R_i^* < \alpha_3 \\ 3 & \text{if } \alpha_3 \leq R_i^* < \alpha_4 \\ 4 & \text{if } \alpha_4 \leq R_i^* < \infty \end{cases}$$

Note that we only consider possessions in which 0, 1, 2, 3 or 4 points are scored - this accounts for more than 99% of the possession from the 2016-17 season. The probability that the i^{th} observation results in outcome j is

$$\begin{aligned} p(R_i = j) &= p(R_i \leq j) - p(R_i \leq j-1) \\ &= p(R_i^* \leq \alpha_j) - p(R_i^* \leq \alpha_{j-1}) \\ &= \frac{e^{\alpha_j - s_i}}{1 + e^{\alpha_j - s_i}} - \frac{e^{\alpha_{j-1} - s_i}}{1 + e^{\alpha_{j-1} - s_i}} \end{aligned}$$

The likelihood of observing the vector of responses R is therefore given by

$$p(R|\Psi, \alpha) \propto \prod_{i=1}^{N} p(R_i|\Psi) \tag{7.3.4}$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{J} p(R_i \leq j) \tag{7.3.5}$$

$$= \prod_{i=1}^{N} \prod_{j=1}^{J} \left(\text{logit}^{-1}(\alpha_j - s_i) - \text{logit}^{-1}(\alpha_{j-1} - s_i)\right)^{z_{ij}} \tag{7.3.6}$$

$$= \prod_{l=1}^{L} \prod_{j=1}^{J} \left(\text{logit}^{-1}(\alpha_j - s_i) - \text{logit}^{-1}(\alpha_{j-1} - s_i)\right)^{\sum_{i:w_i=l} z_{ij}} \tag{7.3.7}$$

where $z_{ij} = \sum \mathbb{I}\{R_i = j\}$ and l indexes the L distinct lineups in the N observations.

7.3.2 Prior for the Cut-points

Prior Over the Proportion of each Level

Although they are not typically the object of interest in analysis, the cut-points α require a prior distribution. By assumption, the prior for α is independent of all other random variables in 7.1. The authors of [35] specify a Dirichlet prior over the probability of each outcome with the predictors evaluated a their sample means. In the context of the current application, the mean of the data refers to a possession played by all players with weights determined by the number of possessions played. Since all input data W is binary, the sample mean is never realized in the data. But, it corresponds to the league averages for all outcomes. For $j \in \{0, 1, 2, 3, 4\}$ let

$$\pi_j = \Pr\{R = j|\bar{x}\} \tag{7.3.8}$$

and

$$p(\pi|\gamma) \propto \prod_{j=1}^{J} \pi_j^{\gamma_j - 1} \tag{7.3.9}$$

where γ is a vector of prior counts. Setting $\gamma_j = 1$ assumes 1 observation in each of the categories. To compute the cut-point j, sum the is given by

$$\alpha_j = \text{logistic}^{-1}\left(\sum_{i=1}^{j} \pi_i\right) = \ln\left(\frac{\sum_{i=1}^{j} \pi_i}{1 - \sum_{i=1}^{j} \pi_i}\right). \tag{7.3.10}$$

Hence cut-points are not estimated directly. To derive the implied prior distribution observe that

$$\sum_{i=1}^{j} \pi_i = \frac{\exp\left\{\alpha_j\right\}}{\exp\left\{\alpha_j\right\} + 1}$$

for $j = 1, \ldots, J$. The prior for α_j is a function of $\sum_{i=1}^{j} \pi_i$ hence the prior distribution over the $\sum_{i=1}^{j} \pi_i$ is need to compute $p(\alpha_j)$. If the first j terms sum to $\tilde{\pi}$, then the vector of proportions is in the set

$$V = \left\{\pi \in \Delta_{K-1} : \sum_{i=1}^{j} \pi_i = \tilde{\pi}, \sum_{i=j+1}^{J} \pi_i = 1 - \tilde{\pi}\right\}$$

and

$$p\left(\sum_{i=1}^{j} \pi_i = \tilde{\pi}\right) = \frac{\text{vol}(V)}{1/J!}.$$

Note that V is a subset of the J dimensional simplex. The vertices $k = 0, \ldots, J$ are given by

$$v_{kl} = \begin{cases} \tilde{\pi} & \text{if } k \leq j, l = k \\ 1 - \tilde{\pi} & \text{if } k > j, l = k \\ 0 & \text{else} \end{cases}$$

Then a change of variable yields

$$\begin{aligned} p(\alpha_j = a) &= \det \frac{d}{d\alpha_j} g^{-1}(\alpha_j) p(g^{-1}(\alpha_j)) \\ &= \frac{\exp\left\{\alpha_j\right\}}{\left(\exp\left\{\alpha_j\right\} + 1\right)^2} \frac{\text{vol}(V)}{1/J!}. \end{aligned}$$

If they are available, aggregate counts or proportions from prior seasons can be used.

Direct Prior Over α

While the prior distribution over the probability of each outcome is quite intuitive, there are alternative formulations that may perform better. Recall the cutpoints must satisfy

$$\alpha_0 < \cdots < \alpha_4. \tag{7.3.11}$$

so that any prior distribution over α must satisfy

$$\{a : p(\alpha_i = a) > 0\} \cap \{a : p(\alpha_{i+1} = a) > 0\} = 0$$

for $i = 0, 1, 2, 3$. Let $\bar{\alpha}$ be a vector of means satisfying

$$\bar{\alpha}_0 < \cdots < \bar{\alpha}_4.$$

Then a joint prior for α can be defined by

$$\begin{aligned} p(\alpha = a) \propto & N(a_0; \bar{\alpha}_0, \sigma_\alpha^2)\mathbb{I}\{a_0 < a_1\} * \\ & N(a_1; \bar{\alpha}_1, \sigma_\alpha^2)\mathbb{I}\{a_0 < a_1 < a_2\} * \\ & N(a_2; \bar{\alpha}_2, \sigma_\alpha^2)\mathbb{I}\{a_1 < a_2 < a_3\} * \\ & N(a_3; \bar{\alpha}_3, \sigma_\alpha^2)\mathbb{I}\{a_2 < \alpha_3 < a_4\}. \end{aligned}$$

Note the prior gives zero weight to any a that does not satisfy 7.3.11. The vector of means can be the observed frequency of each outcome by transformed as in 7.3.10. Under this construction, the likelihood and proposal can be written in terms of α. Details on the proposal distribution are in Section 7.4.

7.3.3 The Network Prior

This section specifies the prior probability distributions used to model dyadic interaction. Two types of models are presented: the exponential random graph in the left panel and the additive-multiplicative model in the right panel.

Exponential Random Graph Prior

The random variable M indicates which dyads in Ψ are nonzero. That is, for all $i, j \in V$ with $j \neq i$, $M_{ij} = 1$ if and only if $\Psi_{ij} \neq 0$. *CoordiNet* assumes Ψ can be partially explained in terms of relational data between nodes. Two conceptualizations are possible. In the first case a prior for model selection is introduced while in the second the dyads

are directly modeled in terms of relational data.

$$p\left(M = m \,\middle|\, \left\{Y_g\right\}, \theta\right) = \frac{\exp\left\{\theta^\top g\left(m \,\middle|\, \left\{Y_g\right\}\right)\right\}}{Z(\theta)} \tag{7.3.12}$$

where $g\left(m| X, \left\{Y_g\right\}\right)$ is a vector of sufficient statistics and

$$Z(\theta) = \sum_{m \in \mathcal{M}} \exp\left\{\theta^\top g\left(m \,\middle|\, \left\{Y_g\right\}\right)\right\}$$

is a normalizing constant. The probability that M takes a given value can depend on a variety of sufficient statistics, many derived from the layers of observed relational data. Some possibilities include

- **a term for sparsity:** the total number of nonzero terms in M;
- **a term for the number of possessions played together:** for edges of types $t \in \{O-O, X-X\}$,

  $$\frac{\sum_{i \in V} \sum_{j \neq i} \text{poss}_{ij}^t \Psi_{ij}^t}{2},$$

  and

  $$\sum_{i \in V} \sum_{j \neq i} \text{poss}_{ij}^{O-X} \Psi_{ij}^{O-X}$$

  for offense-defense interactions;
- **terms for transitivity:** for $t \in \{O-O, X-X\}$ the number of triangles in M are computed by

  $$\frac{\sum_{i \in V} \sum_{j \neq i} \sum_{k \neq i, k \neq j} \mathbb{I}\left\{\Psi_{ij}^t \neq 0\right\} \mathbb{I}\left\{\Psi_{jk}^t \neq 0\right\} \mathbb{I}\left\{\Psi_{ki}^t \neq 0\right\}}{3};$$

- **terms interacting different relationships:** the number of dyads with a term for both offensive and defensive interactions

  $$\frac{\sum_{i \in V} \sum_{j \neq i} \mathbb{I}\left\{\Psi_{ij}^{O-O} \neq 0\right\} \mathbb{I}\left\{\Psi_{ij}^{X-X} \neq 0\right\}}{2};$$

- **terms involving relational events:** counts such as

  $$\sum_i \sum_{j \neq i} \Psi_{ij}^{O-O} Y_{ij}^{\text{ast}}, \ \sum_i \sum_{j \neq i} \Psi_{ij}^{O-O} Y_{ij}^{\text{oreb}}, \ \sum_i \sum_{j \neq i} \Psi_{ij}^{X-X} Y_{ij}^{\text{blkreb}}$$

  for teammates, and

  $$\sum_i \sum_{j \neq i} \Psi_{ij}^{X-O} Y_{ij}^{\text{stl}}, \ \sum_i \sum_{j \neq i} \Psi_{ij}^{O-X} Y_{ij}^{\text{foul}}, \ \sum_i \sum_{j \neq i} \Psi_{ij}^{X-O} Y_{ij}^{\text{blk}}$$

for opponents;

- **terms involving dyadic attributes:** These are terms commonly found in exponential random graph models such as the difference in age between two nodes.

