THE DISTRIBUTION OF BETWEENNESS CENTRALITY IN EXPONENTIAL RANDOM GRAPH MODELS

by

Christina Durón

A Dissertation
Presented to the Faculty of
Claremont Graduate University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
in
Mathematics

Spring 2019

APPROVAL OF THE DISSERTATION COMMITTEE

This dissertation has been duly read, reviewed, and critiqued by the Committee listed below, which hereby approves the manuscript of Christina Durón as fulfilling the scope and quality requirements for meriting the degree of Doctor of Philosophy in Mathematics.

Johanna Hardin, Co-Chair
Department of Mathematics, Pomona College
Professor

Ami Radunskaya, Co-Chair
Department of Mathematics, Pomona College
Professor

Allon Percus
Institute of Mathematical Sciences, Claremont Graduate University
Professor

ABSTRACT OF THE DISSERTATION

THE DISTRIBUTION OF BETWEENNESS CENTRALITY
IN EXPONENTIAL RANDOM GRAPH MODELS

by

CHRISTINA DURÓN

Doctor of Philosophy in Mathematics
Claremont Graduate University, 2019

Although network centrality measures have been employed in the analysis of biological networks to obtain rankings of influential nodes, the statistical interpretation of the centrality measures and their distributions in a random network remain an open question. A framework that utilizes the generalized exponential random graph model is proposed as an approach to determine the distribution of the finite stable betweenness centrality measure. The theoretical underpinnings of generalized exponential random graph models are detailed, with particular attention towards the derivation of the probability model and parameter estimation process. Finally, several different distributions on graphs with 20 nodes and a fixed topology are considered, and the resulting distributions of the finite stable betweenness are compared.

DEDICATION

To my family and husband, Jeffrey Bell.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisors Johanna Hardin and Ami Radunskaya. It has been an honor to be your PhD student. I cannot thank you enough for the endless amount of patience, encouragement, and guidance you have given me over the past few years. I know how much of a privilege it is to have one brilliant and successful female advisor, so to have had two exceptional role models, I am beyond honored. Thank you for all that you have done for me - I cannot fully express my gratitude, but please know that it is endless.

I would like to thank Daniel Pick for funding the Daniel Pick Fellowship in molecular biology, and for his enthusiasm and support of my work.

I would like to thank the professors, and in particular, Marina Chugunova, at Claremont Graduate University for their constant support. I am proud to be a graduate of the university.

For this dissertation, I would like to thank my committee members: Johanna Hardin, Ami Radunskaya, and Allon Percus for their time, interest, and helpful feedback.

Lastly, I would like to thank my parents Suzanne and Zee Durón, my siblings Andrea and Ziyad, and grandparents for the love, encouragement, and support they have given me in the pursuit of my dreams. But most of all, thank you to my loving and incredibly patient husband, Jeffrey, who has been my cheerleader since the day we met. I love you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Understanding the function of biological networks heavily depends upon understanding the network's underlying structure. As a result, centrality measures have been utilized in the analysis of biological networks, as they provide rankings of influential nodes within the network and thus, a better idea of the properties, features, and sub-networks that contribute to the network's biological complexity (Breitkreutz et al., 2012; Mistry et al., 2017; Ramadan et al., 2016; Zhang et al., 2013). While there exist many centrality measures to identify influential nodes within the network, a statistical interpretation of the centrality values associated with each measure (e.g., is a betweenness value of 100 statistically significant?) and their distributions in a random network remain an open question.

Although formally defined in Chapter 1 along with other notations and definitions, a biological network is a collection of nodes representing biological molecules (e.g., genes or proteins) that are joined by edges representing tissue-specific functional associations between nodes (Greene et al., 2015; Wang et al., 2018). In regards to biological applications, genes whose role significantly changes as tissue transitions from a healthy to diseased state may potentially play a role in the development of a disease. Because nodes in biological networks can be characterized by a set of centrality measures, these measures may be used to identify structural differences between biological networks representing two different states of a particular tissue.

To that end, Chapter 2 details a methodology to identify genes that are essential to the structure of diseased tissue. In particular, the framework utilizes gene expression data to construct two weighted networks, one from a group of healthy tissue samples (i.e., the

Normal network) and the other from a group of diseased samples (i.e., the Tumor network). The betweenness centrality measure, which quantifies the involvement of a node in shortest paths within the network, is applied to each weighted network to identify genes that comprise a regulatory network unique to low-grade brain tumors arising in the optic nerves of *NF1* mutant mice. Using datasets from both sets of tissue samples simultaneously, the *Etv5* network is identified as a defining feature of the diseased state in mouse and human low-grade glioma tumors.

If different tissue samples had been collected, thus altering the edge-weights of both the Normal and Tumor networks, and the aforementioned approach repeated, an important question to consider is whether the betweenness centrality would have identified *Etv5* as a gene essential to the structure of low-grade brain tumors. To address this question, Chapter 3 examines the variability of the betweenness centrality measure to identify genes whose role is different in tissues from a healthy as compared to a diseased state. Using a previously constructed regulatory network, gene expression data from pediatric brain tumors are used to create two separately weighted networks, one based upon each of the healthy and diseased sets of samples. Following the methodology detailed in Chapter 2 to obtain a set of genes essential to the topology of the Tumor network, the variability of the betweenness measure is analyzed using two edge-weight perturbation methods. The results of the perturbation methods indicate a robustness of the betweenness centrality in the identification of genes essential to the structure of diseased tissue.

Although the betweenness measure can be utilized in the identification of structurally important genes, the large range of betweenness values that result from edge-weight perturbations is concerning and suggests an instability of the measure. While formally proven

in Chapter 4, the betweenness measure is unstable in the sense that an arbitrarily small change to edge-weights can cause large fluctuations in its value. Motivated by the sensitivity of this measure to edge-weight perturbations, Chapter 4 defines an alternative, *stable*, centrality measure, the *finite stable betweenness*, whose distribution will be explored in generalized exponential random graph models (GERGMs). The theoretical underpinnings of both exponential and generalized exponential random graph models, which specify the probability distribution of a set of unweighted and weighted networks, respectively, are discussed in Chapter 5, with particular attention given to the derivation of the GERGM probability model and coefficient estimation process.

The utilization of GERGMs as a model for the distribution of the finite stable betweenness measure is highlighted in Chapter 6. In particular, the proposed framework to determine a distribution rests upon the assumption inherent in the GERGM probability model, that the structure of the observed network may be explained by network configurations. Because the betweenness centrality captures some global structural properties of a network (Abbasi et al., 2012; Alahakoon et al., 2011), the proposed framework to determine a distribution of finite stable betweenness focuses on incorporating the measure into a GERGM probability model as a network configuration. Various models to determine the distribution of the measure are proposed and compared to the distribution of finite stable betweenness of a small 20-node network constructed from a known structure. Finally, conclusions, future work, and supplementary material are highlighted in Chapter 7.

CHAPTER 1

NETWORK DEFINITIONS

A biological network is formally defined as an abstract, undirected, and weighted graph $G = (V, E)$ where V is the set of nodes representing biological molecules (e.g., genes or proteins), and E is the set of edges representing the functional, causal, or physical interactions between nodes that is associated with a weight function $w : E \to \mathbb{R}^n \geq 0$. Denote the set of edges $E = \{e_{j,k} | v_j, v_k \in V\}$ by the set of connections between nodes v_j and v_k of strength $w(e_{j,k}) = w_{j,k}$.

Any two nodes connected by an edge are considered *adjacent*, and a *path* $P(v_j, v_k)$ between nodes v_j and v_k is defined as a sequence of edges that connect adjacent nodes:

$$P(v_j, v_k) \quad = \quad \{e_{j=j_0,j_1}, e_{j_1,j_2}, ..., e_{j_{n-1},j_n=k}\}$$

where edge $e_{j_i,j_{i+1}}$ connects adjacent nodes v_{j_i} and $v_{j_{i+1}}$ for $0 \leq i \leq n-1$.

