University of Texas at El Paso

ScholarWorks@UTEP

Open Access Theses & Dissertations

2022-12-01

Evaluation Of Effect Of Preprocessing Algorithms On Resting State Fmri Data

Hortencia Josefina Hernandez
University of Texas at El Paso

Follow this and additional works at: https://scholarworks.utep.edu/open_etd

Part of the Statistics and Probability Commons

Recommended Citation

Hernandez, Hortencia Josefina, "Evaluation Of Effect Of Preprocessing Algorithms On Resting State Fmri Data" (2022). *Open Access Theses & Dissertations*. 3685.
https://scholarworks.utep.edu/open_etd/3685

This is brought to you for free and open access by ScholarWorks@UTEP. It has been accepted for inclusion in Open Access Theses & Dissertations by an authorized administrator of ScholarWorks@UTEP. For more information, please contact lweber@utep.edu.

EVALUATION OF EFFECT OF PREPROCESSING ALGORITHMS ON RESTING STATE FMRI DATA

HORTENCIA JOSEFINA HERNANDEZ

Master's Program in Statistics

APPROVED:

Amy Wagler, Chair, Ph.D.

Richard Ortiz, Ph.D.

Art Duval, Ph.D.

Stephen Crites, Ph.D.
Dean of the Graduate School

To my younger siblings,
'¡Sí Se Puede!'

EVALUATION OF EFFECT OF PREPROCESSING ALGORITHMS ON RESTING STATE FMRI DATA

by

HORTENCIA JOSEFINA HERNANDEZ, B.S. in Mathematics

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Master's Program in Statistics

THE UNIVERSITY OF TEXAS AT EL PASO

December 2022

Acknowledgements

I would like to express my deepest and sincerest gratitude to Dr. Amy Wagler. Throughout this past semester and summer she has encouraged, motivated, and inspired me with her immense knowledge. She is truly an inspiration to me.

To my committee members, Dr. Art Duval and Dr. Richard Ortiz, for their insight and guidance. Learning from Dr. Duval and being able to apply his teachings. Thank you to Dr. Richard Ortiz, for being the inspiration to this thesis. This is a very interesting topic and I hope to continue working together in the future.

While studying, I am truly grateful for all the connections I have made here and having a family away from home.

Lastly, I am most grateful for my parents who pushed me throughout school, and to my four younger siblings who push me everyday to be someone they can look up to. Without them I would not be where I am today.

Thank you.

Abstract

Graph theory modeling is a common modeling approach in neurobiology research studies. These models are useful since they describe patterns of connection for regions of interest in the brain using resting state fMRI images. The standard rule of thumb is to threshold the observed activation levels prior to model building. It is reasonable to assume that the use of this threshold affects the statistical distribution of commonly reported centrality metrics from the graph theory model, such as degree, betweenness, and closeness. In this study we examine the differential effect of using the standard approaches versus alternative direct thresholds and incorporation of thresholds through soft and hard covariance estimation. Along with the way it is viewed we care about the way we preprocess the data. Results indicate that direct thresholding is a more reliable preprocessing strategy, but soft thresholding of the covariance matrix may be a promising alternative in particular settings.

Table of Contents

Page

Acknowledgements iv

Abstract v

Table of Contents vi

List of Tables viii

List of Figures ix

Chapter

1 Introduction 1

1.1 Graph Theory Modeling 1

1.1.1 Degree 2

1.1.2 Betweenness 2

1.1.3 Closeness 3

1.2 Applications of Graph Theory 3

1.3 Preprocessing 4

1.3.1 Thresholding 5

2 Methodology 6

2.1 Data Strategies 6

2.2 Soft versus Hard Thresholds for Covarince Matrix Estimation 7

2.3 Comparing KI vs IL 8

2.4 Random Graphs 8

2.5 Compare Simulations 9

3 Results 10

3.1 KI vs IL 10

3.1.1 Case 0: 10

3.1.2 Case 1: 12

3.1.3 Case 2: 13
3.1.4 Case 3.1: 15
3.1.5 Case 3.2: 16
3.1.6 Case 4.1: 17
3.1.7 Case 4.2: 19
3.2 Random Simulations 20
3.2.1 n = 10 20
3.2.2 n = 50 21
3.2.3 n = 100 21
4 Analysis 23
4.1 KI vs IL 23
4.2 Random Simulations 23
5 Concluding Remarks 25
References 33

Appendix

A Graphs 34
Curriculum Vitae 40

List of Tables

3.1 Case 0 (* < 0.05 and ** < 0.01) 10

3.2 Case 1 (* < 0.05 and ** < 0.01) 12

3.3 Case 2(* < 0.05 and ** < 0.01) 13

3.4 Case 3.1(* < 0.05 and ** < 0.01) 15

3.5 Case 3.2 (* < 0.05 and ** < 0.01) 17

3.6 Case 4.1 (* < 0.05 and ** < 0.01) 18

3.7 Case 4.2 (* < 0.05 and ** < 0.01) 19

5.1 Statistical Difference in KI vs IL 26

5.2 Statistical Difference for Random Simulations (where $p_1 = p_2 = 0.1$) 27

5.3 Statistical Difference for Random Simulations (where $p_1 = p_2 = 0.5$) 28

5.4 Statistical Difference for Random Simulations (where $p_1 = p_2 = 0.9$) 29

5.5 Statistical Difference for Random Simulations (where $p_1 = 0.1$ vs. $p_2 = 0.5$) 30

5.6 Statistical Difference for Random Simulations (where $p_1 = 0.1$ vs. $p_2 = 0.9$) 31

5.7 Statistical Difference for Random Simulations (where $p_1 = 0.5$ vs. $p_2 = 0.9$) 32

List of Figures

3.1 KI vs IL: Case 0 11

3.2 KI vs IL: Case 1 13

3.3 KI vs IL: Case 2 14

3.4 KI vs IL: Case 3.1 16

3.5 KI vs IL: Case 3.2 17

3.6 KI vs IL: Case 4.1 18

3.7 KI vs IL: Case 4.2 20

A.1 n = 10: Differences in Betweenness for $p_1 = p_2 = 0.5$ 34

A.2 n = 10: Differences in Degree for $p_1 = p_2 = 0.9$ 35

A.3 n = 10: Differences in Degree for $p_1 = 0.5$ vs $p_2 = 0.9$ 35

A.4 n = 50: Differences in Closeness for $p_1 = p_2 = 0.1$ 36

A.5 n = 50: Differences in Closeness for $p_1 = p_2 = 0.9$ 36

A.6 n = 50: Differences in Degree for $p_1 = 0.1$ vs $p_2 = 0.9$ 37

A.7 n = 50: Differences in Closeness for $p_1 = 0.5$ vs $p_2 = 0.9$ 37

A.8 n = 100: Differences in Betweenness for $p_1 = p_2 = 0.9$ 38

A.9 n = 100: Differences in Degree for $p_1 = 0.1$ vs $p_2 = 0.5$ 38

A.10 n = 100: Differences in Degree for $p_1 = 0.5$ vs $p_2 = 0.9$ 39

Chapter 1

Introduction

There are many ways to analyze and view data. Visually, graphs help with seeing how the data is related The use of graph theory modeling allows the user to see the data as a network of nodes and edges. The graph theory models reveal how a network's nodes are connected and how they directly and indirectly affect one another. Not only that, but when given data, researchers tend to 'clean' up the data. As in, we want to focus on what we want to find as well as if there is any data missing or inputted incorrectly. Thus, there are different preprocessing strategies that researchers use to find the variables they care about, or that are deemed important.

Through this paper we will be testing several different preprocessing strategies and how they affect the centralities of graphical theory models. We ask the question on if different preprocessing strategies will effect the analysis and reporting of centrality measures of graphical theory models.

1.1 Graph Theory Modeling

Graphical Theory Modeling uses networks to view how the data is connected. The data is viewed as a network of nodes and edges, where the data point is a node and an edge is the connection between nodes.

While there are many ways to utilize graph theory modeling to understand relationships in data, we focus on the centrality graph metrics in this study. The centrality metric measures the importance of nodes within the network, and there are many ways to find what is considered to be central [3]. Here, we are focused particularly on degree, betweenness, and closeness since these are measures commonly reported when analyzing brain imaging data. These centrality metrics are commonly used to analyze nodes connection and connective attributes within the network. Although all three measures help determine centrality and node importance, they actually measure distinct attributes of these centrality and importance relations. In fact, assessing multiple centralities simultaneously tells a complete story about node connection and importance. For instance, a node could have a low degree and have high betweenness, as in the node may not have many edges connected to it, but if taken out it will affect the rest of the network. Similarly, a node with low betweenness and high degree would have many connections, but not place a role in other node connections. In the following, these three network metrics are described qualitatively and via mathematical formula.

1.1.1 Degree

The degree of a network is the measure of how many connections a node has or how many edges connect to a node. Usually the node with the highest degree means that it has the most connections or edges to it. Although the simplest centrality, it only determines the quantity of edges that are connected to a node. This can be defined as

$$D = \sum_{j=1}^{n} A_{ij}$$

where n is the number of rows in the adjacency matrix $\mathbf{A}$ and A_{ij} is the number of edges between nodes i and j. [4]

1.1.2 Betweenness

The betweenness centrality measures the path between nodes. Those nodes with high measure for betweenness usually mean that those nodes are of high importance within the network. The betweenness can be calculated as follows:

$$x_i = \sum_{st} \frac{n^i_{st}}{g_{st}}$$

where n_{st}^{i} is the number of shortest paths from s to t through i, and g_{st} is the total number of shortest paths. [3]

1.1.3 Closeness

The closeness of a network is determined by “measuring the mean distance from a node to other nodes” [p.170 Networks] This can be expressed through the harmonic mean distance by the following equation:

$$C_i' = \frac{n}{\sum_{j(\neq i)} d_{ij}}$$

where d_{ij} is the shortest distance from node i to node j [3]

1.2 Applications of Graph Theory

Graph theory can be used in many different settings. In general, graph theory modeling is useful in order to show “patterns of interactions between parts of a system.” [3]. For example, one can view social networks, seeing how many connections a person has in different settings. or one can also see biological networks, for instance a neural network. Moreover, networks build to model brain images and connections between brain regions of interest (ROI) is an important application of graph theory modeling. An example of this is in the article “Functional Connectivity Differences between Two Culturally Distinct Prairie Vole Populations: Insights into the Prosocial Network.” [4]

Ortiz [4] explored the idea of using graph theory to better understand functional connectivity of two different behaviorally distinct populations of prairie voles (Microtus Ochrogaster). Prairie voles are an excellent pre-clinical laboratory animal model used to better understand human-relevant social behaviors, as they display a suite of prosocial behaviors that are similar to humans. He analyzed the connectivity in three “cores” of neural regions, the prosocial core, olfactory core, and a control ‘core’, between the Kansas-Illinois (KI) and Illinois (IL) male brain. He also included a social bonding core and integrated it with the prosocial core. Focusing on three graph metrics – degree, betweenness, and closeness centrality— helped analyze which nodes are significantly connected between the cores and highlighted differences between the KI and IL cores.