Sparsity is in the models selected is appealing as there exists sparsity of data. That is, many offense-offense, offense-defense and defense-defense combinations rarely appear in the data. Hence many interaction terms should be set to zero. Interacting sections of the latent networks permits connectedness on defense to influence connectedness on offense and vice versa. The use of relational events in the prior allows observed events to nudge the model towards those concomitant with the events. For example, if the corresponding coefficient is positive, the number of assists between two players, the higher the probability their offensive interaction is selected. More examples of possible sufficient statistics can be found in [59].

The dyads included in the model relate to observed relational data. As discussed in Section 7.2, specifying the prior probability as an exponential random graph models Ψ in terms of Y indirectly. Equation 7.3.12 establishes the prior. Choices of the sufficient statistics have also been introduced. More specifically, the model m includes models for the offense-offense network, Ψ^{O-O}, the defense-defense network, Ψ^{X-X}, and the offense-defense network, Ψ^{O-X}. Then

$$m = (m^{O-O}, m^{X-X}, m^{O-X}) \in \mathcal{M}^{O-O} \times \mathcal{M}^{X-X} \times \mathcal{M}^{O-X}$$

where each $\mathcal{M}^t$ is the set of feasible models for dyads of type t. If the sufficient statistics in the exponential random graph model governing the three networks are separable, then the priors over these three submatrices are independent. This implies selection of a dyadic offensive term is independent of selection of a dyadic defensive term. A violation of this condition is demonstrated in the following example.

Dependence between Prior Distributions Suppose the vector of sufficient statistics $g(m|\theta, Y)$ contains just the two terms

$$(g_1, g_2) = \left(\sum_{i<j} m_{ij}^{O-O} m_{ij}^{X-X}, \sum_{i \neq j} m^{O-X} \right).$$

The term on the left counts the number of two-way dyads in the model; i.e., it measures reciprocity in the model space and accords with the intuitive notion that dyads who coordinate well with one task likely coordinate well with another. The term on the right is the number of offense-defense dyads included in the model. Then the prior over the model

space can be expressed as

$$\begin{aligned} p(m^{O-O}, m^{X-X}, m^{O-X}|\theta, Y) &\propto \exp\left\{\theta^\top g(m|\theta, Y)\right\} \\ &= \exp\{g_1\theta_1 + g_2\theta_2\} \\ &= \exp\{g_1\theta_1\}\exp\{g_2\theta_2\} \\ &= p(m^{O-O}, m^{X-X}|\theta, Y)p(m^{O-X}|\theta, Y). \end{aligned}$$

The offense-offense and defense-defense networks depend upon one another. Dependence between dyads implies the dependence between networks. Such nuances may be important for particular applications, but they introduce computational complexity to estimation.

Additive and Multiplicative Effects Prior

Model selection is possible without introducing a new random variable to represent the set of active interactions in the model. The value of dyads in Ψ can depend on relational events between nodes directly. An alternative to the exponential random graph is an AMEN model as developed in [57]. Consider a mixture prior such that the elements of Ψ are equal to zero with some probability; otherwise they are given by an additive-multiplicative model for relational data. Let the prior probability over the interaction terms for a symmetric matrix Ψ be

$$\begin{aligned} p(\Psi|\omega, \theta_\Psi) &= \prod_{i,j<i} p(Y_{ij}|\omega_\Psi, \omega) \\ &= \prod_{i,j<i} p(\Psi_{ij} = 0|\omega, \theta_\Psi) + p(\Psi_{ij} \neq 0|\omega, \theta_\Psi) \\ &= \prod_{i,j<i} \delta_{ij}(\omega)\mathbb{I}\left\{\Psi_{ij} = 0\right\} + (1 - \delta_{ij}(\omega))P(\Psi_{ij}|\Psi \neq 0, \theta_\Psi). \end{aligned} \tag{7.3.13}$$

This mixture distribution places an atom of mass at zero. The probability that dyad ij is excluded from the model is $\delta_{ij} = \frac{1}{1+\exp\{-\omega n_{ij}\}}$ where n_{ij} is the number of possessions during which i and j shared the court. This type of prior will simultaneously perform model selection while informing the value of Ψ via relational data.

The continuous probability density for the value of Ψ when it is nonzero is given by the following equations. For dyad ij of type offense-offense or defense-defense, suppose

$$\Psi_{i,j} = \theta_\Psi^\top Y_{i,j} + \mathbf{u}_i^\top \mathbf{\Lambda} \mathbf{u}_j + \epsilon_{i,j} \tag{7.3.14}$$

with $\epsilon \sim N(0, \Sigma_\epsilon)$ where

$$\Sigma_{\epsilon,ij} = \begin{cases} 1 & \text{if } i = j \\ \rho & \text{else} \end{cases}$$

. Then

$$p\left(\Psi_{ij} = \psi | Y; \theta_\Psi, \Sigma_\epsilon\right) = N(\theta_\Psi^\top Y_{i,j} + \mathbf{u}_i^\top \mathbf{\Lambda} \mathbf{u}_j, \sigma_\epsilon^2).$$

While for dyad ij of type offense-defense the prior specifies that

$$\Psi_{i,j} = \theta_\Psi^\top Y_{i,j} + \mathbf{u}_i^\top \mathbf{v}_j + \epsilon_{i,j}. \tag{7.3.15}$$

Under the same assumptions for the error structure,

$$p(\Psi_{ij} = \psi | Y; \theta_\Psi) = N(\theta_\Psi^\top Y_{i,j} + \mathbf{u}_i^\top \mathbf{v}_j, \sigma_\epsilon^2).$$

Suppose the systematic variation not captured by the dyadic features or the individual effects is represented by the symmetric matrix M. Then the eigenvalue decomposition implies

$$M = U \Lambda U^\top$$

where Λ is a diagonal matrix of eigenvalues and U is an orthonormal matrix of eigenvectors. Then the product

$$\mathbf{u}_i^\top \mathbf{\Lambda} \mathbf{u}_j$$

is the effect of the latent factors $\mathbf{u}_i$ and $\mathbf{u}_j$. For the asymmetric offense-defense matrix, the latent factors can be expressed as

$$\mathbf{u}_i^\top \mathbf{v}_j$$

where

$$M = UV^\top$$

for square matrices. Finally the terms used in $Y_{i,j}$ can be the same as those which appear in the exponential random graph prior.

The advantages are that the network prior is real-valued. Hence it models the values of Ψ directly. The coefficients θ therefore indicate not the effect on the probability an edge is included, but the effect on the value of the dyad itself. This procedure, however, increases the number of parameters in an already large model. Equation 7.3.13 contains the random variable δ. Assumptions on the mass at zero affect the number of new parameters. Suppose a matrix of δ are estimated, then $N(N-1)$ parameters are added to the model. 7.3.14 introduces a latent vector for every node in the population while 7.3.15 introduces two latent vectors for every node. If this is too demanding for the data, then the latent factors can be omitted. Furthermore, if model selection is not necessary, then prior simply becomes

$$\Psi_{ij} \sim N(\theta_{\Psi}^{\top} Y_{i,j}, \sigma_{\epsilon}^{2}).$$

Then the model has the same number of parameters as the exponential random graph model. The crucial difference is the role relational data plays in determining Ψ.

7.3.4 The Prior for Main and Dyadic Effects

The model exists to provide estimates of node and dyad effects. Regularized plus-minus methods shrink parameter values to improve the model fit. Models like the elastic net, set a number of terms equal to zero while shrinking the rest of the coefficients. Let β and Ψ have normal priors centered at zero with variance τ. That is, for all nodes $v \in V$,

$$p(\beta_v|\tau) = N(0, \tau)$$

and for all dyads $v, v' \in V$ with $v' \neq v$,

$$p(\Psi_{vv'}|m, \tau) = \begin{cases} 1 & m_{vv'} = 0 \wedge \Psi_{vv'} = 0 \\ 0 & m_{vv'} = 0 \wedge \Psi_{vv'} \neq 0 \\ N(0, \tau) & m_{vv'} = 1 \end{cases}$$

so that for $t \in \{O-O, X-X\}$,

$$\begin{aligned} p(\Psi^t|m, \tau) &= \prod_{(i,j):i<j, m_{ij}=1} p(\Psi_{ij}^t|m, \tau) \\ &= \prod_{(i,j):i<j, m_{ij}=1} N(\Psi_{ij}^t; 0, \tau) \end{aligned}$$

and

$$p(\Psi^{O-X}|m,\tau) = \prod_{(i,j):i\neq j,m_{ij}=1} p(\Psi_{ij}^{O-X}|m,\tau)$$
$$= \prod_{(i,j):i\neq j,m_{ij}=1} N(\Psi_{ij}^{O-X};0,\tau).$$

Since $(\Psi^{O-X})^\top = \Psi^{X-O}$, $p(\Psi^{O-X}|m,\tau) = p(\Psi^{X-O}|m,\tau)$. Note the same shrinkage parameter applies to the main and dyadic effects. CoordiNet and TriadNet share this features with methods like the Lasso or the Elastic net; a single λ is chosen to regularize all variables. Section 7.6 reviews alternative choices for a fixed τ and proposes a method to estimate the shrinkage parameter as one would any other hyperparameter. This formulation shrinks coefficients towards zero; i.e., it assumes nodes and dyads have no effect in expectation. If estimates from prior periods of play are available, they can be used in place of zero. Posterior samples are shrunk towards values estimated in prior periods.