Define the *length of a path* $P(v_j, v_k)$ as the sum of the edge-weights of the edges in $P(v_j, v_k)$:

$$\text{len}\left(P(v_j, v_k)\right) \quad = \quad \sum_{i=0}^{n-1} w\left(e_{j_i,j_{i+1}}\right), \qquad \text{with } j_0 = j \text{ and } j_n = k$$

Let $\mathcal{P}(v_j, v_k)$ be the set of all paths between nodes v_j and v_k. Define $P_s(v_j, v_k) \in \mathcal{P}(v_j, v_k)$ as a *shortest path* between nodes v_j and v_k to be a path of minimum length such

that

$$\text{len}\left(P_s(v_j, v_k)\right) \leq \text{len}\left(P(v_j, v_k)\right), \forall P \in \mathcal{P}$$

Finally, define $s(v_j, v_k)$ as the *length of the shortest path* $P_s(v_j, v_k)$ between nodes v_j and v_k:

$$s(v_j, v_k) = \text{len}\left(P_s(v_j, v_k)\right)$$

An example of a weighted network can be seen in Figure 1.1. Although a path from node 1 to node 4 may be defined as $P(v_1, v_4) = \{e_{1,2}, e_{2,3}, e_{3,4}\}$, a shortest path between the node pair is $P_s(v_1, v_4) = \{e_{1,3}, e_{3,4}\}$.

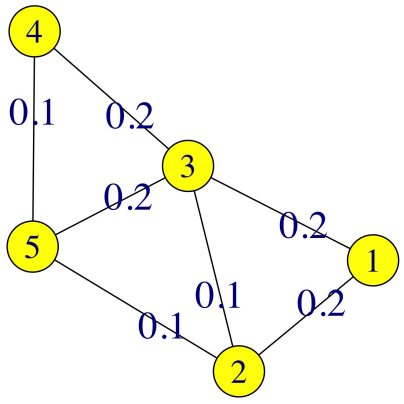

Figure 1.1. A weighted sample network with nodes $V = \{1, 2, 3, 4, 5\}$ and edges $E = \{e_{1,2}, e_{1,3}, e_{2,3}, e_{2,5}, e_{3,4}, e_{3,5}, e_{4,5}\}$ with corresponding edge-weights $\{0.2, 0.2, 0.1, 0.1, 0.2, 0.2, 0.1\}$.

1.0.1 Network Centrality Measures

Network nodes can be characterized by several centrality measures, all of which evaluate the importance of each node through a partial ranking based upon the network's topological features and edge-weights. In a biological context, a gene (i.e., a node) that lies on a large fraction of shortest paths between other genes may indicate the extent of the gene's influence and ability to control the flow of information within the network (Breitkreutz et al., 2012). To identify the gene(s) that lie on these communication pathways, the betweenness centrality measure may be utilized in the analysis of the biological network. A thorough discussion of centrality measures is provided by Boccaletti et al. (2006); Newman (2003).

Betweenness Centrality

The betweenness centrality measure quantifies the involvement of a node v_i in the shortest paths within a network by calculating the sum of the fractions of shortest paths that pass through v_i (Freeman, 1977). The betweenness centrality of node v_i, $C_B(v_i)$, in a network is formally defined in Equation 1.1 and can be computed using Algorithm 1. Note that for every pair of nodes in a connected network, there exists at least one shortest path.

Definition. The **betweenness** centrality of node $v_i \in V$ is given by:

$$C_B(v_i) \quad = \sum_{j \neq k \neq i} \frac{\sigma_{v_j, v_k}(v_i)}{\sigma_{v_j, v_k}} \tag{1.1}$$

where $v_j, v_k \in V$, σ_{v_j, v_k} is the total number of shortest paths from node v_j to node v_k, and $\sigma_{v_j, v_k}(v_i)$ is the number of those paths that pass through v_i.

Algorithm 1 Calculating $C_B(v_i)$: the betweenness centrality of node v_i

1: For each node pair (v_j, v_k), calculate the shortest paths between them.

2: For the fixed node v_i, for each $j \neq k \neq i$, determine the fraction of shortest paths between the node pair (v_j, v_k) that pass through node v_i.

3: Sum these fractions over all node pairs (v_j, v_k).

Consider the weighted network in Figure 1.1. Nodes 2 and 5 have the largest betweenness centrality of the node set, as they both lie on the largest fraction of shortest paths within the network. The shortest paths of each node pair in the weighted network are listed in Table 1.1, while the subsequent betweenness centrality values of each node in the weighted network are listed in Table 1.2.

Table 1.1. The shortest paths calculated for each node pair in the weighted network given in Figure 1.1.

Node pair (v_j, v_k)	**Shortest Path(s)**	**Node(s)** v_i **on Shortest Path(s)**
(1,2)	$e_{1,2}$	None
(1,3)	$e_{1,3}$	None
(1,4)	$e_{1,3}, e_{3,4}$ and $e_{1,2}, e_{2,5}, e_{5,4}$	2, 3, 5
(1,5)	$e_{1,2}, e_{2,5}$	2
(2,3)	$e_{2,3}$	None
(2,4)	$e_{2,5}, e_{5,4}$	5
(2,5)	$e_{2,5}$	None
(3,4)	$e_{3,4}$	None
(3,5)	$e_{3,5}$	None
(4,5)	$e_{4,5}$	None

Table 1.2. The betweenness centrality of each node in the weighted sample network given in Figure 1.1.

Node v_i	**Node Pair** (v_j, v_k)	**Fraction of Shortest Paths through Node** v_i	$C_B(v_i)$
1	(2,3), (2,4), (2,5), (3,4), (3,5), (4,5)	$\frac{0}{1}, \frac{0}{1}, \frac{0}{1}, \frac{0}{1}, \frac{0}{1}, \frac{0}{1}$	0
2	(1,3), (1,4), (1,5) (3,4), (3,5), (4,5)	$\frac{0}{1}, \frac{1}{2}, \frac{1}{1}, \frac{0}{1}, \frac{0}{1}, \frac{0}{1}$	1.5
3	(1,2), (1,4), (1,5) (2,4), (2,5), (4,5)	$\frac{0}{1}, \frac{1}{2}, \frac{0}{1}, \frac{0}{1}, \frac{0}{1}, \frac{0}{1}$	0.5
4	(1,2), (1,3), (1,5) (2,3), (2,5), (3,5)	$\frac{0}{1}, \frac{0}{1}, \frac{0}{1}, \frac{0}{1}, \frac{0}{1}, \frac{0}{1}$	0
5	(1, 2), (1,3), (1,4) (2,3), (2,4), (3,4)	$\frac{0}{1}, \frac{0}{1}, \frac{1}{2}, \frac{0}{1}, \frac{1}{1}, \frac{0}{1}$	1.5

CHAPTER 2

NETWORK ANALYSIS USING BETWEENNESS CENTRALITY

2.1 INTRODUCTION

As with other ecological systems, mammalian tissues can be considered as complex biological systems made up of a multitude of elements that each contribute to the overall biological function. In this regard, both healthy and diseased tissues contain distinct, yet interacting, cell types and molecular components that establish distinctive functional states for diseased tissue relative to their healthy counterparts. Therefore, a natural implication of this conceptualization is the idea that healthy and diseased tissues can be defined in an objective manner using computational approaches. Some examples include algorithms that have been used to classify diseased states, to assess risk as a function of specific factors such as gender, age, and environmental exposures (Bajenaru et al., 2003, 2005; Kaul et al., 2014), and to identify individualized treatments based on gene expression profiles (Daginakatte and Gutmann, 2007; Daginakatte et al., 2008; Hegedus et al., 2008).

But computational modeling can also be utilized to identify interactions that exist within the tissue. As discussed in Chapter 1, one example of this modeling consists of classifying the interactions and relationships as a biological network, where the nodes represent proteins, transcription factors, or genes, and the edges connecting nodes represent communication pathways. Networks that highlight distinct differences between these healthy and diseased tissues may serve as a way to identify unexpected relationships critical to the maintenance of a disease, such as a tumor.

To investigate these potential networks and relationships, an authenticated murine

model of a brain tumor (optic glioma) that arises in children with the neurofibromatosis type 1 (*NF1*) cancer predisposition syndrome is analyzed (Bajenaru et al., 2003, 2005). The optic glioma tumors (World Health Organization grade I pilocytic astrocytomas) are low grade neoplasms that develop in early childhood (Listernick et al., 2007). Because the tumors are not removed or biopsied in children with *NF1*, the *NF1* genetically-engineered mouse low-grade glioma model system best embodies many of the features present in the human condition and has been successfully utilized to evaluate promising targeted therapies now in clinical trials for children with the tumors (Daginakatte and Gutmann, 2007; Daginakatte et al., 2008; Hegedus et al., 2008; Kaul et al., 2014) (http://clinicaltrials.org; NCT01089101, NCT01158651 and NCT01734512).