1.3 Preprocessing

In recent studies, it has been found that resting state functional MRI (rs-fMRI) is a powerful tool used in neuro-imaging to evaluate functional connectivity patterns. [1] In the article, "Evaluating the reliability of different preprocessing steps to estimate graph theoretical measures in resting state fMRI data" they use different preprocessing steps rather than post-processing methods to help calculate the connectivity of the brain. They test the preprocessing with a reliability and reproducibility of commonly reported graph theory metrics.

Preprocessing is the process by which researchers clean the data. This stage in analyzing data allows the user to fix mistakes that may have happened and be able to easily find trends. Preprocessing also is useful for de-noising the data to allow for important signals in the graph theoretic model to emerge. In Ortiz [4] a threshold of 2.3 was used to de-noise his data. In other words, if a connection had a measure of anything less than 2.3 he made those zeros. The threshold follows from a gaussian random field theory where the 2.3 equates to a connectivity significantly greater or less than 0.[7]

However, this threshold was used as a standard rule-of-thumb from other influential studies [7]. This may not be the optimum threshold to use for de-noising data. Moreover, a data-dependent threshold may be ideal and useful for just about any data setting when preparing for graph theory modeling.

In the article, "Evaluating the reliability of different preprocessing steps to estimate graph theoretical measures in resting state fMRI data" they used seven different preprocessing strategies to compare the data with two different test. Although they used seven different preprocessing strategies, they were mixed with 5 different conditions. Bandpass filtering filtered data between 0.01 and 0.1 Hz, White Matter (WM), Cerebrospinal Fluid (CSF), and Global Signal Regression is a multiple regression step where the extracted CSF, WM, or Global signals were nuisance variables.

A. None
B. Bandpass Filtering
C. CSF and WM Regression
D. Bandpass and CSF and WM Regression
E. Bandpass, CSF and WM Regression, and Scrubbing with Motion
F. Bandpass, CSF and WM Regression, and Scrubbing with Outliers
G. Bandpass, CSF and WM Regression, and Global Signal Regression

Here, they did three different tests to check the reliability and reproducibility. One of the tests was on motion versus the graphical theory measures correlation. [1]. They found that “adding or removing different preprocessing schemes greatly affect the final results”[1]. Using four graphical theory measurements they tested the reliability and the reproducibility of each strategy. This end results they found were that strategy “F” should be used as it increased the reliability of using it.

Using a similar idea, we will use seven different preprocessing strategies to help analyze the data and the results.

1.3.1 Thresholding

The portion of preprocessing we are focused on is the type of threshold we are using on the data. Threshold, in simplicity is the cut off point. If values are less than a certain number they will yield to zero. Will changing the threshold affect the conclusion of the data? The thresholds used here are the 2.3 threshold, hard and soft threshold based on the data set, and folded normal distribution. Each uses a process to determine the threshold and apply it to the data.

Chapter 2

Methodology

This study is designed to examine the effect of different preprocessing strategies on the analysis and on the reporting of graphical theory models. To achieve this goal, we will follow these steps:

- Test data from Ortiz's paper with 7 preprocessing strategies (thresholds)
- Compare the KI and IL results (from Ortiz's paper)
- Simulate random graphs with n = 10, 50, and 100 with p = 0.1, 0.5, and 0.9
- Apply the cases to the simulated random graphs
- Compare same to same
  - Generate two n = 10, with p = 0.1 for both
  - Similar for each n for each p
- Compare the different graphs
  - Compare each 0.1, 0.5, and 0.9 for each n.

2.1 Data Strategies

To start we are going to look at the data presented in Ortiz's [4] paper of the KI and IL voles brains. As stated he studied the connections within the brain ROIs and used graph modeling to view and analyse the results. Using his data we are going to change the threshold to see if looking at different thresholds will affect the graphical model, as in that the thresholds will lead to different results.

There are seven cases that we begin to look at:

0. Raw Data using Standard Covariance
1. 2.3 Threshold (Ortiz's Threshold)
2. Assumes folded normal distribution since activation levels are bounded by 0, rather than normal.
3. Raw Data with Soft Thresholding

   A threshold based on the size of the data estimated by [6]

   $$a = 2 - (\frac{2 + \log(1)}{\log(n)})^{\frac{1}{2}}$$

   where n is the size of the data.

   3.1 Uses 10-fold validation for optimizing parameter estimates

   3.2 Sequencing from .01 to the threshold a
4. Raw Data with Hard Thresholding

   A threshold based on the size of the data found in Case 3

   4.1 Uses 10-fold validation for optimizing parameter estimates

   4.2 Sequencing from .01 to the threshold a

2.2 Soft versus Hard Thresholds for Covarince Matrix Estimation

To define soft and hard thesholds let us define a generalized thresholding operator. For any $\lambda \geq 0$, define a generalized thresholding operator to be a function $s_\lambda : R \to R$ such that for all $z \in R$ the following are satisfied:

i. $|s_\lambda(z)| \leq |z|$

ii. $s_\lambda(z) = 0$ for $|z| \leq \lambda$

iii. $|s_\lambda(z) - z| \leq \lambda$

Hard thresholding is appropriate for sparse covariance and applies to off-diagonal elements of the sample covariance matrix. The entry of sample covariance matrix $S_{i,j} = 0$ if $|S_{i,j}| \leq \lambda$ where λ is a thresholding value. This means that for hard thresholding everything is set to zero except the largest values. [2]

The soft thresholding method for covariance estimation takes off-diagonal elements z of sample covariance matrix and applies

$$h(z) = \mathrm{sgn}(z)(|z| - \lambda)_+$$

where $\mathrm{sgn}(z)$ is a sign of the value z, and

$$(x)_+ = \max(x, 0).$$

This essentially applies a lasso penalty to provide maximum shrinkage. It essentially shrinks all the values towards zero, not just a specific type of value. [2]

2.3 Comparing KI vs IL

Of each threshold, we compared the IL vs the KI data, by looking at the differences on a histogram, as well as comparing their mean, median, and standard deviation of each of the centrality measures.

2.4 Random Graphs

After comparing the KI vs IL data, we created random graphs with 10, 50, and 100 nodes with a probability of 0.1, 0.5, and 0.9 of having an edge. We created 500 simulations of the data to compare.

2.5 Compare Simulations

After creating the 500 simulations, we pulled the maxs of each simulation of the degree, closeness, and betweenness. We then compared each p with itself and between the other p's.

As in, for each n we compared it with the same probablity and between the different probabilities.

Chapter 3

Results

3.1 KI vs IL

For each case we output the table of the mean, median and standard deviation (StdDev) of each centrality metric. We also include the graphs of the differences of each metric.

3.1.1 Case 0:

Table 3.1: Case 0 (* < 0.05 and ** < 0.01)

	Mean Degree**	Mean Between	Mean Close**
KI	31.75	51.89	0.00
IL	27.95	53.05	0.00
	Median Degree	Median Between	Median Close
KI	31.77	26	0.00
IL	28.28	30	0.00
	StdDev Degree	StdDev Between	StdDev Close
KI	7.47	69.18	0.00
IL	6.80	64.22	0.00

For the KI the max (or node with the highest measure) for degree and closeness is the Paraflocculus Cerebellum with a degree of 46.9602 and closeness of 0.00346054. The highest betweenness is found at the Reticular Nucleus with a measure of 394.

For the IL the max for degree, betweenness, and closeness is the Caudate Putamen

Striatum with a degree of 42.02879, betweenness of 322 and closeness of 0.003194883.

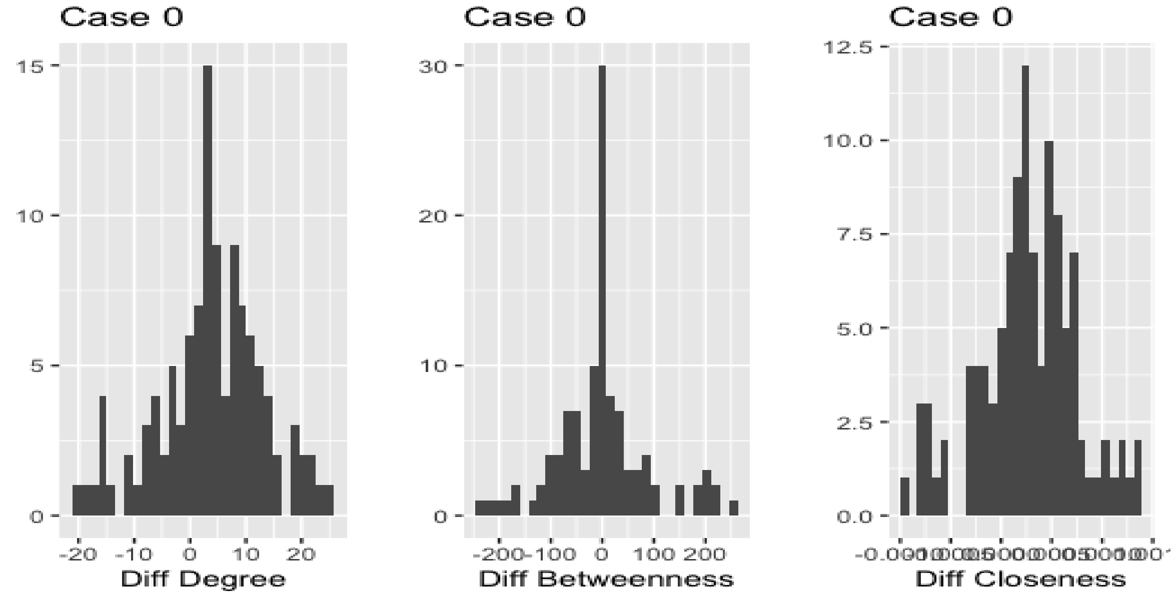

Figure 3.1: KI vs IL: Case 0

From the graph we can see that the differences in degree and betweenness are relatively normally distributed where closeness looks to be slightly right skewed.