7.3.5 The Model Selection Hyperparameter

Exponential Random Graph Prior

The vector of sufficient statistics θ drives the model selection in the methodology. Two methods to estimate θ are discussed in this subsection. The first choice is to make the model fully Bayesian by placing a prior distribution on θ. Many alternative have been proposed in the literature on Bayesian analysis of statistical networks. The conjugate prior is a standard choice. Assume

$$p(\theta|\nu,n_0) \propto \exp\left\{\theta^\top \nu - n_0 A(\theta)\right\} \tag{7.3.16}$$

where $A(\theta) = \log(Z(\theta))$. With the conjugate prior, the marginal prior of m is given by

$$p(m|Y) = \int_\Theta p(m,\theta|Y)\,d\theta$$
$$= \int_\Theta p(m|\theta,Y)p(\theta|\nu,n_0)\,d\theta$$
$$= \int_\Theta \exp\left\{\left(\nu + g\left(m\,\middle|\,\{Y_g\}\right)\right)^\top \theta - (n_0+1)A(\theta)\right\} d\theta.$$

The posterior distribution for the θ is

$$p(\theta|R) = \sum_{m\in\mathcal{M}} p(\theta,m|R)$$
$$\propto \sum_{m\in\mathcal{M}} p(R|m)p(m|\theta)p(\theta)$$

where $p(R|m)$ is the expression in 7.3.1. But, due to factors such as computational efficiency, θ is estimated by Markov chain Monte Carlo maximum likelihood developed in [39]. The sampler treats the current model as if it were data. Then given the current model and relational data Y, choose

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \mathcal{L}(m_{\text{current}}; \theta, Y).$$

Doing this is equivalent to sampling the most likely part of the conditional distribution. For large networks this method confers substantial computational benefit.

Dyadic Independent Models When the prior distribution for over the model space satisfies dyadic independence, then it is equivalent to logistic regression. In this case, standard methods can be used to apply priors such as the prior suggested in [38]. For such models, the choice between the two should be made on the availability of prior information. A simple choice is

$$p(\theta) = N(0, \nu^2)$$

with ν^2 fixed.

The Additive Multiplicative Latent Factor Model

Recall the prior introduced in 7.3.13 such that the offense-offense and defense-defense interaction terms satisfy

$$\Psi_{ij} = \theta_\Psi^\top Y_{ij} + \mathbf{u}_i^\top \mathbf{\Lambda} \mathbf{u}_j + \epsilon_{i,j}$$

with

$$(\epsilon_{i,j}, \epsilon_{j,i}) \sim N\left(\mathbf{0}, \sigma_\epsilon^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

where $\mathbf{u}_i$ and $\mathbf{u}_j$ are latent factors, and

$$\theta_\Psi \sim N(0, I).$$

The priors over the vector of parameters θ_Ψ is

$$p(\theta_\Psi = t) = N(t; 0, \sigma_\theta^2).$$

[58] develops a method to simulate orthonormal matrices capturing the latent factors. Given the complexity of the model a low-rank approximation is sufficient to capture unobserved data relating pairs in V. A rank-one approximation is simply

$$\lambda_1 uu^\top$$

where u is the first eigenvector of Z. The prior for the matrix Z is

$$\begin{aligned} p(Z|U,\Lambda) &= p(U\Lambda U^\top + E|U,\Lambda) \\ &= p(E|U,\Lambda) \\ &\propto \exp\left\{\text{tr}\frac{-(Z-U\Lambda U^\top)^\top(Z-U\Lambda U^\top)}{4}\right\} \end{aligned}$$

The prior for the diagonal matrix Λ is

$$p(\Lambda|Z,U) = \prod_{r=1}^{R} N\left(\lambda_r; \frac{\tau^2 U_r^\top Z U_r}{(2+\tau^2)}, \frac{2\tau^2}{(2+\tau^2)}\right).$$

And, the prior for orthonormal matrix U is

$$p(U|Z,\Lambda) \propto \exp\left\{\text{tr}\frac{Z^\top U\Lambda U^\top}{2}\right\}$$

The offense-defense matrix is asymmetric. Hence the latent factors will be as well. Then

$$p(Z|U,V) \propto \exp\left\{\text{tr}\frac{-(Z-U\Lambda V^\top)^\top(Z-U\Lambda V^\top)}{4}\right\}$$

The prior distributions on U and V are uniform on the space of orthonormal $N \times R$ matrices.

7.3.6 Alterations for *TriadNet*

Denote the cube of three-term interactions as Φ. The main issues involve the prior over the additional parameters. One particularly simple strategy is to suppose Φ is an exponential random graph given Ψ; i.e.,

$$p(\Phi|\Psi) \propto \exp\left\{\zeta^\top h(\Phi = \phi|\Psi, Y)\right\}. \tag{7.3.17}$$

Since the prior for Φ is set conditionally, it can be used to introduce hierarchy conditions to the set of interactions included in the model. By assumption, the triadic model includes all main effects, but not all dyadic interaction effects. Hence model hierarchy depends upon which dyadic effect terms are included in a particular model. Consider

$$p(\Phi|\Psi) \propto \mathbb{I}\{\mathcal{H}(\Phi|\Psi) = 1\} \exp\left\{\zeta^\top h(\Phi = \phi|\Psi, Y)\right\} \tag{7.3.18}$$

where $\mathcal{H}(\Phi|\Psi)$ indicates whether hierarchy condition $\mathcal{H}$ holds for Φ given Ψ. If the prior values of Φ only depend on whether the terms in Ψ are nonzero, the distribution can be written

$$p(\Phi|m_\Psi, \zeta) \propto \mathbb{I}\{\mathcal{H}(\Phi|m_\Psi) = 1\} \exp\left\{\zeta^\top h(\Phi|m_\Psi, Y)\right\}.$$

if hierarchy is encoded into the sufficient statistics $h(m_\Phi|m_\Psi, Y)$, then the prior is given by 7.3.17 with sufficient statistics such as an indicator for a hierarchy conditions or the proportion of terms that satisfy the hierarchy condition. Relational data in *CoordiNet* records events that transpire between two individuals. In a triadic model, ternary relational events are used to inform the prior over models. Interestingly, many ternary motifs or relational events can be decomposed into binary events. A triangle in a social network provides an example. Let i, j and k form a triangle if and only if $Y_{ij} = Y_{jk} = Y_{ik} = 1$. A double team in basketball follows the same logic; let i and j be a double team k if and only if i and j defend k. This observation aids in the intuition of a triadic term learned by *TriadNet*. A triadic term captures systematic variation that cannot be decomposed into the dyads that comprise it.

Dependence between m_Ψ and m_Φ Assumptions about the interaction hierarchy conditions to which the model is subject induces dependency between Ψ and Φ. But, one can imagine various mechanisms connecting these random variables.

The priors over the values of the selected triadic interaction terms is

$$p(\Phi_{vv'v''}|m,\tau) = \begin{cases} 1 & m_{vv'v''} = 0 \wedge \Phi_{vv'} = 0 \\ 0 & m_{vv'v''} = 0 \wedge \Phi_{vv'} \neq 0 \\ N(0,\tau) & m_{vv'v''} = 1 \end{cases}$$

Hence the joint prior for m_Φ, m_Ψ, Φ and Ψ given θ, ζ and, Y will be an exponential random graph. A conjugate prior for ζ can be assessed or the maximum likelihood estimate can be used within the sampling scheme. The next section reviews the MCMC sampling strategy.

7.4 Posterior Inference and Sampling

Estimation procedures for the dyadic and triadic interaction models are developed. The former obtains a sample from $\{p(\Psi,\beta,\alpha,m)\}_{m\in\mathcal{M}}$ and the latter obtains a sample from $\{p(\Phi,\Psi,\beta,\alpha,m_\Phi,m_\Psi)\}_{m_\Phi,m_\Psi\in\mathcal{M}}$.