Although applications of network analysis to understand cancer biology are relatively new, its analysis can be divided into three methodological types: (1) enrichment of fixed gene sets, (2) *de novo* subnetwork construction and clustering, and (3) network-based modeling (Creixell et al., 2015). Although all three methods can be applied to characterize murine *NF1* optic gliomas, only the network-based approach (i.e., the third type) is able to use relational network information to define specific regulator (i.e., nodal) connections.

2.2 NETWORK CONSTRUCTION

A gene regulatory network (GRN) depicts how some genes that encode regulatory molecules, such as transcription factors or microRNAs, control the expression of other genes (Narang et al., 2015). The glioma regulatory network, which is used as the base network in the following analysis, was inferred from the Rembrandt microarray data set available from GEO GSE68848 by Madhavan et al. (2009) and constructed by Margolin et al. (2006). The

regulatory network was derived from a transcription (RNA) dataset (i.e., the human dataset) representing a variety of different gliomas to capture the regulatory interactions. Four hundred twenty-seven human glioma gene expression profiles were obtained from the Rembrandt data repository (Madhavan et al., 2009), and were combined to create the glioma regulatory network according to the ARACNe-AP algorithm (Lachmann et al., 2016; Margolin et al., 2006). Note that the Rembrandt data were generated through the Glioma Molecular Diagnostic Initiative and include 874 glioma specimens.

The ARACNe-AP (Algorithm for the Reconstruction of Accurate Cellular Networks) algorithm reconstructs gene regulatory networks from large-scale gene expression data (Lachmann et al., 2016). Although the steps of the ARACNe-AP algorithm are summarized in Algorithm 2, additional details are discussed in Madhavan et al. (2009).

Algorithm 2 The ARACNe-AP algorithm

1: The input to the algorithm is a list of transcription factors and gene expression profile data.

2: The gene expression data is pre-processed in order to determine a mutual information (MI) threshold. Specifically, all pairwise MI scores between gene expression profiles are estimated, and then their significance is assessed by comparing them to a null dataset. Note that the significance level depends on the sample size of the input.

3: A random sample is selected from the input gene expression profiles.

4: The gene expression profiles are rank-transformed, and then the MI for each transcription factor/target pair is calculated.

5: The MI threshold from Step 2 is used to remove any connections that are not statistically significant.

6: Indirect interactions are removed using the Data Processing Inequality tolerance filter, as described by Margolin et al. (2006).

Using the glioma regulatory network as a base network, previously generated RNA expression data (i.e., the murine dataset) from the optic nerves of $NF1^{\text{flox/flox}}$ (N, healthy control group, $n = 4$) and $NF1^{\text{flox/mut}}$; GFAP-Cre (OPG-1, tumor diseased group, $n = 11$) (Pan et al., 2017) were used to create two separately weighted networks: Normal (i.e., healthy) and Tumor (i.e., diseased). Because the glioma regulatory network (generated according to the ARACNe-AP algorithm) provided the topology used in their construction, the Normal and Tumor networks maintained an identical topological structure. Furthermore, genes that were not expressed in one of the two groups were removed from both networks. Although structurally identical, the Normal and Tumor networks differ in the weights assigned to each edge, as each weight is calculated based upon the distance between two nodes (i.e., between two genes) v_j and v_k. This distance is calculated as $1 - |\rho(x_j, x_k)|$, where $\rho(x_j, x_k)$ is the minimum of the Pearson and Spearman correlations between the RNA expression levels of gene v_j and gene v_k. Additional details concerning correlation measures are provided in Section 3.3.2.

2.3 NETWORK CENTRALITY MEASURES

Network centrality measures are commonly employed to identify nodes that may potentially play important roles in weighted networks (Zhang and Horvath, 2005). Some commonly used measures include closeness, betweenness, and entropy; for example West et al. (2012) employ differential entropy between diseased and healthy tissue to detect relevant genes.

Recall the betweenness centrality of gene v_i (i.e., node v_i), $C_B(v_i)$, is a measure derived

from the fraction of shortest paths in the network that pass through gene v_i:

$$C_B(v_i) \ = \sum_{j \neq k \neq i} \frac{\sigma_{v_j, v_k}(v_i)}{\sigma_{v_j, v_k}}$$

where $v_j, v_k \in V$, σ_{v_j, v_k} is the total number of shortest paths connecting genes v_j and v_k, and $\sigma_{v_j, v_k}(v_i)$ is the number of shortest paths that pass through gene v_i. For a closer look at betweenness, along with a concrete example of how to calculate the value of each node in a weighted network, refer to Chapter 1.

In both the Normal and Tumor networks, the length of the path between any two genes is provided by the sum of the distances, or the weights, of its edges. Because the edge-weights (i.e., distances) are based upon the RNA expression data, the betweenness measures differ between the Normal and Tumor networks, and may be utilized to identify genes essential to the structure of the diseased tissue. Although other centrality measures could have been employed, such as $C_C(v_i)$, the *closeness centrality* for node v_i, defined as

$$C_C(v_i) \ = \frac{1}{\sum_k d(v_i, v_k)}, \quad \text{where } d(v_i, v_k) = \text{ the distance between nodes } v_i \text{ and } v_k$$

the large range of the betweenness values led to a clear discrimination between the Normal and Tumor networks.

Figures 2.1 and 2.2 show the comparison of using the betweenness and closeness measures to differentiate the Normal and Tumor networks. As is evident from both figures, the betweenness measure identifies genes that are substantially different across the two networks, while the closeness measure does not identify any such stand-out genes.

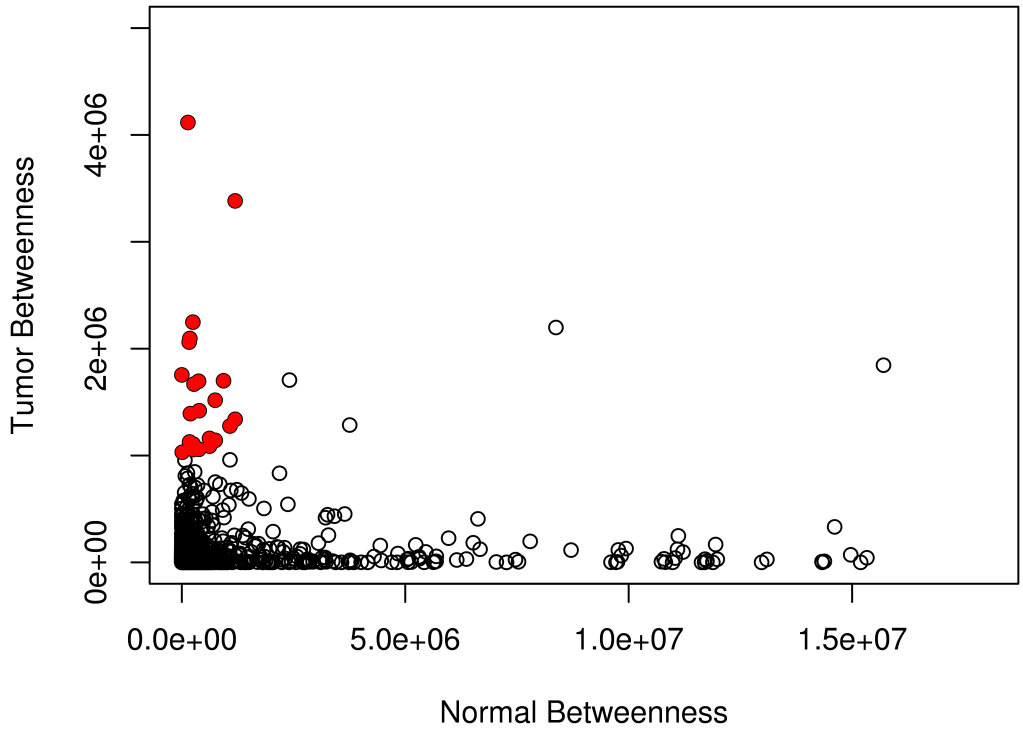

Figure 2.1. A comparison of betweenness measures in the Normal and Tumor networks. Filled (red) circles indicate genes whose betweenness measure is both at least 1.1 times as large in the Tumor network as in the Normal network, and whose Tumor betweenness value is greater than 1e6. These genes are listed in Table 2.1, and are shown in pink in Fig 2.3.