3.1.2 Case 1:

Table 3.2: Case 1 (* < 0.05 and ** < 0.01)

	MeanDeg**	MeanBetween*	MeanClose**
KI	9.66	145.12	0.00
IL	12.09	106.58	0.00

	MedDeg	MedBetween	MedClose
KI	9.79	108	0.00
IL	12.00	72	0.00

	StDevDeg	StDevBetween	StDevClose
KI	1.63	131.48	0.00
IL	2.18	113.20	0.00

For the KI the max for betweenness and closeness is the Superior Colliculus with a betweenness of 818 and closeness of 0.001272408. The max degree is found at the Pontine Reticular Nucleus Oral with a measure of 14.70715.

For the IL the max for degree, betweenness, and closeness is the Caudate Putamen Striatum with a degree of 42.02879, betweenness of 322 and closeness of 0.003194883.

From the graphs we can see that the differences of betweenness look slightly right skewed and the differences of closeness and degree look slightly left skewed.

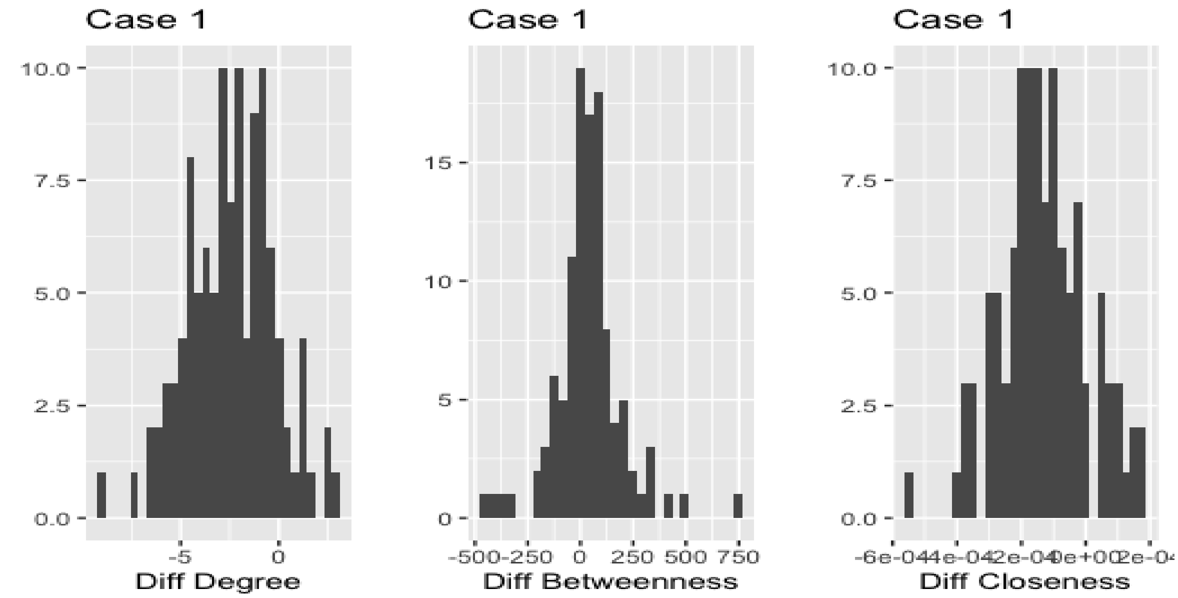

Figure 3.2: KI vs IL: Case 1

3.1.3 Case 2:

Table 3.3: Case 2(* < 0.05 and ** < 0.01)

	MeanDeg**	MeanBetween*	MeanClose**
KI	7.51	186.40	0.00
IL	10.51	143.55	0.00

	MedDeg	MedBetween	MedClose
KI	7.44	140	0.00
IL	10.63	112	0.00

	StdDevDeg	StdDevBetween	StdDevClose
KI	2.03	191.53	0.00
IL	1.92	133.75	0.00

For the KI the max for betweenness and closeness is the Anterior Cingulate Ctx with a betweenness of 1088 and closeness of 0.001316408. The max degree is found at the Pontine

Reticular Nucleus Oral with a measure of 13.56319.

For the IL the max for degree and closeness is the Reticular Formation with a degree of 14.79103 and closeness of 0.001360383. The max betweenness is found at the Pontine REticular Nucleus Oral with a measure of 848.

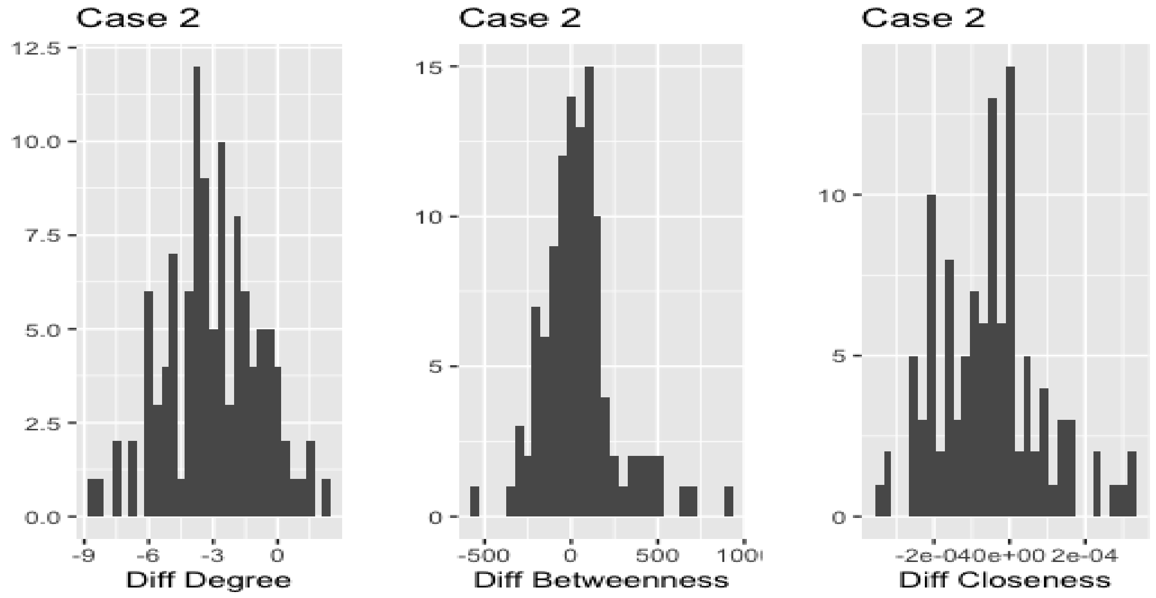

Figure 3.3: KI vs IL: Case 2

From the graphs it is easy to see that the difference of betweenness is right skewed.

3.1.4 Case 3.1:

Table 3.4: Case 3.1($* < 0.05$ and $** < 0.01$)

	MeanDeg	MeanBetween	MeanClose
KI	6.71	185.64	0.00
IL	5.73	160.77	0.00

	MedDeg	MedBetween	MedClose
KI	4.61	0	0
IL	3.63	0	0

	StDevDeg	StDevBetween	StDevClose
KI	6.67	626.13	0.00
IL	5.78	465.31	0.00

For the KI the max for betweenness is found at Reticular Nuclues with a measure of 4506. The max for degree is found at Crus Ansiform Lobule with a measure of 29.28603

For the IL the max for betweenness is found at Reticular Nuclues of 3034 and degree is found at Pontine Reticular Nucleus Caudal with a measure of 24.95146 For both populations, the closeness is 0 for all nodes.

From the table and the and the graph, when applying the soft thresholding with a threshold based on the data, all closeness measures become zero. This means that the average distance between each node is 0. The difference of betweenness and degree look mostly normally distributed.

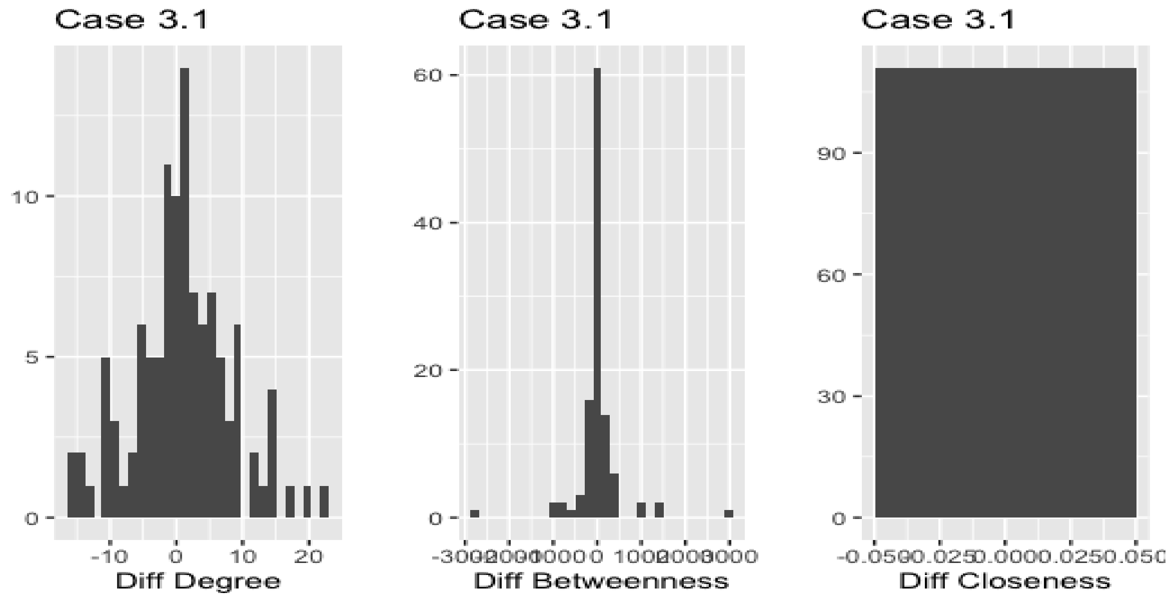

Figure 3.4: KI vs IL: Case 3.1

3.1.5 Case 3.2:

For the KI the max for degree and closeness is the Crus Ansiform Lobule with a degree of 81.1653 and closeness of 0.00572378. The max betweenness is found at the Reticular Nucleus with a measure of 1402.