7.4.1 Posteriors of Interest

In this subsection the objects of primary interest are reviewed and methods to sample from the posterior distribution of each are presented. Focus resides on three posterior computations: computation of model probabilities, model-averaged parameter estimates and the posterior predictions. Comparison of model probabilities requires good estimates of the posterior distribution of the parameters dependent on a given model m. Hence whenever the sampler moves to a new model, it must sample from the conditional distribution of the relevant parameters. For each m_i with $i = 1,\ldots,D_m$ draw $(\alpha_i^{(j)},\beta_i^{(j)},\Psi_i^{(j)})$ for $j = 1,\ldots,S$ from the conditional posterior $p(\alpha,\beta,\Psi|m,R)$ and compute

$$\begin{aligned} p(R|m_i) &= \int\int\int p(R|\alpha_i^{(j)},\beta_i^{(j)},\Psi_i^{(j)},m_i)p(\alpha)p(\beta)p(\Psi|m)\,d\alpha d\beta d\Psi \\ &\approx \sum_j p(R|\alpha_i^{(j)},\beta_i^{(j)},\Psi_i^{(j)},m_i)p(\alpha_i^{(j)})p(\beta_i^{(j)})p(\Psi_i^{(j)}|m_i). \end{aligned}$$

Then, combining all sampled models, compute the posterior model probability

$$p(m_i|R) \approx \frac{p(R|m_i)p(m_i)}{\sum_{i=1}^{D_m} p(R|m_i)p(m_i)}.$$

The sampled models can be ranked according to their probabilities. The parameters sampled at each model can be used to infer characteristics of the posterior distribution of highly likely models. Usually, the parameters β and Ψ are of greater interest than α. The joint posterior of these parameters can be computed for a given model, or the

model-averaged parameters can be used to analyze the main and dyadic effects. The latter can be computed by

$$\begin{aligned} p(\Psi, \beta | R) &= \sum_{i=1}^{M} \int p(\alpha, \beta, \Psi | m_i, R) p(m_i | R) \, d\alpha \\ &\approx \sum_{i=1}^{M} \sum_{j} p(\alpha_i^{(j)}, \beta, \Psi | m_i, R) p(m_i | R) \end{aligned}$$

Posterior Predictive

Samples drawn from the posterior distribution of the main effects, β, the network Ψ and the cut-point parameters α can be used to compute the posterior predictive distribution. Given the sequence of values $R_1, \ldots, R_L$, a new point with design $\tilde{w}$ is distributed as

$$\begin{aligned} p(\tilde{R} | \tilde{w}, R_1, \ldots, R_L, W) &= \int \int \int p(\tilde{R}, \alpha, \beta, \Psi | \tilde{w}, R, W) \, d\alpha d\beta d\Psi \\ &\approx \frac{\sum_{j=1}^{S} p(\tilde{R} | \alpha^{(j)}, \beta^{(j)}, \Psi^{(j)}) \prod_{l=1}^{L} p(R_l | \alpha^{(j)}, \beta^{(j)}, \Psi^{(j)}) p(\alpha^{(j)}, \beta^{(j)}, \Psi^{(j)})}{\sum_{j=1}^{S} \prod_{l=1}^{L} p(R_l | \alpha^{(j)}, \beta^{(j)}, \Psi^{(j)}) p(\alpha^{(j)}, \beta^{(j)}, \Psi^{(j)})} \end{aligned} \tag{7.4.1}$$

where $\left\{\alpha^{(j)}, \beta^{(j)}, \Psi^{(j)}\right\}_{j=1}^{S}$ are samples from the posterior distributions of α, β and Ψ, respectively. Note the models have been averaged out of this expression. 7.4.1 can be used to compare the performance of *CoordiNet* to other modeling techniques.

7.4.2 Algorithm Details

Gibbs sampling provides a method to draw from the posterior distribution of $p(\alpha, \beta, \Psi, m | R)$. In order to average over models, the sampler must explore the space of models and explore the parameter space for every model considered. A Gibbs sampler of the conditional distribution $p(\alpha, \beta, \Psi | m, R)$ is regular and therefore converges to the unique stationary distribution.

Choose initial values for the random variables $\alpha, \beta^O, \beta^X, \Psi^{O-O}, \Psi^{X-X}, \Psi^{O-X}, \theta^O, \theta^X, \theta^{OX}, \tau$. Note that M^{O-O} and M^{X-X} are implicitly defined as $M_{ij} = \mathbb{I}\left\{\Psi_{ij} \neq 0\right\}$ for $i < j$, $i, j \in V$. Consider a Gibbs sampling algorithm for the model. To sample from $p(\theta | m, \Psi, \beta, \alpha, R)$ when $r < r_1$, solve

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \, \mathcal{L}(m; \theta, Y)$$

and set $Q(\theta'|\theta) = N(\hat{\theta}, \sigma_\theta^2)$. Otherwise, if $r > r_1$, set $Q(\theta'|\theta) = N(\theta, \sigma_\theta^2)$. For the latter condition, the acceptance ratio for the offense-offense and defense-defense dyadic effect models is

$$\frac{p(m|\theta')p(\theta')Q(\theta|\theta')}{p(m|\theta)p(\theta)Q(\theta'|\theta)} = \frac{\prod_{i<j}\left(\pi'_{ij}\right)^{m_{ij}}(1-\pi'_{ij})^{1-m_{ij}}N(\theta';0,\sigma_\theta^2)}{\prod_{i<j}\left(\pi_{ij}\right)^{m_{ij}}(1-\pi_{ij})^{1-m_{ij}}N(\theta;0,\sigma_\theta^2)}$$

where $\pi_{ij} = \frac{\exp\{\theta^\top g_i j\}}{1+\exp\{\theta^\top g_i j\}}$. This simplifies to

$$\prod_k \exp\left\{-\frac{\left((\theta'_k)^2-\theta_k^2\right)}{2\sigma_\theta^2}\right\}\prod_{i<j}\left(\frac{\exp\left\{-(\theta)^\top g_{ij}\right\}+1}{\exp\left\{-(\theta')^\top g_{ij}\right\}+1}\right)^{m_{ij}}\left(\frac{\exp\left\{(\theta)^\top g_{ij}\right\}+1}{\exp\left\{(\theta')^\top g_{ij}\right\}+1}\right)^{1-m_{ij}}.$$

After updating θ^{O-O}, θ^{X-X} and θ^{O-X}, perform a Gibbs run through each feasible dyad in Ψ^{O-O}, Ψ^{X-X} and Ψ^{O-X}. For each model, sample m and Ψ jointly to improve mixing. For some fixed $\delta > 0$ propose $\Psi'_{ij} = 0$ so that $m'_{ij} = 0$; otherwise propose $\Psi' \sim N(\Psi, \sigma_\Psi^2)$ and $m'_{ij} = 1$.

In to Out Suppose the latent networks Ψ and Ψ' differ by a single edge. Let $M_{ij} = 1$ so that $\Psi_{ij} \neq 0$, and $M'_{ij} = 0$ so that $\Psi'_{ij} = 0$. Then the acceptance probability is

$$\begin{aligned}
\frac{p(\Psi', M'|\beta,\alpha,R)Q(\Psi|\Psi')}{p(\Psi, M|\beta,\alpha,R)Q(\Psi'|\Psi)} &= \frac{p(R|\Psi',\beta,\alpha)p(\Psi'|M',\tau)p(M'|\theta)}{p(R|\Psi,\beta,\alpha)p(\Psi|M,\tau)p(M|\theta)}\frac{(1-\delta)N(\Psi_{ij};\Psi'_{ij},\sigma_\Psi^2)}{\delta} \\
&= \frac{\prod_{l\in L_{ij}}p(R_l|\Psi',\alpha,\beta)}{\prod_{l\in L_{ij}}p(R_t|\Psi,\alpha,\beta)}\frac{p(\Psi'_{ij}=0|M'_{ij}=0,\tau)}{p(\Psi_{ij}=\psi|M_{ij}=1,\tau)}\frac{\exp\{\theta^\top f(M')\}}{\exp\{\theta^\top f(M)\}}\frac{(1-\delta)N(\Psi_{ij};\Psi'_{ij},\sigma_\Psi^2)}{\delta} \\
&= \exp\left\{-\theta^\top f(M_{ij})\right\}\frac{(1-\delta)}{\delta}\prod_{l\in L_{ij}}\left(\frac{p(R_l|\Psi',\alpha,\beta)}{p(R_l|\Psi,\alpha,\beta)}\right)\frac{N(\Psi_{ij};0,\sigma_\Psi^2)}{N(\Psi_{ij};0,\tau)}
\end{aligned}$$

where the last line holds whenever the network prior satisfies dyadic independence and L_{ij} is the set of lineups containing dyad ij.

Out to In Let $M_{ij} = 0$ so that $\Psi_{ij} = 0$, and $M'_{ij} = 1$ so that $\Psi'_{ij} \neq 0$. Then the acceptance probability is

$$\frac{p(\Psi', M'|\beta,\alpha,R)Q(\Psi|\Psi')}{p(\Psi, M|\beta,\alpha,R)Q(\Psi'|\Psi)} = \exp\left\{\theta^\top f(M_{ij})\right\}\frac{\delta}{(1-\delta)}\prod_{l\in L_{ij}}\left(\frac{p(R_l|\Psi',\alpha,\beta)}{p(R_l|\Psi,\alpha,\beta)}\right)\frac{N(\Psi'_{ij};0,\tau)}{N(\Psi'_{ij};0,\sigma_\Psi^2)}$$

In to In In this case $M_{ij} = M'_{ij} = 1$, and $\Psi_{ij} \neq \Psi'_{ij}$. The acceptance probability becomes

$$\frac{p(\Psi', M'|\beta, \alpha, R)Q(\Psi|\Psi')}{p(\Psi, M|\beta, \alpha, R)Q(\Psi'|\Psi)} = \prod_{l \in L_{ij}} \left(\frac{p(R_l|\Psi', \alpha, \beta)}{p(R_l|\Psi, \alpha, \beta)} \right) \frac{N(\Psi'_{ij}; 0, \tau)}{N(\Psi_{ij}; 0, \tau)}$$

Then many draws are taken given the fixed model. Doing so provides the samples necessary to compute model probabilities. Sampling for the Ψ and β parameters is straightforward. Sample from the conditionals

$$p(\Psi_{ij}|\Psi_{-ij}, \tau, \beta, \alpha, \theta)$$

and

$$p(\beta_i|\beta_{-i}, \tau, \psi, \alpha, \theta).$$

Sample α by sampling proportions of observations falling into each of the categories π' from the $J - 1$ unit simplex. Observe that $p(\pi|R, \Psi)$ can be sampled via a run of Metropolis-Hastings with proposal distribution

$$q_\epsilon(\pi'|\pi) = \frac{\Gamma(J)}{\Gamma(1)^J} \mathbb{I}\{d(\pi, \pi') \leq \epsilon\}$$

and accepted with probability

$$\min\left(1, \frac{p(\pi'|\Psi, R)q_\epsilon(\pi|\pi')}{p(\pi)q_\epsilon(\pi'|\pi)}\right) = \min\left(1, \frac{p(\pi')}{p(\pi)}\right).$$

Given $d(\pi', \pi) \leq \epsilon$ the instrumental densities cancel. Then the parameters are transformed.