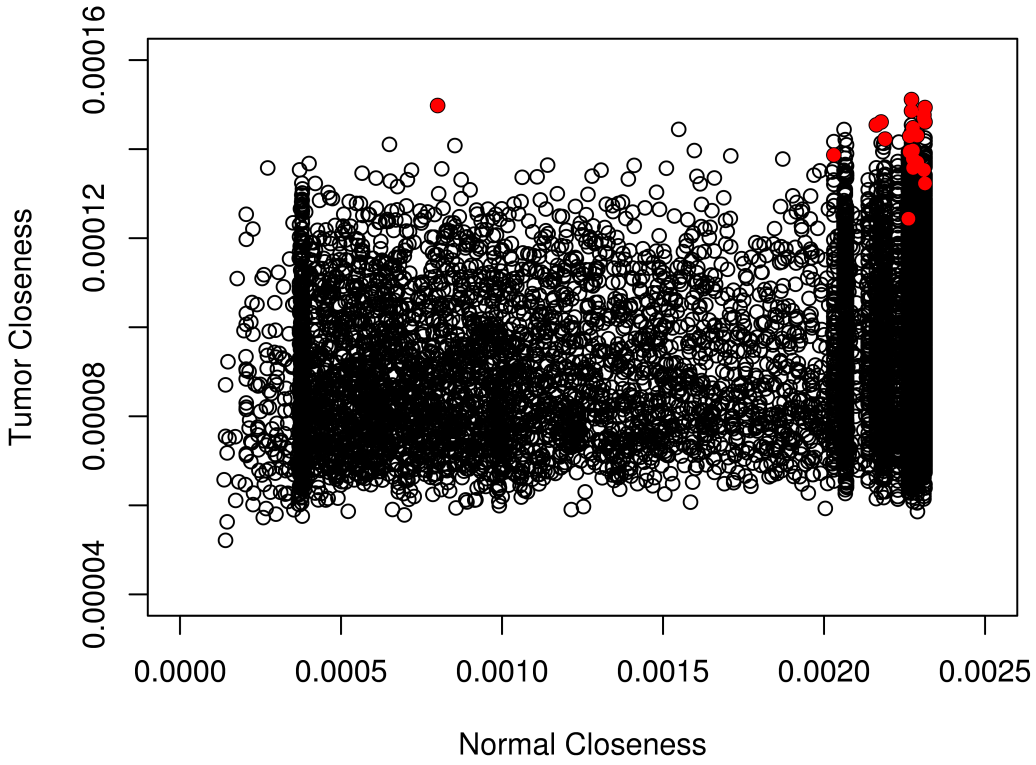

Figure 2.2. Filled (red) circles indicate genes whose betweenness measure is both at least 1.1 times as large in the Tumor network as in the Normal network, and whose Tumor betweenness value is greater than 1e6. These genes are listed in Table 1, and shown in pink in Figure 2.3. Note that the closeness measure does not differentiate the Tumor and Normal networks as well as betweenness (see Fig 2.1).

2.4 BIOINFORMATIC APPROACH

In order to identify genes that may play a role in disease evolution, it is critical to identify the changes in the biological networks that describe cellular processes. Genes whose role significantly changes across healthy versus diseased tissue states are classified as "potential genes of interest" and are identified through network centrality analysis. In this work, the betweenness centrality measure is used to identify genes of potential interest

Using the murine dataset, twenty-three genes are identified as having betweenness values at least 1.1 times as large in the Tumor network as in the Normal network, along with Tumor betweenness values greater than 1e6. These genes are shown as filled, red circles in Fig 2.1, are listed in Table 2.1, and are denoted as "essential" to the glioma regulatory network.

Table 2.1. The "essential genes" are the transcription factors that have been identified as having betweenness values at least 1.1 times as large in the Tumor network as in the Normal network, and along with Tumor betweenness values greater than 1e6.

Cebpz	*Etv5*	*Spen*	*Zcchc14*	*Camta1*	*Chd5*	*Cers2*
Hnrnpab	*Ilf2*	*Zcchc17*	*Zc3h15*	*Tulp4*	*Purb*	*Rpl7*
Tcf3	*Tead1*	*Cnbp*	*Prdm2*	*Sarnp*	*Zranb2*	*Gcsh*
Ift74	*Myl12B*					

Using the four healthy samples and eleven diseased samples, differential expression analysis was employed to test the null hypothesis that the expected count of a given gene is the same for all samples, where each gene of the twenty-three essential genes and their targets

were considered. Note that count denotes the number of sequence fragments that have been assigned to each gene in each sample. Differential expression calculates the coefficients β of a generalized linear model that best fit the observed count K:

$$K_{ij} \sim NB(s_j \mu_{ij}, a_i)$$

with expected gene expression strength,

$$\log \mu_{ij} = \beta_{i0} + x_j \beta_{iT}$$

More specifically, the counts K_{ij} for gene i in sample j are modeled using a negative binomial distribution with fitted mean $s_j \mu_{ij}$ and a gene-specific dispersion parameter a_i (Love et al., 2014). The fitted mean is the product of a parameter μ_{ij} (the expectation value of the observed counts for gene i in sample j) and a sample-specific size factor s_j (the sequencing depth of each sample j), where the purpose of the size factor is to make the counts from different samples, which may have been sequenced to different depths, comparable (Anders and Huber, 2010). Note that the coefficients β_{iT} represent the fold-changes for gene i, while x_j equals 0 or 1, if j is a healthy or diseased treatment sample, respectively. Given that the null hypothesis is $\beta_{iT} = 0$, if the value of β_{iT} is significantly different from zero, then the alternative model that there is a difference in fold-change is accepted. (Love et al., 2014).

From the results of the differential expression analysis, one gene, *Etv5*, emerged as having a significant differential effect across the two expression datasets on the murine dataset (Table 2.2). Furthmore, *Etv5* was found to have the highest percentage of differentially

expressed targets of all of the twenty-three essential genes (Table 2.3).

A list of the thirty-one differentially expressed target genes of *Etv5* are listed in Table 2.4. The network consisting of the essential regulators (i.e., a gene involved in controlling the expression of one or more genes), along with the differentially expressed targets of *Etv5*, is shown in Figure 2.3. The combination of the different computational techniques utilized in the analysis that identified *Etv5* as an essential regulator is listed in Algorithm 3.

Algorithm 3 A computational and bioinformatic approach: Identifying the *Etv5* network

1: A reference network is identified for all transcription factors. In this analysis, the regulatory network was created with the ARACNe-AP algorithm and is used as the reference network.

2: Weights are added to the edges of the reference network using one minus the minimum of the Pearson or Spearman correlation of independently generated samples of RNA-Seq data. Two different weighted networks are created - one with diseased RNA-Seq data (i.e., the Tumor network), and one with healthy RNA-Seq data (i.e., the Normal network).

3: From each weighted network, the betweenness value for each gene is calculated.

4: Genes with betweenness values at least 1.1 times as large in the Tumor network relative to the Normal network, along with Tumor betweenness values greater than 1e6, are identified. These genes are considered "essential" in the Tumor network (Table 2.1).

5: Of the "essential" genes, the most differentially expressed gene is identified (*Etv5*).

6: Genes which represent targets of the *Etv5* transcription factor and are differentially expressed across the two treatments (i.e., diseased and healthy) are identified (Table 2.4).

2.5 CONCLUSION

Using the betweenness network analysis and the murine dataset, *Etv5* and its associated genes are discovered as a differentially regulated transcriptional network in low-grade brain tumors relative to healthy brain tissue. This type of network analysis is extremely useful in identifying potential network regulators essential to the diseased state. Taken together, the methodology discussed in Algorithm 3 is a valuable tool for identifying genes essential to the tumor ecosystem, and it nicely illustrates the value of combining both computational and bioinformatic approaches to characterize the diseased state relative to its healthy counterpart. The official publication of the work detailed in this chapter may be found in Pan et al. (2018).