For the IL the max for degree and closeness is the Pontine Reticular Nucleus Caudal with a degree of 76.97324 and closeness of 0.005686382. The max betweenness is found at the Reticular Nucleus with a measure of 1154.

Difference of degree looks to be slightly left skewed and difference of betweenness looks to be mostly normal.

Table 3.5: Case 3.2 (* < 0.05 and ** < 0.01)

	MeanDeg	MeanBetween	MeanClose
KI	45.12	65.17	0.00
IL	43.98	66.23	0.00

	MedDeg	MedBetween	MedClose
KI	45.57	4	0.00
IL	44.68	8	0.00

	StDevDeg	StDevBetween	StDevClose
KI	14.70	191.14	0.00
IL	14.06	173.50	0.00

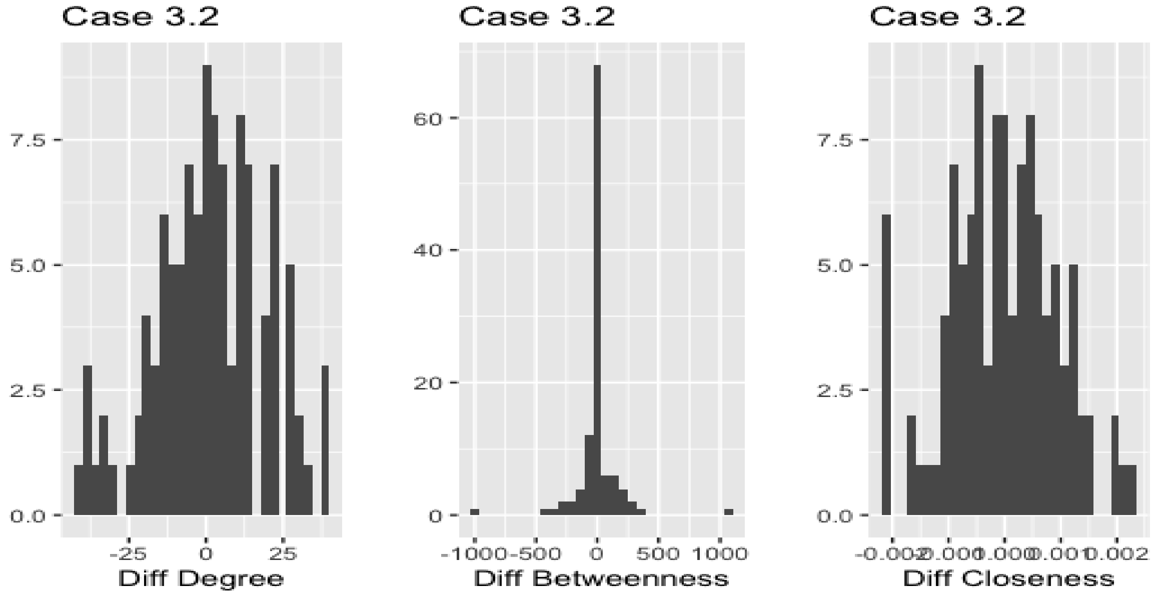

Figure 3.5: KI vs IL: Case 3.2

3.1.6 Case 4.1:

For the KI the max for betweenness is 1842 and degree is 68.53373

For the IL the max for betweenness is 1610 and degree is 66.54229 For both populations, the closeness is 0 for all nodes.

Table 3.6: Case 4.1 (* < 0.05 and ** < 0.01)

	MeanDeg	MeanBetween	MeanClose
KI	24.82	92.18	0.00
IL	22.58	89.68	0.00

	MedDeg	MedBetween	MedClose
KI	22.67	4	0
IL	19.72	4	0

	StDevDeg	StDevBetween	StDevClose
KI	17.70	254.57	0.00
IL	16.71	238.79	0.00

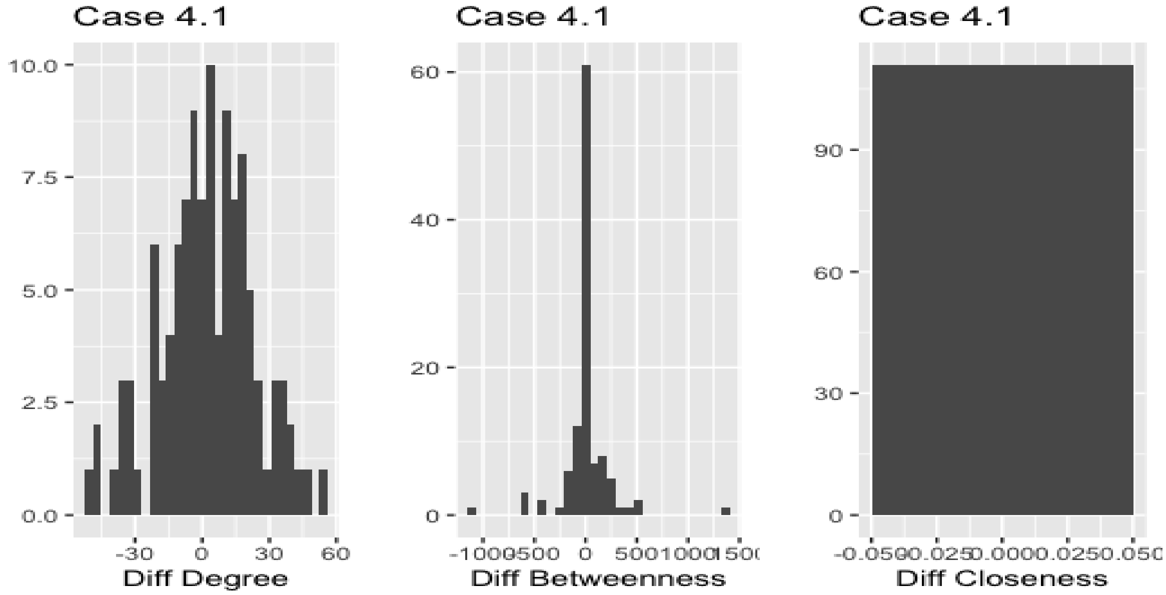

Figure 3.6: KI vs IL: Case 4.1

Similar to case 3.1 (soft thresholding using a threshold based on the data size) the difference of degree and betweenness look normal where the difference of closeness is zero. Every node has a measure of 0 for its closeness.

3.1.7 Case 4.2:

Table 3.7: Case 4.2 (* < 0.05 and ** < 0.01)

	MeanDeg	MeanBetween	MeanClose
KI	46.20	63.77	0.00
IL	45.06	64.92	0.00

	MedDeg	MedBetween	MedClose
KI	46.67	4	0.00
IL	45.77	8	0.00

	StDevDeg	StDevBetween	StDevClose
KI	14.71	185.82	0.00
IL	14.07	169.41	0.00

For the KI the max for degree and closeness is the Crus Ansiform Lobule with a degree of 82.2553 and closeness of 0.005807079. The max betweenness is found at the Reticular Nucleus with a measure of 1344.

For the IL the max for degree and closeness is the Pontine Reticular Nucleus Caudal with a degree of 78.05324 and closeness of 0.005763311. The max betweenness is found at the Reticular Nucleus with a measure of 1118.

Similar to case 3.2, the difference of betweenness looks to be normally distributed and the difference of degree looks to be slightly left skewed.

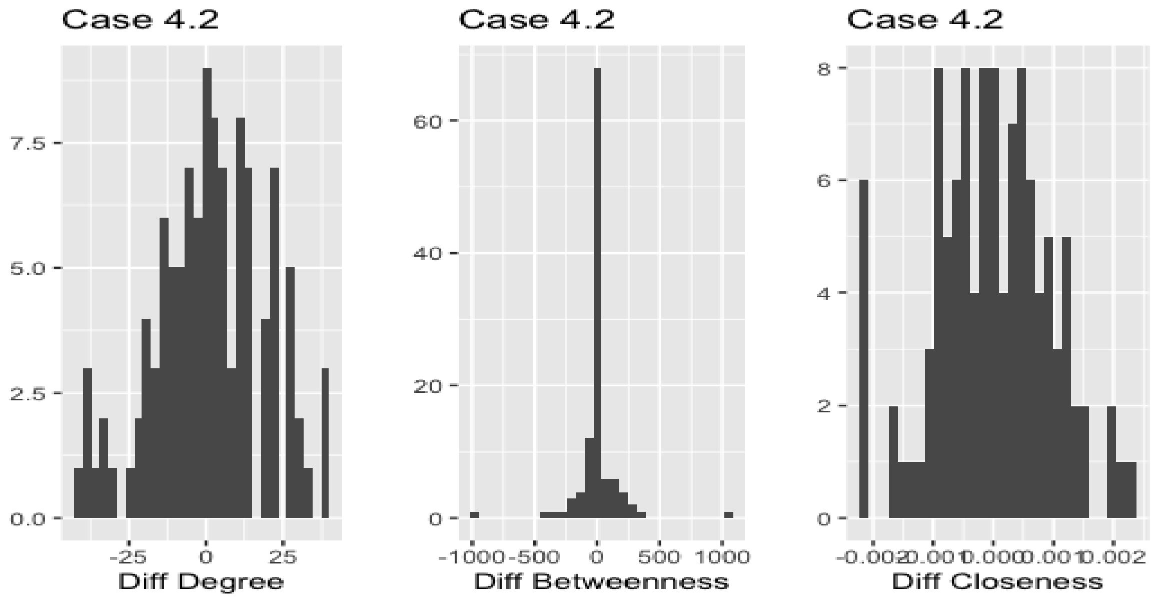

Figure 3.7: KI vs IL: Case 4.2

3.2 Random Simulations

Following the strategy using Ortiz's data, we randomly simulate our own data using the erdos.renyi.game function in the iGraph package. This function creates random graphs with n nodes and p probability of the node having an edge. Testing for n = 10, 50, and 100 each with probability p = 0.1, 0.5, and 0.9, we ran this simulation 500 times comparing for each n with the same p and with different p's.