Note this choice is not necessary. One could sample m without Ψ. Doing so can lead to many dyads ij with $m_{ij} = 1$, but $\Psi = 0$. The acceptance probability for these models.

1. Sample τ' from $p(\tau|R, \alpha, \beta^O, \beta^X, \Psi^{O-O}, \Psi^{X-X}, \theta^O, \theta^X)$.

2. Sample Ψ (and implicitly sample M'). Iterate over each dyad in the network and with probability $\delta > 0$ set $\Psi'_{ij} = 0$. With probability $(1 - \delta)$ sample Ψ'_{ij} from

$$p(\Psi_{ij}|\Psi_{-ij}, \tau, \beta, \alpha, \theta) \tag{7.4.2}$$

3. Iterate over each node and sample β_i from

$$p(\beta_i|\beta_{-i}, \tau, \psi, \alpha, \theta) \tag{7.4.3}$$

4. Sample α by sampling proportions of observations falling into each of the categories π' from the $J-1$ unit simplex. Observe that $p(\pi|R, \Psi)$ can be sampled via a run of Metropolis-Hastings with proposal distribution

$$q_\epsilon(\pi'|\pi) = \frac{\Gamma(J)}{\Gamma(1)^J}\mathbb{I}\{d(\pi, \pi') \leq \epsilon\}$$

and accepted with probability

$$\min\left(1, \frac{p(\pi'|\Psi, R)q_\epsilon(\pi|\pi')}{p(\pi)q_\epsilon(\pi'|\pi)}\right) = \min\left(1, \frac{p(\pi')}{p(\pi)}\right).$$

Given $d(\pi', \pi) \leq \epsilon$ the instrumental densities cancel. Then the parameters are transformed.

5. Update θ by maximum likelihood estimation of the simulated values or the appropriate conditional posterior. Again, Metropolis-Hastings can be used.

7.5 Application: Eastern Conference Semifinals

CoordiNet was applied to the play-by-play data generated during the Eastern Conference Semifinals in 2016-17 in the NBA. Two teams played seven games against one another. The prior for the main effects used data on the player's attributes like; the prior for the dyads is based upon an intercept and the total height of a pair. Results for the additive-multiplicative network prior are demonstrated. Figure **??** presents a histogram of the number of terms selected. Figures 7.6 and 7.7 show a run of MCMC for dyadic and main effects. Some of the information can be presented in network plots with the color of node indicative of the sign of the estimated main effect. Edges are colored green if their parameters are positive and red otherwise. Dyads not included in the model do not share an edge. Figures 7.8 and 7.9 display network plots for the Eastern Conference Semifinals.

7.5.1 Prediction of Counterfactual Dyads

Prediction of the response for dyads that were actually observed can be accomplished by sampling from 7.4.1. This type of prediction implicitly assumes the potential outcome of a lineup is independent of the other lineups in which they were observed. Under this assumption, every lineup composed of dyads that have played together in the observed play-by-play can be computed.

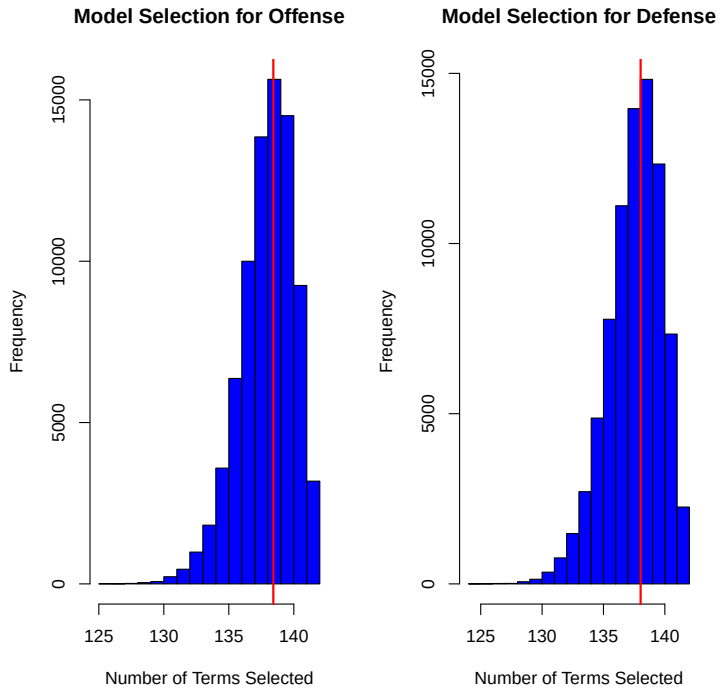

Figure 7.5: Selection of terms in the direct prior *CoordiNet*.

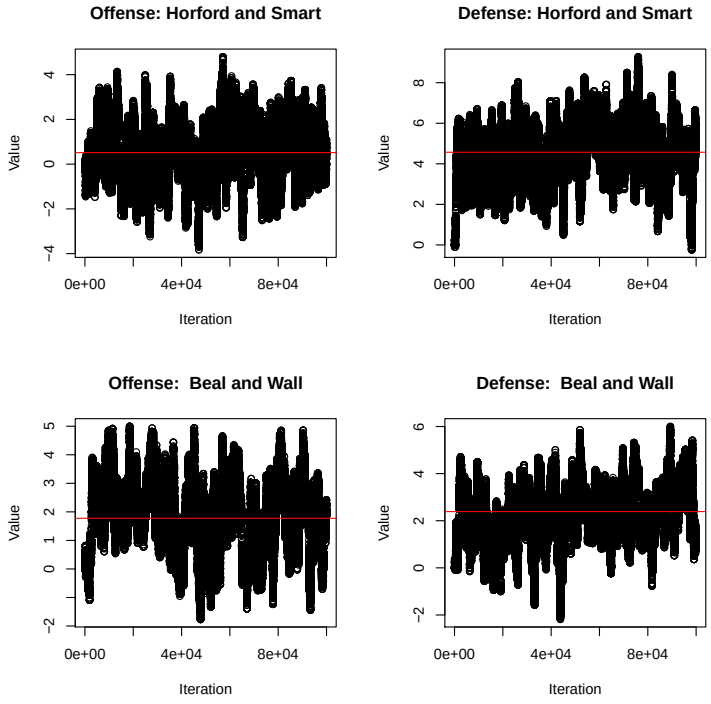

Figure 7.6: Posterior draws for dyadic terms in the direct prior *CoordiNet*.

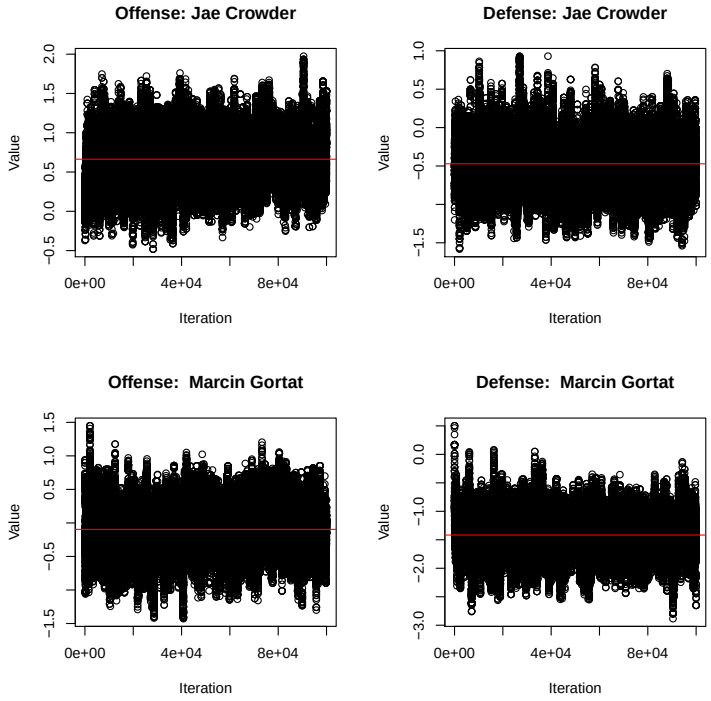

Figure 7.7: Posterior draws for main effects in the direct prior *CoordiNet*.