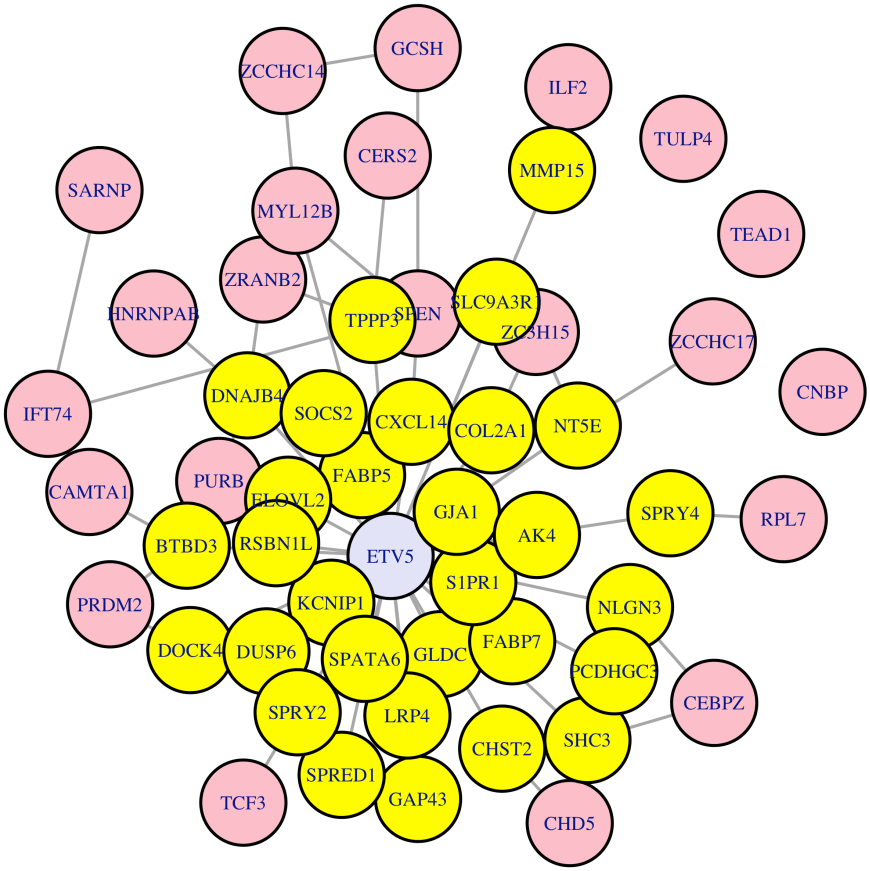

Figure 2.3. The *Etv5* network is comprised of *Etv5* (in lavender in the center), and its differentially expressed targets (shown in yellow). The remaining twenty-two essential genes, identified by their high betweenness measures in the Tumor network relative to the Normal network, are shown on the periphery in pink.

Table 2.2. The results of the differential expression analysis to identify which of the twenty-three essential genes identified using the betweenness measure are significantly different between the diseased and healthy tissue samples. Note that fold-change is a measure of the change in the expression level of a gene.

Gene	***p*-value**	**log2 Fold-Change**
Etv5	1.34E-09	1.4474
Myl12B	6.81E-04	-0.5850
Zc3h15	4.03E-03	-0.5603
Camta1	4.22E-03	-0.4705
Spen	4.61E-03	0.5714
Tulp4	6.11E-03	0.4845
Sarnp	7.36E-03	0.7022
Zcchc17	9.34E-03	-0.6699
Ift74	9.66E-03	-0.5475
Tcf3	1.29E-02	0.4928
Zcchc14	3.82E-02	0.3660
Rpl7	3.83E-02	-0.5100
Cnbp	4.16E-02	-0.3724
Zranb2	6.71E-02	-0.3515
Tead1	1.40E-01	0.2794
Ilf2	1.51E-01	-0.3077
Cebpz	1.60E-01	-0.3255
Gcsh	3.92E-01	-0.1581
Cers2	5.36E-01	-0.0923
Hnrnpab	6.23E-01	0.0784
Purb	6.80E-01	0.0573
Prdm2	8.75E-01	0.0207
Chd5	8.98E-01	-0.0334

Table 2.3. The percent of target genes from the regulatory network that are differentially expressed (i.e., that are significantly different between the diseased and healthy tissue samples) by each of the essential genes that are identified using the betweenness measure.

Gene	**# Significant**	**Total Targets**	**% Significant**
Etv5	31	449	0.0690
Cers2	32	496	0.0645
Sarnp	15	255	0.0588
Zcchc14	16	361	0.0433
Tcf3	28	640	0.0437
Tead1	19	443	0.0429
Zc3h15	27	652	0.0414
Tulp4	17	412	0.0413
Rpl7	10	247	0.0405
Purb	27	678	0.0398
All	529	14926	0.0354
Hnrnpab	19	543	0.0350
Cnbp	15	458	0.0328
Camta1	35	1093	0.0320
Chd5	27	852	0.0317
Zcchc17	8	257	0.0311
Prdm2	20	652	0.0307
Spen	12	392	0.0306
Cebpz	10	379	0.0264
Ilf2	13	555	0.0234
Zranb2	11	486	0.0226
Gcsh	0	0	0
Ift74	0	0	0
Myl12B	0	0	0

Table 2.4. The thirty-one targets of *Etv5* that are identified as differentially expressed in the murine dataset.

Gene	***p*-value**	**log2 Fold-Change**
Spry2	1.13E-08	0.916
Dnajb4	2.63E-05	-0.457
Col2a1	1.07E-04	1.235
Spred1	1.04E-05	0.800
Dusp6	1.50E-04	0.637
S1pr1	6.84E-17	1.358
Ak4	5.38E-05	0.959
Fabp5	1.39E-09	1.087
Fabp7	6.22E-05	0.956
Rsbn1	2.16E-04	-0.428
Btbd3	7.12E-09	0.755
Gap43	1.94E-05	-0.797
Gja1	3.64E-10	0.600
Gldc	6.60E-06	1.161
Kcnip1	3.06E-04	-0.797
Igfbp6	3.07E-04	-0.942
Lrp4	5.38E-05	0.671
Mmp15	2.11E-04	1.003
Nt5e	1.18E-04	-0.744
Pcdhgc3	2.12E-06	0.932
Tppp3	6.01E-05	-0.912
Shc3	2.71E-04	0.903
Nlgn3	1.82E-05	0.705
Spata6	1.62E-04	-0.552
Elovl2	1.62E-11	1.908
Spry4	5.77E-05	1.104
Socs2	1.77E-04	-0.814
Slc9a3r1	1.41E-06	0.722
Chst2	2.28E-11	1.163
Cxcl14	4.88E-17	1.425
Dock4	2.88E-05	0.642

CHAPTER 3

THE VARIABILITY OF BETWEENNESS CENTRALITY IN THE IDENTIFICATION OF ESSENTIAL GENES

3.1 INTRODUCTION

Current technology provides clinicians with high-dimensional measurements that describe features of a particular tissue, or cellular environment. An example of high-dimensional measurements include gene expression data for samples of cells that can be divided into "healthy" (i.e., Normal) and "diseased" (i.e., Tumor) groups. Figure 3.1 shows a heat map of gene expression data from thirty tissue samples (i.e., the glioma dataset) which are publicly available from GEO GSE42656. Of the thirty tissue samples, sixteen samples are of healthy tissue, while fourteen are of diseased samples of low-grade gliomas.

Using the glioma dataset in Figure 3.1, one question of interest to address is: "What is the difference between the healthy and the diseased groups?". Chapter 2 highlighted the use of a computational and bioinformatic approach that involved the analysis of both the betweenness centrality and differential expression on the murine dataset (defined in Section 2.2) to identify *Etv5* as an essential gene in the development of optic glioma. In that context, the term *essential* was defined as a way to identify genes that met two threshold criteria based on the betweenness centrality measure. In particular,

Definition. Let M and M' be two networks with identical topology (i.e., with the same nodes and edges connecting the nodes), but different edge-weights, and let T_1 and T_2 be two

pre-determined threshold values. A node v_i is defined as **essentially different** if:

$$\frac{C_B^M(v_i)}{C_B^{M'}(v_i)} > T_1, \qquad \text{and} \quad C_B^M(v_i) > T_2 \tag{3.1}$$

where $C_B^M(v_i)$ and $C_B^{M'}(v_i)$ are the betweenness centrality values of node v_i in networks M and M', respectively.

A discussion on how the threshold values T_1 and T_2 can be determined is provided in Section 3.4. Additional details are provided by Pan et al. (2018).

Although *Etv5* was experimentally validated as a regulator (i.e., a gene involved in controlling the expression of one or more genes) essential to the structure of diseased tissue in the murine dataset (Chapter 2 and Pan et al. (2018)), an important question to consider is: "Would the same gene have been identified had a different sample of healthy and diseased tissues been collected?" Thus, the motivation of this chapter is to develop an analytic framework to gauge the robustness of the betweenness centrality as a method for identifying key regulators and provide computational validation when experimental validation is not possible. The glioma dataset is used to assess the variability of the procedure.