3.2.1 n = 10

For comparing with the same p, the average difference of closeness is 0. For all cases except case 0, the closeness is 0 for each node. For case 0, the difference of closeness is distributed normally. For difference in degree, for each case they are normally distributed where the mean looks to be 0. The higher p is, the more case 0 looks normally distributed and for each case of the difference in betweenness the mean is 0.

For different p's, the mean of each centrality is 0, whereas for the null case, most of the differences in the centrality measures are normally distributed.

3.2.2 n = 50

For comparing with the same p, cases 0 to 2 the difference of degrees and betweenness are normally distributed. Interesting to note that for case 3.2 the difference of degree has a mean of 0 whereas 3.1 and 4 the degree is 0 for all. As for the difference of closeness, all cases except case 0 has a measure of 0. Case 0 is normally distributed which we can easily see that as p gets higher, the more noticable it is.

Comparing $p = 0.1$ to $p = 0.5$ and $p = 0.9$ for difference in degree cases 0, 1 and 2 all look left skewed where case 3.1 and 4 has a measure 0 and 3.2 has an average of approximately 0. For the difference of closeness the measures are 0 for all cases except case 0 which is normally distributed. The difference of betweenness for case 1 and 2 look to be left skewed where cases 3 and 4 are 0. For the null case it is normally distributed comparing $p = 0.5$ to $p = 0.9$ and right skewed for comparing $p = 0.5$ and $p = 0.9$ to $p = 0.1$.

3.2.3 n = 100

Comparing the same p, similar to n = 50, the difference of degree the higher p is the more it is noticable that it is distributed normally for cases 0, 1, and 2. Whereas for case 3.1 and 4.1 is 0 and for 3.2 and 4.2 the mean is approximately 0. For difference in closeness all of the cases have a measure of 0 except the null case which is normally distributed. As for difference in betweenness the higher the p is, the more the distribution is normal, where case 3 and 4 are 0.

Comparing $p = 0.1$ to $p = 0.5$ and $p = 0.9$ for difference in degree cases 0, 1 and 2 all look left skewed where cases 3 and 4 are either 0 or have an average of 0. The difference of closeness are all 0 except case 0. Comparing Comparing $p = 0.5$ to $p = 0.9$ it is normally distributed where as for $p = 0.1$ versus $p = 0.5$ or $p = 0.9$ is left skewed. For difference of betweenness, $p = 0.5$ versus $p = 0.9$ case 0 is normally distributed and for $p = 0.1$ versus

$p = 0.5$ or $p = 0.9$ it is right skewed. Case 1 and 2 are similarly left skewed and case 3 and 4 have a measure of 0.

Chapter 4

Analysis

Comparing the results between each case.

4.1 KI vs IL

Between the cases 3 and 4, both yielded similar results, and comparing to the null case (based solely on the original data using standard covariance) the mean degree gets smaller between cases 0 and 2 the mean betweenness gets larger. It is also interesting to note that comparing case 0 with 3.2 and 4.2 the measures are larger than the null case. Case 3.1 and 4.1 have very similar distributions and 3.2 and 4.2 also have very similar distributions. In fact, for all of case 3 and 4 they have the same node that has the max centrality (excluding the closeness centrality since for 3.1 and 4.1 those come out to zero).

Since case 3 and 4 yield similar results, such that the areas where the maxes are found are also in the same region. I would argue that case 3.2 and 4.2 can be interchangable as the results are very similar and yield to a similar conclusion between the two.

Similarly case 1 and 2 yield similar results, not precisely the same. they have slight differences. For instance the degree measures for both the KI and IL in case 1 and 2 are the same, with the measure being slightly different.

4.2 Random Simulations

For the random simulations Cases 3.1 and 4.1 (10 fold of the adaptive threshold) are both similar and yield similar results which is why 4.1 was not added to the output.

As more nodes came through, closeness for the null case was normally distributed whereas for the rest of the cases, the closeness measures all became 0.

As for case 4 as a whole, many of the differences became 0 especially for betweenness and closeness which aligns with saying that the distances between each of the nodes is the same for each of the probabilities no matter the size of the network.

Actually it is very interesting to note that for almost all n's the null case is normally distributed for each centrality metric. The some of the null cases that it is not normally distributed is found in the following:

- All difference of metrics in n = 10 are right skewed
- Difference of Betweenness
  - n = 50
    - $p_1 = 0.1$ vs $p_2 = 0.9$ (Right Skewed)
  - n = 100
    - $p_1 = 0.1$ vs $p_2 = 0.5$ (Right Skewed)
    - $p_1 = 0.1$ vs $p_2 = 0.9$ (Right Skewed)

The more nodes we can compare the more likely it is to see the distribution and the more likely the distribution is going to be normal. For cases 3 and 4, the measures are either 0 or the average difference is approximately zero which is normal in the case of 3.2.

Chapter 5

Concluding Remarks

There were major differences in the thresholding approaches and effect on the network statistical estimates. In general, cases 3 and 4 tended to pull all betweenness and closeness estimates to 0. As case 3 and 4 are based on soft and hard thresholding of the covariance estimate based on the observed activiation levels, this may indicate that a direct threshold is more appropriate. However in the KI vs IL data, it could be argued that either thresholding approach (direct or via covariance matrix thresholding) operates similarly as they uniformly lead to similar results and generally yield the same conclusion. Generally, cases 1 and 2 yield similar results for these reported metrics. These simulations provide support for using either the 2.3 or multiplicity (node size) corrected direct thresholds on activation levels. The variation among some of the differences in each case (excluding case 0) is 0 thus resulting in NA. Future research could focus on improving the soft and hard thresholding for the covariance estimation approaches and incorporate a more realistic graph theory simulation model for comparison.

Table 5.1: Statistical Difference in KI vs IL

*: p < 0.05, and **: p < 0.01

	Degree	Betweenness	Closeness
Case 0	Difference**	No Difference	Difference**
Case 1	Difference**	Difference*	Difference**
Case 2	Difference**	Difference*	Difference**
Case 3.1	No Difference	No Difference	No Difference
Case 3.2	No Difference	No Difference	No Difference
Case 4.1	No Difference	No Difference	No Difference
Case 4.2	No Difference	No Difference	No Difference

Table 5.2: Statistical Difference for Random Simulations (where $p_1 = p_2 = 0.1$)

*: p < 0.05, and **: p < 0.01

		Degree	Betweenness	Closeness
Case 0	n = 10:	Difference *	No Difference	Difference **
	n = 50:	No Difference	No Difference	Difference **
	n = 100:	Difference**	No Differencec	Difference**
Case 1	n = 10:	No Diffference	No Difference	NA
	n = 50:	No Difference	No Difference	NA
	n = 100:	Difference**	Difference*	NA
Case 2	n = 10:	No Difference	NA	NA
	n = 50:	Difference *	No Difference	NA
	n = 100:	No Difference	No Difference	NA
Case 3.1	n = 10:	No Difference	NA	NA
	n = 50:	NA	NA	NA
	n = 100:	NA	NA	NA
Case 3.2	n = 10:	No Difference	No Difference	NA
	n = 50:	No Difference	NA	NA
	n = 100:	No Difference	NA	NA
Case 4.1	n = 10:	Difference**	NA	NA
	n = 50:	NA	NA	NA
	n = 100:	NA	NA	NA
Case 4.2	n = 10:	No Difference	No Difference	NA
	n = 50:	NA	NA	NA
	n = 100:	NA	NA	NA

Table 5.3: Statistical Difference for Random Simulations (where $p_1 = p_2 = 0.5$)

*: p < 0.05, and **: p < 0.01

		Degree	Betweenness	Closeness
Case 0	n = 10:	No Difference	No Difference	Difference **
	n = 50:	No Difference	No Difference	Difference*
	n = 100:	No Difference	No Difference	No Difference
Case 1	n = 10:	Difference*	No Difference	NA
	n = 50:	No Difference	No Difference	NA
	n = 100:	No Difference	No Difference	NA
Case 2	n = 10:	No Difference	No Difference	NA
	n = 50:	Difference *	No Difference	NA
	n = 100:	No Difference	No Difference	NA
Case 3.1	n = 10:	No Difference	NA	NA
	n = 50:	NA	NA	NA
	n = 100:	NA	NA	NA
Case 3.2	n = 10:	No Difference	No Difference	NA
	n = 50:	No Difference	No Difference	NA
	n = 100:	No Difference	Difference**	NA
Case 4.1	n = 10:	No Difference	NA	NA
	n = 50:	NA	NA	NA
	n = 100:	NA	NA	NA
Case 4.2	n = 10:	No Difference	NA	NA
	n = 50:	NA	NA	NA
	n = 100:	NA	NA	NA

Table 5.4: Statistical Difference for Random Simulations (where $p_1 = p_2 = 0.9$)

*: p < 0.05, and **: p < 0.01

		Degree	Betweenness	Closeness
Case 0	n = 10:	No Difference	No Difference	No Difference
	n = 50:	No Difference	No Difference	No Difference
	n = 100:	No Difference	No Difference	Difference**
Case 1	n = 10:	No Difference	No Difference	NA
	n = 50:	Difference*	No Difference	NA
	n = 100:	Difference*	No Difference	NA
Case 2	n = 10:	No Difference	No Difference	NA
	n = 50:	No Difference	No Difference	NA
	n = 100:	No Difference	No Difference	NA
Case 3.1	n = 10:	No Difference	No Difference(p val = 1)	NA
	n = 50:	NA	NA	NA
	n = 100:	NA	NA	NA
Case 3.2	n = 10:	No Difference	No Difference	Difference**
	n = 50:	No Difference	No Difference	NA
	n = 100:	No Difference	No Difference	NA
Case 4.1	n = 10:	No Difference	No Difference	NA
	n = 50:	NA	NA	NA
	n = 100:	NA	NA	NA
Case 4.2	n = 10:	No Difference	NA	NA
	n = 50:	NA	NA	NA
	n = 100:	NA	NA	NA

Table 5.5: Statistical Difference for Random Simulations (where $p_1 = 0.1$ vs. $p_2 = 0.5$)