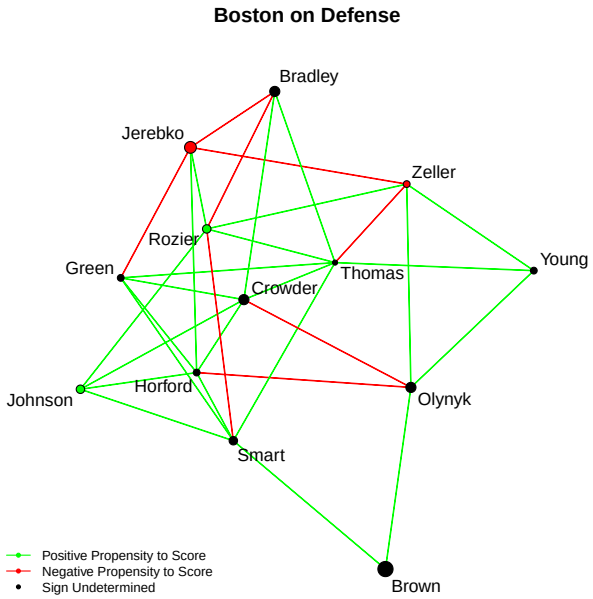

Figure 7.8: Boston's inferred defensive network via the direct prior *CoordiNet*.

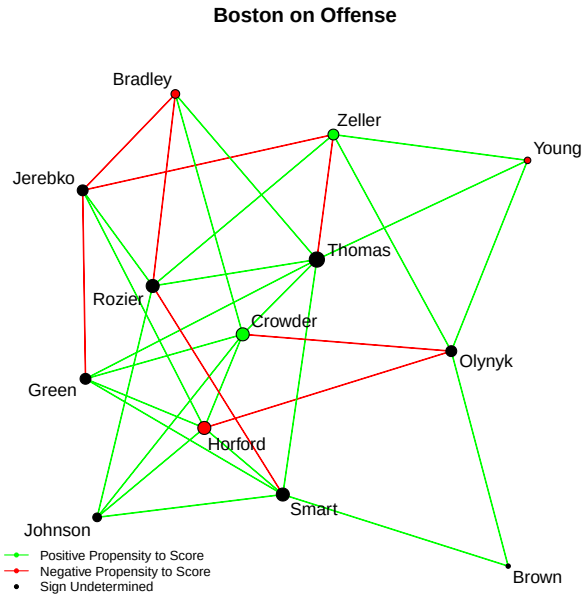

Figure 7.9: Boston's inferred offensive network via the direct prior *CoordiNet*.

Prediction of counterfactual lineups composed of players who have never played with one another requires stronger assumptions. The exponential random graph prior model can provide priors over the probability of an edge, but it cannot determine the sign and magnitude. In the Normal model, there is no object representing the model. Suppose the prior for dyad ij is

$$p(\Psi_{ij}|x) \sim N(x_{ij}^{\top}\theta, \sigma^2)$$

and $\left\{\theta^{(j)}\right\}_j$ are samples from $p(\theta|R)$. Then for counterfactual dyad Ψ' with associated data x',

$$p(\Psi'|R, x') \approx \sum_{j=1}^{S} p(\Psi, \theta^{(j)}|R).$$

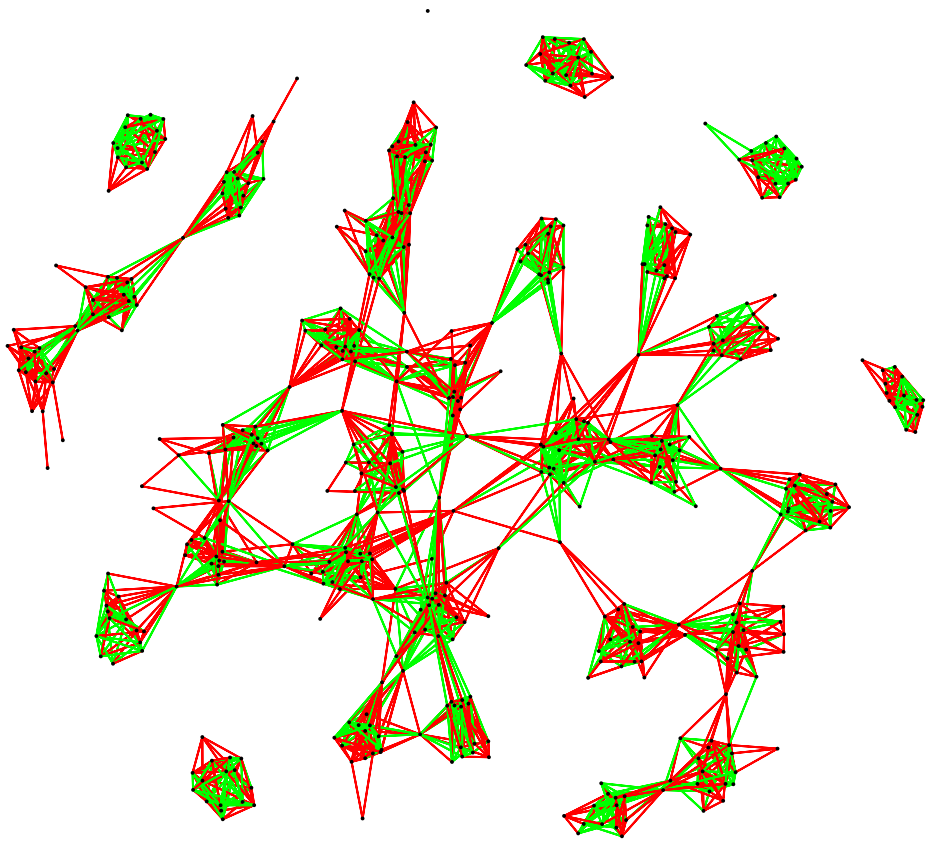

Figure 7.10: League Plot

7.6 Discussion

This chapter outlines methods to estimate models with dyadic and triadic interactions, and fits the former using the Eastern Conference Semifinals data from 2016-17. For *CoordiNet* two priors are suggested for dyadic interactions. In the first the prior over the dyads included in the regression is an exponential random graph. This has the feature of modeling dependence between the dyads included in the model. The second prior assumes dyads are distributed independently conditioned on data x. No model term must be introduced in this version. Moreover, this method provides a way to simulate counterfactual dyads.

There are several issues which merit further investigation. The model assumes the responses $\{R_l\}$ are distributed independently of one another. A more general model allows responses to be correlated in time. Although the proportional odds logistic regression provides interpretable parameters, there are some issues associated with it. One obvious problem with this formulation is that lineups can score four points in a single possession frequently, but may perform poorly otherwise so the sum of their parameters is quite low. They will be expected to score four points fewer times than a lineup who never scored four points. This can be an issue when certain lineups are used strategically for short periods of time. Some lineups specialize in scoring one or two points while others score three or none. The assumption of ordinal data obscures this phenomenon. One solution is to remove the assumption that the data is ordinal. In this case far more parameters are required.

The estimated parameters provide a network structure. They can be analyzed to create measures of centrality within teams. Let $g = 1, \ldots, G$ be a collection of teams to which the nodes belong. Consider the spectral decomposition of the dyads contained within a team. For all g and $\Psi_g \in \left\{\Psi_g^{O-O}, \Psi_g^{X-X}\right\}$, the component associated with the largest eigenvalue of an adjacency matrix is called the eigenvalue centrality, the best one-dimensional approximation of the adjacency matrix. Intuitively those with positive values are highly connected to those with positive values.

The relational events used to inform the prior are assumed to non-stochastic in the model. But, this is clearly not a realistic assumption. The outcomes of possessions result from the myriad actions that occur. Not only are these actions indicative of important dyadic and triadic effects, the efficacy of relational events vary across the dyads. *CoordiNet* and *TriadNet* assume relational events are only associated with the dyads included in the model. Taking these events seriously enables modeling of different sequences of events with different lineups. Outcomes can be modeled as a function of the relational events and the lineup in which it occurs.

7.7 Appendix

7.7.1 The Prior for τ

In the exponential random graph model the components of the parameters are distributed independently and normally around zero with standard deviation τ. Since τ is a hyperparameter - a parameter in a prior distribution - its value can be fixed or can vary according to a prior distribution. Results in the main text use a fixed value of τ to reduce computation. Alternatively, let the prior for τ be

$$p(\tau; \underline{\tau}, \overline{\tau}) = \frac{\mathbb{I}\left\{\tau \in [\underline{\tau}, \overline{\tau}]\right\}}{\overline{\tau} - \underline{\tau}}.$$

Propose all values within 2ϵ of τ with equal probability. Define the proposal distribution as

$$q(\tau'|\tau) = \begin{cases} \frac{\mathbb{I}\{\tau \in (0, \tau+\epsilon]\}}{2\epsilon - \tau} & \tau < \epsilon \\ \frac{\mathbb{I}\{\tau \in (\tau-\epsilon, \tau+\epsilon)\}}{2\epsilon} & \overline{\tau} - \epsilon \geq \tau \geq \epsilon \\ \frac{\mathbb{I}\{\tau \in (\tau-\epsilon, \overline{\tau}]\}}{2\epsilon - (\overline{\tau} - \tau)} & \overline{\tau} - \epsilon < \tau \end{cases}$$

There are 5 types of moves possible in the sampling scheme. Each has a different scheme.