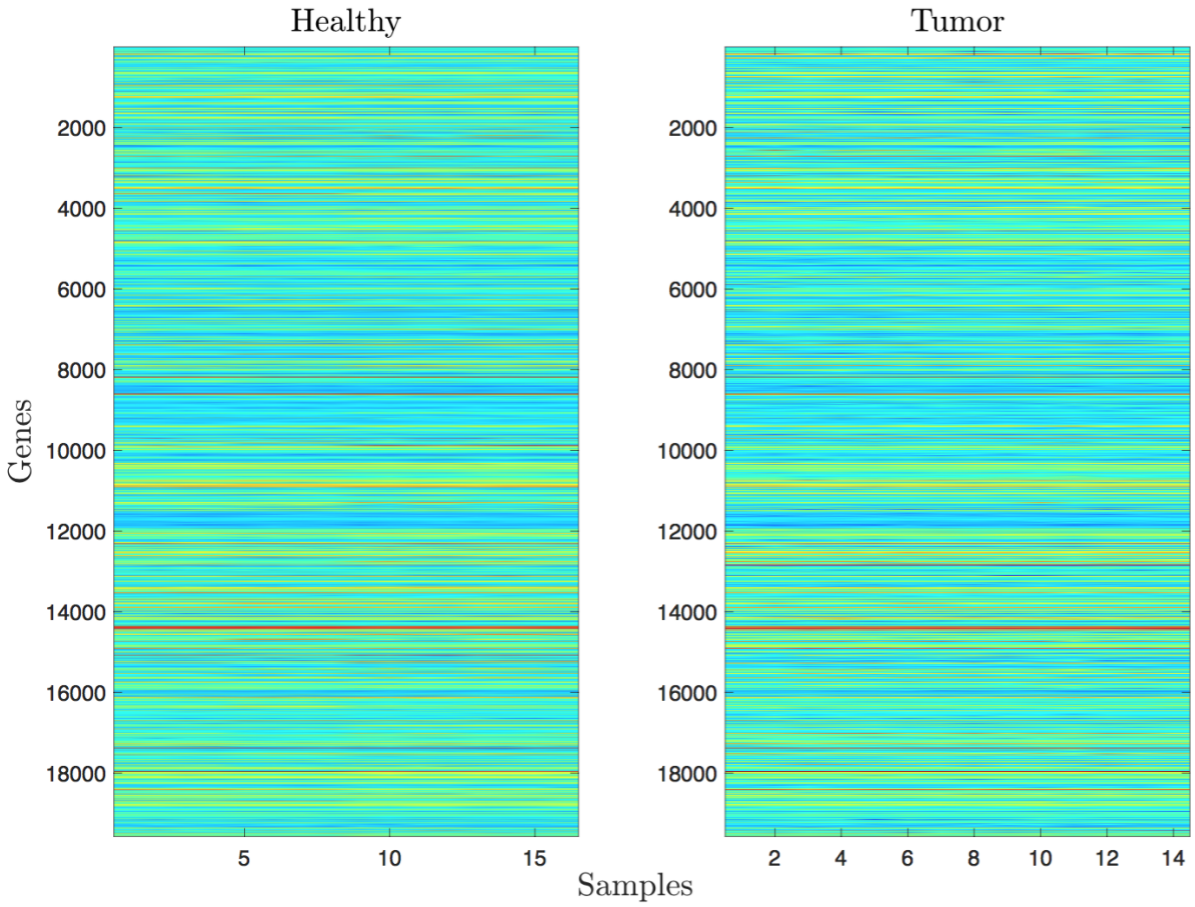

Figure 3.1. A heat map showing the gene expression from the glioma dataset. Each horizontal row represents one gene, while each column represents a distinct sample. The sixteen samples on the left are from healthy tissues, while the fourteen samples on the right are from diseased tissues. The colors indicate gene expression levels.

3.2 VARIABILITY IN BETWEENNESS CENTRALITY

Because an experiment can be thought of as a sample of observations from a larger population, a network constructed from the RNA expression data is an estimate of the population network, defined as a theoretical construct given by gene expression data on all possible samples in a population of interest.

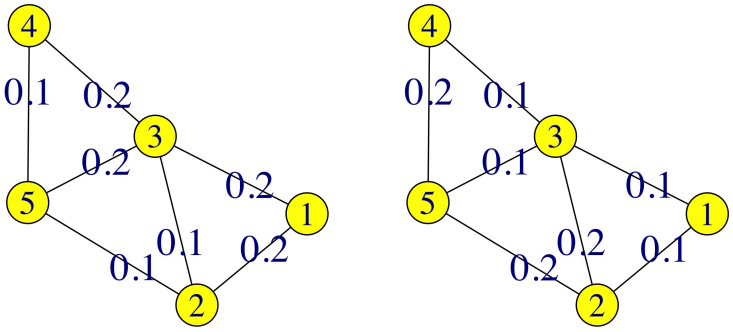

Figure 3.2. Two structurally identical weighted networks, true (left) and observed (right), with nodes $V = \{1, 2, 3, 4, 5\}$ and edges $E = \{e_{1,2}, e_{1,3}, e_{2,3}, e_{2,5}, e_{3,4}, e_{3,5}, e_{4,5}\}$, but differing edge-weights. The betweenness values for the nodes in the true network are provided in the table at left, while the betweenness values for the nodes in the observed network are provided in the table at right.

Node v_i	$C_B(v_i)$ **Parameter**
1	0
2	1.5
3	.5
4	0
5	1.5

Node v_i	$C_B(v_i)$ **Statistic**
1	0.25
2	0
3	3
4	0
5	0

Figure 3.2 (left) provides an example of the betweenness centrality values of a small, hypothetical population network. Recall the definition of the betweenness centrality of node $v_i \in V$:

$$C_B(v_i) = \sum_{j \neq k \neq i} \frac{\sigma_{v_j, v_k}(v_i)}{\sigma_{v_j, v_k}}$$

where $v_j, v_k \in V$, σ_{v_j, v_k} is the total number of shortest paths from node v_j to node v_k, and

$\sigma_{v_j,v_k}(v_i)$ is the number of those paths that pass through v_i.

Using the partial ranking provided by the betweenness centrality, nodes 2 and 5 lie on an equally large fraction of shortest paths in the network. Yet, consider Figure 3.2 (right), which is a perturbation of the population network representing the observed network based on sample data. Although both networks have identical topology, their difference in edge-weights results in different betweenness centrality values. This example motivates the following question: "How robust is betweenness centrality to sampling variability?"

A variety of network centrality measures, some in combination with other biological information, have been used to identify genes that are important in the development of a disease, where examples of such studies include Breitkreutz et al. (2012); Mistry et al. (2017); Ramadan et al. (2016); Zhang et al. (2013). The potential impact of the work in this chapter, coupled with the previously posed questions, point to a need for a theoretical framework to understand the variability of centrality measures in biological networks. Some work has already been done in this direction.

In Segarra and Ribeiro (2016), the authors provide formal definitions of stability and continuity for centrality measures in weighted networks. They show that betweenness centrality is unstable according to their definition, and propose an alternative stable definition of betweenness. Additional details of the stable betweenness centrality measure (Segarra and Ribeiro, 2016) is provided in Chapter 4.

Epskamp et al. (2017) take a statistical approach, discussing the stability of centrality rankings in terms of how the rankings can change with fewer observations. They also describe how to test whether differences in centrality measures between groups are significant by introducing a bootstrapped difference test for centrality measures. The proposed methods

are applied to a psychological network.

3.3 ANALYTIC FRAMEWORK

3.3.1 Notations and Definitions

To put the discussion in a statistical context, the betweenness network centrality measure will be considered as a *parameter* describing a 'population network' derived from the entire population of, for example, tumor patients of interest.

Definition. A **parameter** is a numerical summary value of the population from which data are obtained.

As in Chapter 2, recall that the topology, or structure of the network, is fixed: all of the networks in the collection have the same nodes and edges connecting the nodes. A particular weighted network from a collection, say, of tumor patients, can be characterized by the centrality measures of its set of nodes. In other words, the centrality measure of each node is a parameter of this family of networks. Given the theoretical assumption that one weighted network represents all possible "healthy" tissue, and another represents all possible "diseased" tissue, the value of the parameter representing the centrality of a particular node might distinguish between healthy and diseased tissue.