*: p < 0.05, and **: p < 0.01

		Degree	Betweenness	Closeness
Case 0	n = 10:	Difference**	Difference**	Difference**
	n = 50:	Difference**	Difference**	Difference**
	n = 100:	Difference**	Difference**	Difference**
Case 1	n = 10:	Difference*	No Difference	NA
	n = 50:	Difference**	Difference**	NA
	n = 100:	Difference**	Difference**	NA
Case 2	n = 10:	Difference**	NA	NA
	n = 50:	Difference**	Difference**	NA
	n = 100:	Difference**	Difference**	NA
Case 3.1	n = 10:	Difference**	No Difference	NA
	n = 50:	NA	NA	NA
	n = 100:	NA	NA	NA
Case 3.2	n = 10:	Difference**	Difference**	Difference**
	n = 50:	Difference*	NA	NA
	n = 100:	Difference**	Difference**	NA
Case 4.1	n = 10:	Difference**	No Difference	NA
	n = 50:	NA	NA	NA
	n = 100:	NA	NA	NA
Case 4.2	n = 10:	Difference**	No Difference	NA
	n = 50:	No Difference	NA	NA
	n = 100:	NA	NA	NA

Table 5.6: Statistical Difference for Random Simulations (where $p_1 = 0.1$ vs. $p_2 = 0.9$)

*: p < 0.05, and **: p < 0.01

		Degree	Betweenness	Closeness
Case 0	n = 10:	Difference**	Difference**	Difference**
	n = 50:	Difference**	Difference**	Difference**
	n = 100:	Difference**	Difference**	Difference**
Case 1	n = 10:	Difference**	Difference**	NA
	n = 50:	Difference**	Difference**	NA
	n = 100:	Difference**	Difference**	NA
Case 2	n = 10:	Difference**	No Difference	NA
	n = 50:	Difference**	Difference**	NA
	n = 100:	Difference**	Difference**	NA
Case 3.1	n = 10:	Difference**	No Difference	NA
	n = 50:	NA	NA	NA
	n = 100:	NA	NA	NA
Case 3.2	n = 10:	Difference**	Difference**	NA
	n = 50:	Difference**	NA	NA
	n = 100:	Difference**	Difference**	NA
Case 4.1	n = 10:	Difference**	NA	NA
	n = 50:	NA	NA	NA
	n = 100:	NA	NA	NA
Case 4.2	n = 10:	Difference**	Difference*	NA
	n = 50:	NA	NA	NA
	n = 100:	Difference*	NA	NA

Table 5.7: Statistical Difference for Random Simulations (where $p_1 = 0.5$ vs. $p_2 = 0.9$)

*: p < 0.05, and **: p < 0.01

		Degree	Betweenness	Closeness
Case 0	n = 10:	Difference**	Difference**	Difference**
	n = 50:	No Difference	No Difference	Difference**
	n = 100:	No Difference	No Difference	Difference**
Case 1	n = 10:	Difference**	Difference**	NA
	n = 50:	Difference**	Difference**	NA
	n = 100:	Difference**	Difference**	Difference**
Case 2	n = 10:	Difference**	Difference*	NA
	n = 50:	Difference**	Difference**	NA
	n = 100:	Difference**	Difference**	NA
Case 3.1	n = 10:	Difference**	No Difference	NA
	n = 50:	NA	NA	NA
	n = 100:	NA	NA	NA
Case 3.2	n = 10:	Difference**	Difference*	Difference**
	n = 50:	No Difference	No Difference (1)	NA
	n = 100:	Difference**	Difference**	NA
Case 4.1	n = 10:	Difference*	Difference**	NA
	n = 50:	NA	NA	NA
	n = 100:	NA	NA	NA
Case 4.2	n = 10:	No Difference	No Difference	NA
	n = 50:	NA	NA	NA
	n = 100:	No Difference	NA	NA

References

[1] Aurich, Nathassia K. and Alves Filho, José O. and Marques da Silva, Ana M. and Franco, Alexandre R., “Evaluating the reliability of different preprocessing steps to estimate graph theoretical measures in resting state fMRI data”, *Frontiers in Neuroscience*, 2015, Vol. 9, 10.3389/fnins.2015.00048.

[2] Liu, Haoyang and Barber, Rina Foygel. “Between hard and soft thresholding: optimal iterative thresholding algorithms.” *Information and Inference: A Journal of the IMA*, 2019, Vol. 9, 899-933, doi: 10.1093/imaiai/iaz027

[3] Newman, Mark. *Networks*, Oxford University Press, Oxford, United Kingdom, 2018

[4] Ortiz, Richard J., et al. “Functional Connectivity Differences between Two Culturally Distinct Prairie Vole Populations: Insights into the Prosocial Network” *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 2022, Vol. 7, No. 6, pp. 576–587., https://doi.org/10.1016/j.bpsc.2021.11.007.

[5] Rothman, Adam J., Levina, Elizaveta, and Zhu, Ji. “Generalized Thresholding of Large Covariance Matrices.” *Journal of the American Statistical Association*, 2009, Vol. 104, No. 485, 177-186., doi: 10.1198/jasa.2009.0101

[6] Qiu, Yumou, and Janaka S S Liyanage. “Threshold selection for covariance estimation.” *Biometrics*, 2019, Vol. 75, No. 3 , 895-905., doi:10.1111/biom.13048

[7] Worsley, K.J, “Statistical Analysis of Activation Images” *Functional Magnetic Resonance Imaging: An Introduction to Methods*(eds. Jezzard, P., Matthews, P. M. Smith, S. M.). Oxford University, 2001., doi: 10.1093/acprof:oso/9780192630711.003.0014

Appendix A

Graphs

The following are some of the graphs mentioned from chapter 4.