Lower Boundary to Interior Accept a proposal with probability

$$\begin{aligned} \nu &= \min\left\{1, \frac{p(\Psi, \beta|\tau')p(\tau'; \overline{\tau})q(\tau'|\tau)}{p(\Psi, \beta|\tau)p(\tau; \overline{\tau})q(\tau|\tau')}\right\} \\ &= \min\left\{1, \frac{p(\Psi, \beta|\tau')(2\epsilon - \tau)}{p(\Psi, \beta|\tau)2\epsilon}\right\}. \end{aligned}$$

Interior to Lower Boundary Accept a proposal with probability

$$\begin{aligned} \nu &= \min\left\{1, \frac{p(\Psi, \beta|\tau')p(\tau'; \overline{\tau})q(\tau'|\tau)}{p(\Psi, \beta|\tau)p(\tau; \overline{\tau})q(\tau|\tau')}\right\} \\ &= \min\left\{1, \frac{p(\Psi, \beta|\tau')(2\epsilon)}{p(\Psi, \beta|\tau)(2\epsilon - \tau)}\right\}. \end{aligned}$$

Interior to Interior Accept a proposal with probability

$$\nu = \min\left\{1, \frac{p(\Psi, \beta|\tau')}{p(\Psi, \beta|\tau)}\right\}.$$

Upper Boundary to Interior Accept a proposal with probability

$$\nu = \min\left\{1, \frac{p(\Psi, \beta|\tau')(2\epsilon - (\overline{\tau} - \tau))}{p(\Psi, \beta|\tau)(2\epsilon)}\right\}.$$

Interior to Lower Boundary Accept a proposal with probability

$$\nu = \min\left\{1, \frac{p(\Psi, \beta|\tau')(2\epsilon)}{p(\Psi, \beta|\tau)(2\epsilon - (\overline{\tau} - \tau))}\right\}.$$

The likelihood in the acceptance proposal is just the normal distribution with standard deviation τ centered at zero. Then

$$\begin{aligned} p(\Psi, \beta) &= \prod_j \frac{\tau}{\tau'} \exp\left\{-\frac{\Psi_j^2(\tau^2 - (\tau')^2)}{2\tau^2(\tau')^2}\right\} \prod_i \frac{\tau}{\tau'} \exp\left\{-\frac{\beta_i^2(\tau^2 - (\tau')^2)}{2\tau^2(\tau')^2}\right\} \\ &= \left(\frac{\tau}{\tau'}\right)^{2N+2M} \exp\left\{-\frac{\sum_i \beta_i^2(\tau^2 - (\tau')^2) + \sum_j \Psi_j^2(\tau^2 - (\tau')^2)}{2\tau^2(\tau')^2}\right\} \\ &= \left(\frac{\tau}{\tau'}\right)^{2N+2M} \exp\left\{-\frac{(\sum_i \beta_i^2 + \sum_j \Psi_j^2)(\tau^2 - (\tau')^2)}{2\tau^2(\tau')^2}\right\} \end{aligned}$$

Bibliography

[1] Waqar Ali, Anatol E. Wegner, Robert E. Gaunt, Charlotte M. Deane and Gesine Reinert. *Comparison of Large Networks with Sub-sampling Strategies*. Nature, New York, NY, 2000.

[2] Ethan Anderes, Steffen Borgwardt, Jacob Miller. *Discrete Wasserstein Barycenters: Optimal Transport for Discrete Data*. Mathematical Methods of Operations Research, 84(2):389-409, 2016.

[3] Stephen A. Ayidiya and McKee J. McClendon *Response Effects in Mail Surveys*. Public Opinion Quarterly Volume, 54:229-247, 1990.

[4] Jose M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. John Wiley & Sons, New York, NY, 2000.

[5] Christophe Biernacki and Julien Jacques. *A Generative Model for Rank Data Based on Insertion Sort Algorithm*. Computational Statistics and Data Analysis, 58:162-176, 2013.

[6] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, New York, 2006.

[7] Devon D. Brewer. *Forgetting in the Recall-Based Elicitation of Personal and Social Networks*. Social Networks, 22(1):29–43, 2000.

[8] Stephane Boucheron, Gabor Lugosi and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford university press, 2013.

[9] William A. Brock and Steven N. Durlauf. *Identification of Binary Choice Models with Social Interactions*. Journal of Econometrics, 140:52-75, 2007.

[10] William A. Brock and Steven N. Durlauf. *Discrete Choice with Social Interactions*. Review of Economic Studies, 68:235-260, 2001.

[11] Michael L. Burton and Sara B. Nerlove. *Balanced Designs for Triads Tests: Two examples from English*. Social Science Research, 5(3):247-267, 1976.

[12] Jerome R. Busemeyer and Peter D. Bruza. *Quantum Models of Cognition and Decision*. Cambridge University Press, New York, 2012.

[13] Carter T. Butts. *Network Inference, Error, and Informant (In)accuracy: A Bayesian Approach*. Social Networks, 25:103–140, 2003.

[14] Carter T. Butts. *A Relational Event Framework for Social Action*. Sociological Methodology, 38(1):155–200, 2008.

[15] Alberto Caimo, Antonietta Mira. *Efficient computational strategies for doubly intractable problems with applications to Bayesian social networks*. Statistics and Computing, 25(1):113-125, 2015.

[16] A. Colin Cameron and Pravin K. Trivedi. *Microeconomnetrics: Methods and Applications* Cambridge Universty Press, 2005.

[17] Yanshuai Cao, David J. Fleet. *Generalized Product of Experts for Automatic and Principled Fusion of Gaussian Process Predictions*. Modern Nonparametrics 3: Automating the Learning Pipeline.

[18] Miguel A. Carreira-Perpinan and Goeffrey E. Hinton. *On Contrastive Divergence Learning*. Aistats, 10:33-40, 2005.

[19] Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. *Measurement Error in Nonlinear Models*. Chapman and Hall/CRC, Boca Raton, FL, 2006.

[20] Ben Carterette. *On Rank Correlation and the Distance Between Rankings.* Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. ACM, 436-443, 2009.

[21] Shuo Chen, Thorsten Joachims. *Modeling Intransitivity in Matchup and Comparison Data*. Bernoulli, 19(4):1465-1483, 2013. DOI: http://dx.doi.org/10.1145/2835776.2835787

[22] Hugh Chipman. *Bayesian Variable Selection with Related Predictors*. Canadian Journal of Statistics, 24:17–36, 1996.

[23] Aaron Clauset, Cristopher Moore and Mark Newman. *Structural Inference of Hierarchies in Networks*. Statistical network analysis: models, issues, and new directions. Springer, Berlin, Heidelberg, 2007. 1-13.

[24] Hans Crauel. *Random Probability Measures on Polish Spaces*. Taylor and Francis, London, 2002.

[25] Marco Cuturi, Arnaud Doucet. *Fast Computation of Wasserstein Barycenters*. Proceedings of the 31st International Conference on Machine Learning, Beijing,China, 2014. JMLR:W&CPvolume 32.

[26] Bin Dai, Shilin Ding and Grace Wahba. *Multivariate Bernoulli Distribution*. Bernoulli, 19(4):1465-1483, 2013.

[27] Massimiliano Sassoli de Bianchi. *The Observer Effect*. Foundations of Science, 18(2):213-243.

[28] Bruce A. Desmarais and Skyler J. Cranmer. *Statistical Inference for Valued-Edge Networks: The Generalized Exponential Random Graph Model*. Plos One, 7(1):1-12, 2012.

[29] Persi Diaconis and Ronald L. Graham. *Spearman's Footrule as a Measure of Disarray*. Journal of the Royal Statistical Society. Series B (Methodological), 262-268, 1977.

[30] e-Source. *Sample Surveys: Developing a Survey Instrument*. esourceresearch.org/eSourceBook/SampleSurveys/6DevelopingaSurveyInstrument/tabid/484/Default.aspx, 2017.

[31] Ian Fellows and Mark Handcock. *Exponential-family Random Network Models*. arXiv preprint arXiv:1208.0121, 2012.

[32] Anuska Ferligoj and Valentina Hlebec. *Evaluation of Social Network Measurement Instruments*. Social Networks, 21:111–130, 1999.

[33] Patrick M. Fitzpatrick. *Advanced Calculus*. Thompson Brooks/Cole, Belmont, 2006.

[34] Wayne A. Fuller. *Sampling Statistics*. John Wiley & Sons, Inc., Hoboken, New Jersey, 2009.

[35] Jonah Gabry and Ben Goodrich. *Estimating Ordinal Regression Models with rstanarm*. cran.r-project.org/web/packages/rstanarm/vignettes/polr.html, 2017.

[36] Andrew Gelman. *Prior Distributions for Variance Parameters in Hierarchical Models*. Bayesian Analysis, 1(3):515-533, 2006.

[37] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, Cambridge, 2007.

[38] Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau and Yu-Sung Su. *A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models*. The Annals of Applied Statistics, 2(4):1360-1383, 2008.

[39] Charles J. Geyer. *Markov Chain Monte Carlo Maximum Likelihood.* Interface Foundation of North America. Retrieved from the University of Minnesota Digital Conservancy, http://hdl.handle.net/11299/58440, 1991.