The weights on the network edges are assigned using correlations between expression levels of each gene represented as a node, where additional details on the construction of the edge-weights are provided in Section 3.3.2. The theoretical networks have edge-weights determined by the "true" correlations between the genes (i.e., the correlations determined

using all possible instances of "healthy" or "diseased" tissue). Because, in practice, the edge-weights are assigned using correlations between relatively few samples, only an estimate of the true centrality measure of each node can be obtained. In other words, the estimated centrality is a *statistic* that is the best estimate of the true centrality parameter.

Definition. A **statistic** is a numerical summary value of a **sample** from the population.

In what follows in Section 3.3.3, a confidence interval methodology is proposed to analyze the variability of the betweenness statistic to identify genes that are essentially different (refer to Equation (3.1)) in diseased tissue. In particular, two methods to perturb edge-weights are used to analyze the measure's variability.

3.3.2 Correlation Measures

As mentioned in Chapter 2, the edge-weight $w_{j,k}$ between nodes v_j and v_k in a biological network can be defined using correlation,

$$w_{j,k} = 1 - |\rho(v_j, v_k)| \tag{3.2}$$

where $\rho(v_j, v_k)$ is the correlation between the measurements of the expression of genes v_j and v_k. Since $\rho(v_j, v_k)$ is in [-1,1], the edge-weight $w_{j,k}$ is in [0,1]. When a pair of genes or proteins are perfectly correlated, the edge-weight between the corresponding nodes is 0. If two measurements are highly correlated, the weight on the edge between the corresponding nodes is small, increasing the possibility of the pair of nodes lying on a shortest path.

For the networks utilized in the confidence interval methodology (Section 3.3.3), the edge-weights are defined using Equation (3.2), where $\rho(v_j, v_k)$ is the minimum of the Pearson and Spearman-rank correlations between the RNA expression levels of gene v_j and gene v_k.

The **Pearson correlation** measures the extent of the linear association between the data on two nodes v_j and v_k, and is defined as

$$\rho_P(v_j, v_k) = \frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ik}-\bar{x}_k)^2}}$$

where x_{ij} and x_{ik} are the i^{th} observation of the expression level on genes or nodes v_j and v_k, respectively.

The **Spearman-rank correlation** measures the extent of monotonicity between the ranked data on two nodes v_j and v_k, and is defined as the Pearson correlation coefficient of the ranked (i.e., sorted from largest to smallest) observations for each node.

3.3.3 Confidence Interval Methodology

Suppose two weighted networks are constructed, one from a group of healthy tissue samples (i.e., the Normal network) and the other from a group of diseased tissue samples (i.e., the Tumor network). Assume that the structure of the networks describe the interactions between genes or proteins, represented as nodes, as their topology is based upon knowledge of the particular tissue being sampled. Furthermore, although both the Normal and Tumor networks are structurally identical, their edge-weights differ as they are determined by calculating correlations from data using Equation (3.2).

After identifying genes whose roles are substantially different by comparing healthy

and diseased states, and those that are essential to the Tumor network, the variability of the betweenness values of these genes is analyzed by perturbing the edge-weights of the original Normal and Tumor networks. In particular, two separate perturbation methods, non-parametric bootstrapping and the addition random noise to the gene-gene pair correlation values, are used in the analysis of the variability of the *betweenness difference statistic* around the betweenness difference parameter. Refer to Section 3.3.4 for additional details on each perturbation method.

Definition. The **betweenness difference parameter** of gene v_i is defined as the difference in Tumor and Normal log betweenness values of gene v_i from the population network:

$$DC_B(v_i) \quad = \quad \log(C_{B_{\text{Tumor}}}(v_i)) - \log(C_{B_{\text{Normal}}}(v_i)) \tag{3.3}$$

Algorithm 4 details the procedure for analyzing the variability of the betweenness difference statistic through confidence intervals. Recall that genes are considered "essentially different" if they meet two threshold criteria based on the betweenness centrality measure (refer to Equation (3.1)). Genes are defined as "significantly different" if they possess statistically significant betweenness confidence interval(s).

Although the methodology detailed in Algorithm 4 may be applied to other centrality measures, the betweenness measure is selected in this context as the best for distinguishing between the Normal and Tumor networks, as its large range of values leads to a clear distinction between the two networks. Refer to Figures 2.1 and 2.2 for an example of a comparison

Algorithm 4 Confidence interval method

1. Obtain the pre-determined list of essentially different genes identified using thresholds on betweenness values for both Tumor and Normal networks. Refer to Equation (3.1) for the formal definition of "essentially different".

2. Define the betweenness difference statistic of an essentially different gene v_i for the original Tumor and Normal networks as

$$\hat{DC}_B(v_i) \;=\; \log\left(\hat{C}_{B_{\text{Tumor}}}(v_i)\right) - \log\left(\hat{C}_{B_{\text{Normal}}}(v_i)\right)$$

for each of the essentially different genes.

3. Generate 100 simulated Tumor networks and 100 simulated Normal networks.

4. Define the betweenness difference statistic of an essentially different gene v_i from each of the 100 simulated networks as

$$\tilde{DC}_B(v_i) \;=\; \log\left(\tilde{C}_{B_{\text{Tumor}}}(v_i)\right) - \log\left(\tilde{C}_{B_{\text{Normal}}}(v_i)\right)$$

for each of the essentially different genes.

5. Construct 95% confidence intervals (CI) for the betweenness difference parameter of an essentially different gene v_i, $DC_B(v_i)$, where

$$CI \;=\; \hat{DC}_B(v_i) \pm 2\sigma_{\tilde{DC}_B(v_i)}$$

such that $\sigma_{\tilde{DC}_B(v_i)}$ is the standard error of $\tilde{DC}_B(v_i)$.

6. Classify an essentially different gene as significantly different if its confidence interval excludes 0.

of the betweenness and closeness measures applied to two identically structured networks generated from the murine dataset.

For Steps 2 and 4 in Algorithm 4, the log transformation is applied to the betweenness difference statistic in order to reduce the variability of the betweenness measure. Figure 2.1 highlights the variability of the betweenness measure.

Additionally, if an essentially different gene's confidence interval includes zero, no conclusions can be made about the value of the true betweenness ratio. However, essentially different genes whose confidence intervals exclude zero, as specified in Step 6 of Algorithm 4, are considered statistically different in the Tumor network as compared to the Normal network among the genes previously identified as essentially different from Step 1. Note that intervals which do not contain zero necessarily have only positive values due to the thresholds, which define essential genes.

3.3.4 Edge-Weight Perturbation Methods

Two perturbation methods, which produce variability in the edge-weights of the network differently, are used to simulate the distribution of the betweenness difference statistic. The two perturbation methods are discussed below.

Non-parametric Bootstrapping

Non-parametric bootstrapping is a general resampling technique that constructs a sampling distribution for a bootstrap statistic by resampling the data with replacement, thereby allowing duplicate selections in each sample. Bootstrapping generally follows three steps of (1) resampling a dataset with replacement, (2) calculating the statistic for each of the boot-

strapped samples, and (3) estimating the standard error for the bootstrap statistic using the standard deviation of the bootstrapped statistics.

Random Noise Using the Standard Error of Correlation

Given a bivariate normal distribution of samples, the standard error of the Pearson correlation $SE(\rho)$ is:

$$SE(\rho) \quad = \sqrt{\frac{1-\rho^2}{N_s-2}} \tag{3.4}$$

where ρ is the correlation between two nodes (i.e., genes) in a biological network, and N_s is the number of samples in the dataset used to calculate the correlation.

Using Equation (3.4), the theoretical standard error of the correlation for each gene-gene pair is calculated and used to generate the noise added to the correlation values. In general, adding noise using the theoretical standard error allows for network perturbations to remain within the typical theoretical sampling variation given by the estimated correlations. As the simulated noise must be both consistent with the data and constrain the correlations in [-1,1], the noise is generated from a truncated normal distribution centered at the original correlation value with standard deviation equal to standard error of correlation provided in Equation (3.4).

3.4 APPLICATION

Although the methodology detailed in Algorithm 4 may be generalized to a variety of networks and centrality measures, the variability of betweenness on a specific gene regulatory network is illustrated in the following application. As was utilized in the application in

Chapter 2, a previously constructed glioma regulatory network inferred from the human dataset was used as the base network. While the glioma network is publicly available as a supplement to Pan et al. (2018), refer to Section 2.2 for a more detailed discussion on its construction by Margolin et al. (2006).