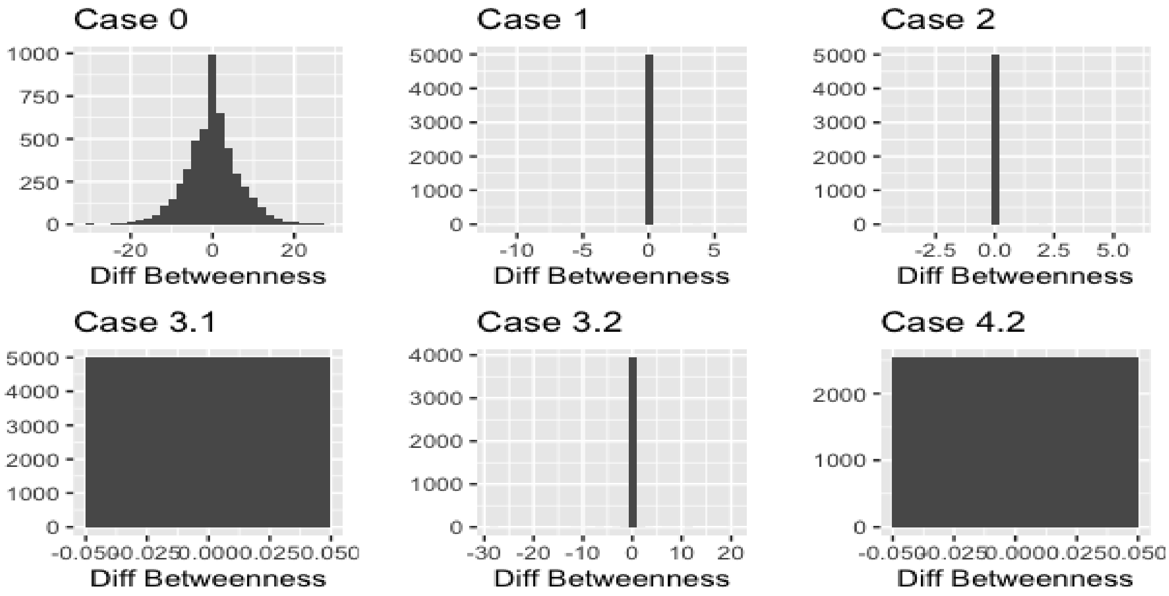

Figure A.1: n = 10: Differences in Betweenness for $p_1 = p_2 = 0.5$

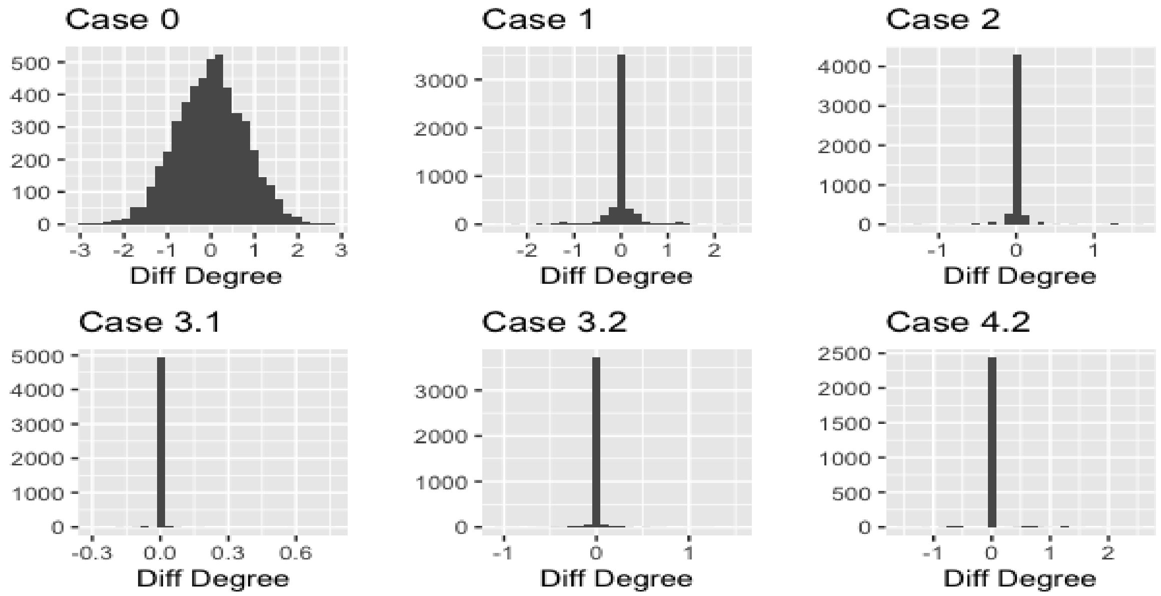

Figure A.2: n = 10: Differences in Degree for $p_1 = p_2 = 0.9$

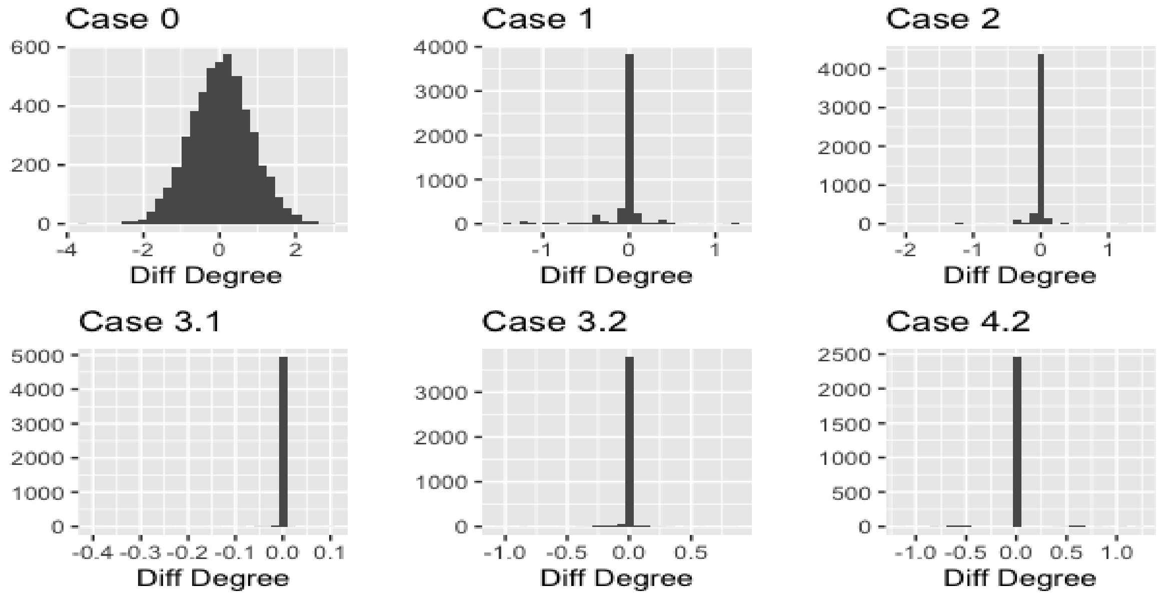

Figure A.3: n = 10: Differences in Degree for $p_1 = 0.5$ vs $p_2 = 0.9$

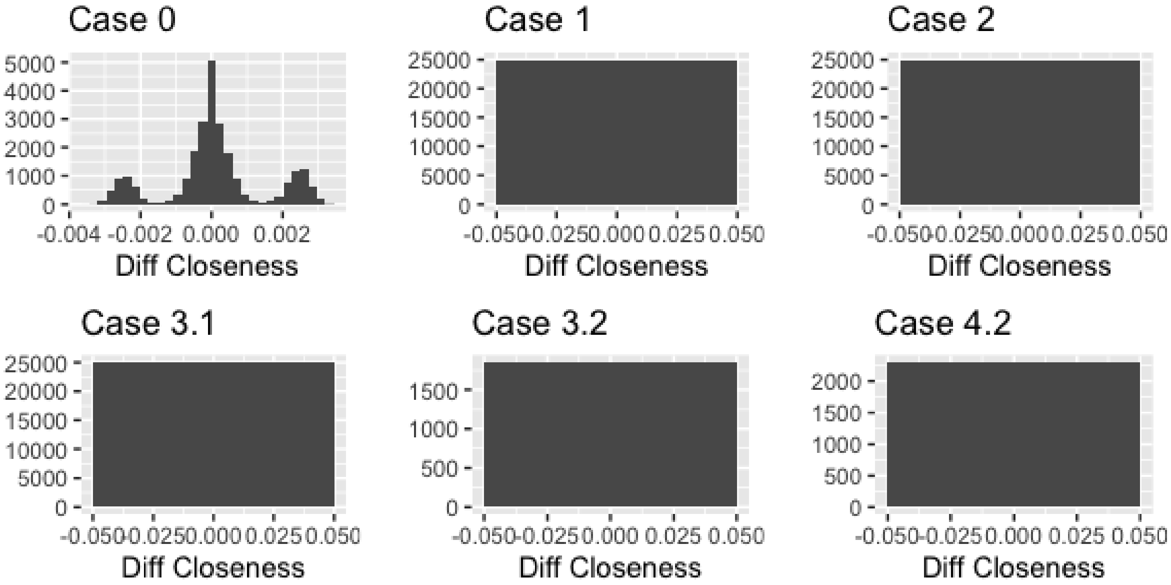

Figure A.4: n = 50: Differences in Closeness for $p_1 = p_2 = 0.1$

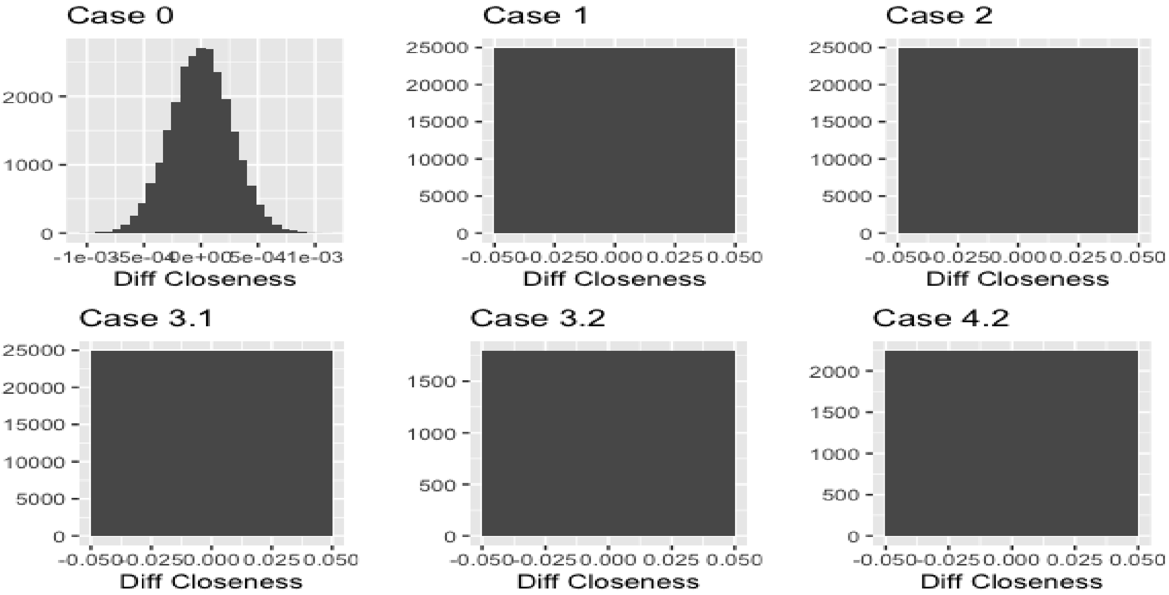

Figure A.5: n = 50: Differences in Closeness for $p_1 = p_2 = 0.9$

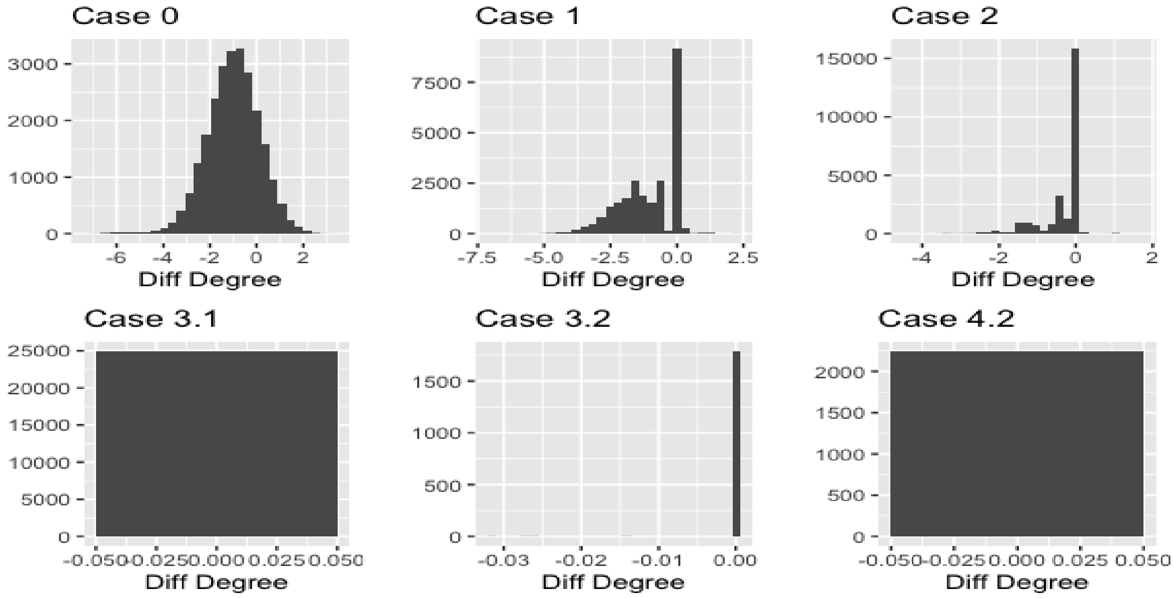

Figure A.6: n = 50: Differences in Degree for $p_1 = 0.1$ vs $p_2 = 0.9$

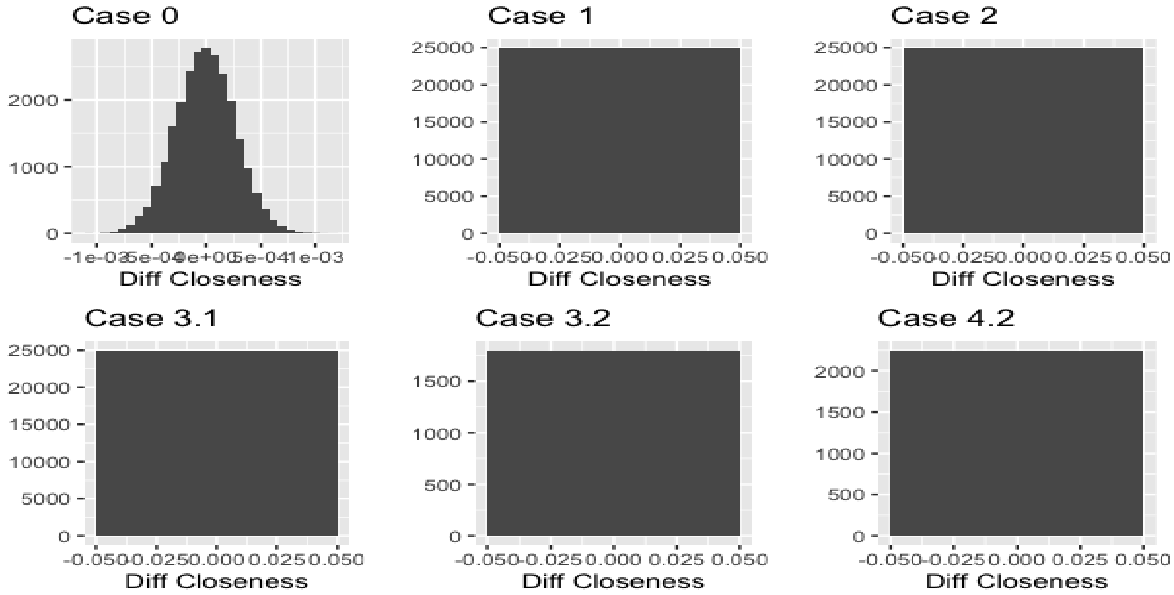

Figure A.7: n = 50: Differences in Closeness for $p_1 = 0.5$ vs $p_2 = 0.9$

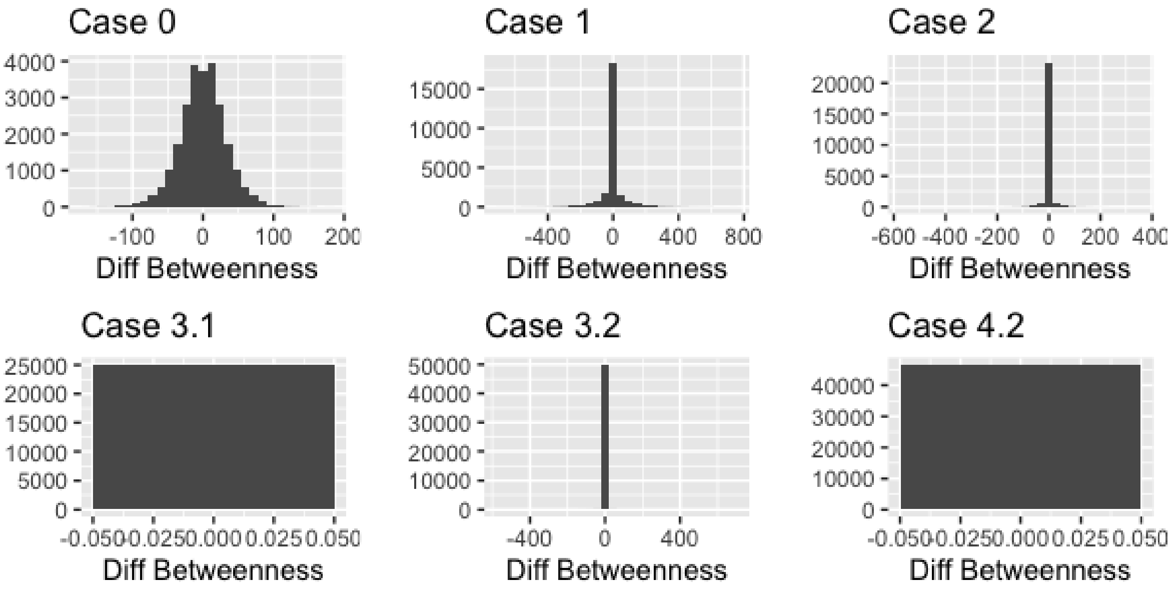

Figure A.8: n = 100: Differences in Betweenness for $p_1 = p_2 = 0.9$

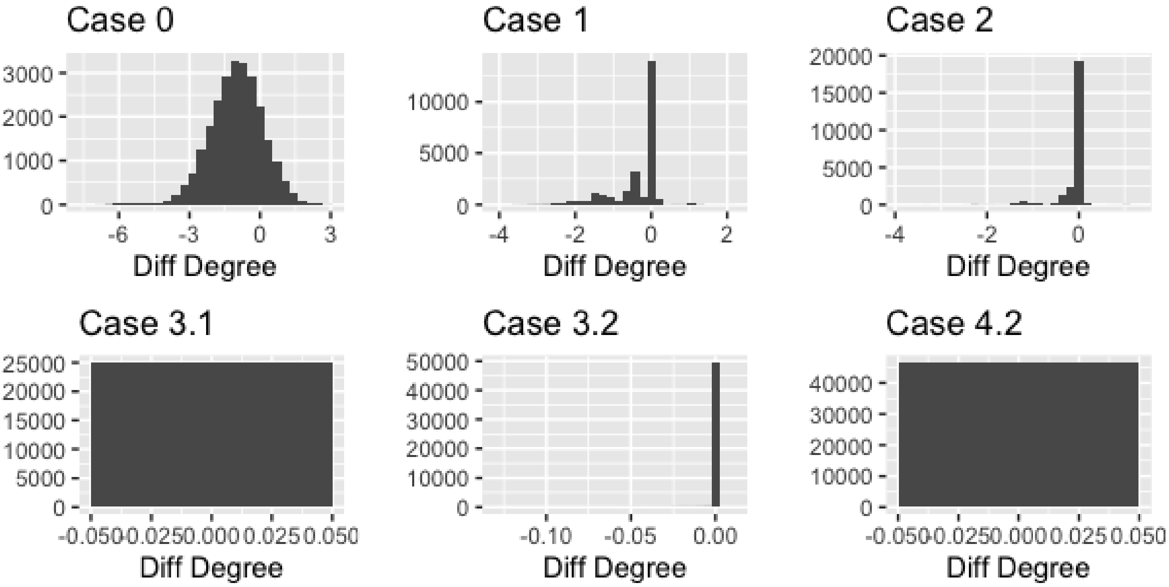

Figure A.9: n = 100: Differences in Degree for $p_1 = 0.1$ vs $p_2 = 0.5$

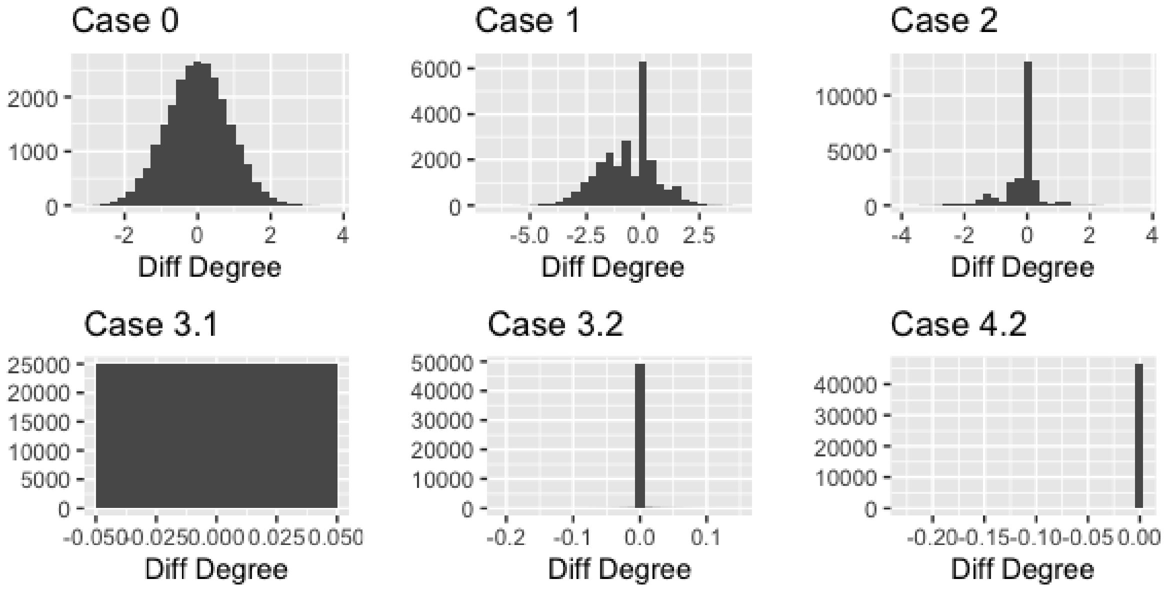

Figure A.10: n = 100: Differences in Degree for $p_1 = 0.5$ vs $p_2 = 0.9$

Curriculum Vitae

Hortencia J. Hernandez was born January 12^{th}, 1998 to Agustin Jose Hernandez IV and Lisa Hernandez in Waco, Texas. She was raised in Waco, Texas until her father's job required them to move to Georgia and then finally to El Paso, Texas where she graduated high school.

Upon graduating high school, she went on to pursue a bachelor of science in mathematics with a music minor at Baylor University in 2016. While in school she served in multiple organizations such as Kappa Kappa Psi, a national honorary band fraternity, and alpha Kappa Delta Phi, an international Asian-interest sorority, while being a manager at a restaurant.

Graduating in the fall of 2019 she continued to work until the pandemic hit, which cut her out of her job. She quickly started working with her sister at an assisted living home in San Antonio, TX. Deciding on wanting to continue her educational journey she enrolled at the University of Texas at El Paso (UTEP) to pursue a Masters of Science in Statistics. Upon graduation, she will start her Ph.D. in Data Science in the Spring of 2023 at UTEP.

In addition to fulfilling her academic achievements, she is currently apart of musical ensembles at UTEP, including the Mariachi and Flute Choir. She works as a Graduate Teaching Assistant as well as a Math/Statistics Tutor at El Paso Community College. Hortencia's email address is hjhernandez2016@gmail.com