[40] Charles J. Geyer and Elizabeth A. Thompson. *Constrained Monte Carlo Maximum Likelihood for Dependent Data.* Journal of the Royal Statistical Society. Series B (Methodological), 54(3):657-699, 1992.

[41] Krista Gile and Mark Handcock. *Respondent-Driven Sampling: An Assessment of Current Methodology*. Sociological Methodology, 40:285-327, 2010.

[42] Phillip Good. *Permutation, Parametric and Bootstrap Tests of Hypotheses*. Springer series in statistics. Statistics 400: 60535, 2004.

[43] Clarence C. Gravlee, H. Russell Bernard, Chad R. Maxwell and Aryeh Jacobsohn. *Mode Effects in Free-list Elicitation: Comparing Oral, Written, and Web-based Data Collection*. Social Science Computer Review, 31(1):119–132, 2013.

[44] Robert M. Groves. *Survey Errors and Survey Costs*. John Wiley & Sons, Inc., New York, New York, 1989.

[45] Robert M. Groves and Lars Lyberg. *Total Survey Error: Past, Present, and Future*. Public Opinion Quarterly, 74(5):849–879, 2010.

[46] Mark S. Handcock, Adrian E. Raftery and Jermey M. Tantrum. *Model-based clustering for social networks* J. R. Statist. Soc. A, 170(2):301–354, 2007.

[47] Mark S. Handcock and Krista J. Gile. *Modeling Social Networks with Sampled or Missing Data*. CSSS Working Paper 75.

[48] Mark S. Handcock and Krista J. Gile. *Modeling Social Networks From Sampled Data*. The Annals of Applied Statistics, 4(1):5-25, 2010.

[49] Mark S. Handcock and Krista J. Gile, *Supplement to "Modeling social networks from sampled data."* arXiv preprint math.PR/0000000, 2010.

[50] David I. Hastie and Peter J. Green. *Model choice using reversible jump Markov chain Monte Carlo*. Statistica Neerlandica, 66(3):309-338, 2012.

[51] Miguel A. Hernán and Jamie M. Robins. *Causal Inference*. Chapman & Hall/CRC, Boca Raton, forthcoming.

[52] Hinne, M., T. Heskes, and M. A. J. van Gerven. *Bayesian inference of whole-brain networks*. arXiv preprint arXiv:1202.1696, 2012.

[53] Hinne, M., Heskes, T., Beckmann, C. F., & van Gerven, M. A. *Bayesian inference of structural brain networks*. Neuroimage, 66:543-552, 2013.

[54] G.E. Hinton. *Training products of experts by minimizing contrastive divergence.* Neural Computation, 14:1771-1800, 2002.

[55] Jacob B. Hirsh, Raymond A. Mar and Jordan B. Peterson *Psychological Entropy: A Framework for Understanding Uncertainty-Related Anxiety*. Psychological Review, 119(2):304-320, 2012.

[56] Peter D. Hoff, Adrian E. Raftery, Mark S. Handcock. *Latent Space Approaches to Social Network Analysis.* Journal of the American Statistical Association, 97(460):1090-1098.

[57] Peter D. Hoff. *Multiplicative Latent Factor Models for Description and Prediction of Social Networks*. Computational and Mathematical Organization Theory, 15(4):261-272, 2009.

[58] Peter D. Hoff. *Simulation of the matrix Bingham-von Mises-Fisher distribution, with applications to multivariate and relational data*. Journal of Computational and Graphical Statistics, 18(2):438-456, 2009.

[59] David R. Hunter, Mark S. Handcock, Carter T. Butts, Steven M. Goodreau, and Martina Morris. *ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks*. Journal of Statistical Software, 24(3), 2008.

[60] Rob J. Hyndman *Computing and Graphing Highest Density Regions*. The American Statistician, 50(2):120-126, 1996.

[61] Julien Jacques and Christophe Biernacki. *Model-based Clustering for Multivariate Partial Ranking Data*. Journal of Statistical Planning and Inference, 149:201-217, 2014.

[62] Pavel N. Krivitsky. *Exponential-family Random Graph Models for Valued Networks*. Electronic Journal of Statistics, 6:1100-1128, 2012.

[63] Jon A. Krosnick and Duane F. Alwin. *An evaluation of a cognitive theory of response-order effects in survey measurement*. Public Opinion Quarterly, 51:201-219, 1987.

[64] Ravi Kumar and Sergei Vassilvitskii. *Generalized Distances between Rankings*. Proceedings of the 19th international conference on World wide web. ACM, 571-580, 2010.

[65] Barbara S. Lawrence. *Organizational Reference Groups: A Missing Perspective on Social Context*. Organization Science, 17(1):80–100, 2006.

[66] Barbara S. Lawrence and Michael J. Zyphur. *Identifying Organizational Faultlines With Latent Class Cluster Analysis*. Organizational Research Methods, 14(1):32–57, 2011.

[67] Scott W. Linderman, Ryan P. Adams. *Discovering Latent Network Structure in Point Process Data*. In International Conference on Machine Learning, 1413-1421, 2014.

[68] Neil Malhotra. *Completion Time and Response Order Effects in Web Surveys*. Public Opinion Quarterly, 72(5):914-934, 2008.

[69] Peter V. Marsden. *Network Data and Measurement*. Annual Review of Sociology, 16:435–463, 1990.

[70] Christopher McCarty, Peter D. Killworth, James Rennell. *Impact of Methods for Reducing Respondent Burden on Personal Network Structural Measures*. Social Networks, 29:300–315, 2007.

[71] M. D. McKay, W. J. Conover, and R. J. Beckman. *A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code*. Technometrics, 21:239–245, 1979.

[72] Karthika Mohan, Judea Pearl, and Jin Tian. *Graphical Models for Inference with Missing Data*. Advances in Neural Information Processing System, 26:1277-1285, 2013.

[73] David W. Moore. *Measuring New Types of Question-Order Effects: Additive and Subtractive*. The Public Opinion Quarterly, 66(1):80-91, 2002.

[74] Stephen L. Morgan and Christopher Winship. *Counterfactuals and Causual Inference*. Cambridge University Press, Cambridge, UK, 2007.

[75] Laurin A. J. Mueller, Matthias Dehmer and Frank Emmert-Streib. *Systems Biology: Integrative Biology and Simulation Tools*. Springer Netherlands, Dordrecht, 2013.

[76] Roger B. Nelsen. *An Introduction to Copulas*. Springer Science & Business Media, 2007.

[77] A. James O'Malley and Peter V. Marsden *The Analysis of Social Networks* Health Service Outcomes Res Methodology, 8(4):222-229.

[78] Art B. Owen. *Quasi-monte Carlo Sampling*. Monte Carlo Ray Tracing: Siggraph 1 (2003):69-88.

[79] Art B. Owen and Seth D. Tribble. *A Quasi-Monte Carlo Metropolis Algorithm*. Proceedings of the National Academy of Sciences of the United States of America, 102(25):8844-8849, 2005.

[80] Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. *Causal inference in statistics: a primer*. John Wiley & Sons, 2016.

[81] Kenneth A. Rasinski. *The Effect of Question Wording on Public Support for Government Spending*. Public Opinion Quarterly, 53:388-394, 1989.

[82] Donald B. Rubin. *Inference and missing data*. Biometrika, 63(3):581-592, 1976.

[83] Donald B. Rubin. *Multiple imputation for nonresponse in surveys*. John Wiley & Sons, 2004.

[84] Paul R. Rosenbaum. *Interference Between Units in Randomized Experiments*. Journal of the American Statistical Association, 102(477):191-200, 2007.

[85] Michael Schweinberger and Mark S. Handcock. *Local Dependence in Random Graph Models: Characterization, Properties and Statistical Inference*. Journal of the Royal Statistical Society B, 77(3):647-676, 2015.

[86] D. Sculley. *Rank aggregation for similar items*. Proceedings of the 2007 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, 587-592, 2007.

[87] Betsy Sinclair, Margaret McConnell and Donald P. Green. *Detecting spillover effects: Design and analysis of multilevel experiments*. American Journal of Political Science, 56(4):1055-1069, 2012.

[88] Michael Stein. *Large Sample Properties of Simulations Using Latin Hypercube Sampling*. Technometrics, 29(2):143-151, 1987.

[89] Daniel Taylor-Rodriguez, Andrew Womack and Nikolay Bliznyuk. *Bayesian Variable Selection on Model Spaces Constrained by Heredity Conditions*. Journal of Computational and Graphical Statistics, 25(2):515-535, 2016.

[90] Steven K. Thompson. *Sampling*. John Wiley & Sons, Inc., Hoboken, NJ, 2012.

[91] Robert Tibshirani. *Regression Shrinkage and Selection via the Lasso*. Journal of the Royal Statistical Society. Series B (Methodological), 58(1):267-288, 1996.

[92] Vasja Vehovar, Katja Lozar Manfreda, Gasper Koren, and Valentina Hlebec. *Measuring Ego-Centered Social Networks on the Web: Questionnaire Design Issues*. Social Networks, 30:213–222, 2008.

[93] Alan L. Yuille. *The convergence of contrastive divergences*. Advances in neural information processing systems, 1593-1600, 2005.

[94] John Zaller and Stanley Feldman. *A Simple Theory of the Survey Response: Answering Questions versus Revealing Preferences*. American Journal of Political Science, 579-616, 1992.