Using the glioma regulatory network as a base network, the glioma dataset is used to create two separate weighted networks, one based on each of the healthy (i.e., the Normal network) and diseased (i.e., the Tumor network) sets of samples. The diseased group consists of fourteen pilocytic astrocytoma samples, a type of low-grade glioma. The healthy group consists of sixteen samples from healthy brain tissue. These two datasets were chosen because they contain healthy (non-neoplastic) tissue for comparison, and allow for comparisons to previously generated low-grade mouse glioma (Pan et al., 2018). Figure 3.1 illustrates the two datasets as side-by-side heat maps. While the networks corresponding to both groups have a fixed and identical structure (i.e., the same nodes and edges connecting the nodes), the weights on their edges differ. As was done in Chapter 2, genes that were not identified in one of the groups were removed from both the Normal and Tumor networks.

Following the procedure detailed in Algorithm 3 in Section 2.4, the set of essentially different genes is obtained based upon the Normal and Tumor networks constructed from the glioma dataset and is listed in Table 3.1. The set of essentially different genes was obtained using thresholds of a betweenness value at least 1.5 times as large in the Tumor network relative to the Normal network, along with a Tumor betweenness value greater than 950,000.

The variability of the betweenness difference statistic to identify the set of essentially different genes (see Table 3.1) is analyzed through the two perturbation approaches on the edge-weights: non-parametric bootstrapping (Section 3.4.1) and the addition of noise to

correlation values (Section 3.4.2). Using each perturbation method separately, 100 simulated Normal and 100 simulated Tumor networks are generated. The betweenness value of each of the essentially different genes listed in Table 3.1 is calculated from each of the 100 simulated Normal and Tumor networks generated by each perturbation method, and the collection of the betweenness values is used to generate a distribution of the betweenness difference statistic for each essentially different gene.

To justify the threshold values used to identify the set of essentially different genes, consider Figures 3.3 and 3.4, along with the formal definition of "accuracy":

$$\text{accuracy} \quad = \quad \frac{\#\text{ significantly different and essentially different}}{\#\text{ essentially different}} \tag{3.5}$$

The plots in Figure 3.3 and 3.4 are generated as follows: Algorithm 4 is applied using each threshold value on the xy-plane: $(\tilde{C}_{B_{\text{Tumor}}}/\tilde{C}_{B_{\text{Normal}}},\ \tilde{C}_{B_{\text{Tumor}}}/100000)$. For each perturbation method, 100 simulated Normal and Tumor networks are generated. The results of Algorithm 4 are two sets of genes: a set of essentially different genes and a set of significantly different genes. Using these two sets of genes, the accuracy at each threshold value is calculated according to Equation (3.5). The accuracy value is multiplied by 100 to give the percent accurate, which determines the height of the graph at the given point (i.e., the pair of threshold values) on the xy-plane. Additional justification for the particular threshold values used to determine the sets of essentially different genes is provided in Durón et al. (2018).

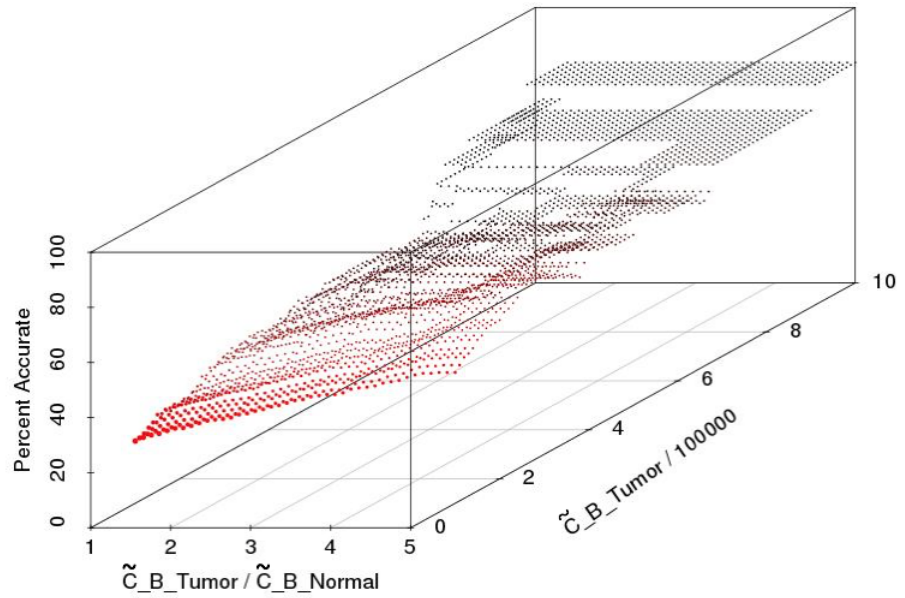

Figure 3.3. The identification accuracy of essential different genes (see Equation (3.5)) associated with non-parametric bootstrapping.

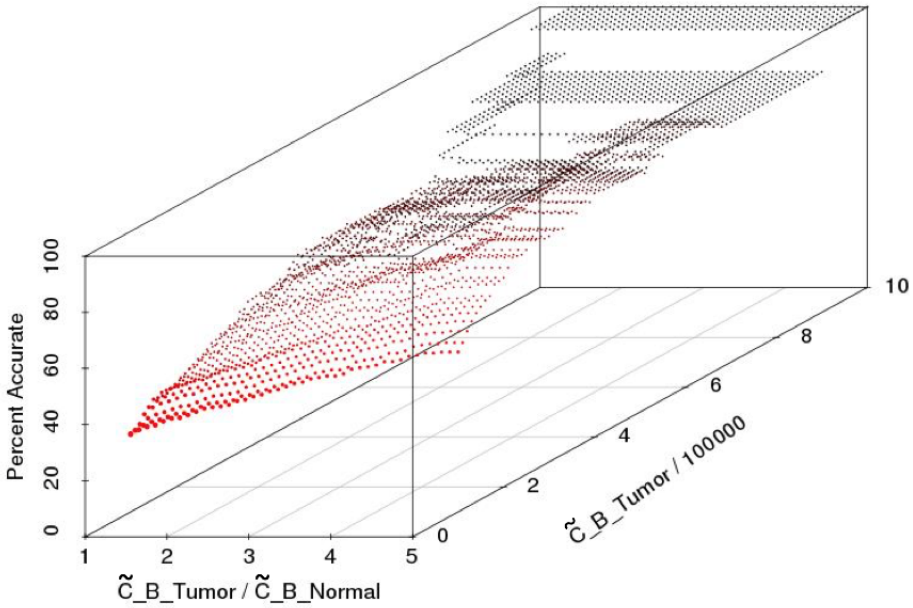

Figure 3.4. The identification accuracy of essential different genes (see Equation (3.5)) associated with the addition of random noise to the correlation values. In particular, noise from a truncated normal distribution centered at ρ with standard deviation given by Equation (3.4) is added to each initial correlation value, ρ.

Furthermore, in order to measure the consistency of the method for identifying essentially different genes as compared to the significance obtained through confidence intervals, the list of genes identified as essentially different is compared to the list of those identified as significantly different. Particular interest lies in the genes which are originally identified by their high betweenness centrality (i.e., the essentially different genes listed in Table 3.1), which are then also identified as significantly different across the simulated Normal and Tumor networks using the confidence interval method.

Table 3.1. The set of essentially different genes v_i identified using only thresholds on betweenness centrality. Recall that $\hat{C}_{B_{\text{Tumor}}}(v_i)$ and $\hat{C}_{B_{\text{Normal}}}(v_i)$ are the Tumor and Normal betweenness value, respectively, of the genes in the original Normal and Tumor networks.

Essentially Different Gene v_i	$\hat{C}_{B_{\mathbf{Normal}}}(v_i)$	$\hat{C}_{B_{\mathbf{Tumor}}}(v_i)$	$\hat{C}_{B_{\mathbf{Tumor}}}(v_i)/\hat{C}_{B_{\mathbf{Normal}}}(v_i)$
Cebpb	170000	1130000	6.65
Olig1	101000	1380000	13.7
Sox8	203000	1250000	6.16
Sp100	18700	1120000	59.9
Thra	80600	1550000	19.2

3.4.1 Perturbation with Non-Parametric Bootstrapping

By resampling with replacement of fourteen and sixteen samples from the diseased and healthy datasets, 100 bootstrapped Normal and 100 bootstrapped Tumor networks are generated. By applying the confidence interval procedure in Algorithm 4 to samples generated by non-parametric bootstrapping, the highest achievable level of accuracy is 80%, as

supported by Figure 3.3. The results of the bootstrapping perturbation method are listed in Table 3.2. Of the five essentially different genes listed in Table 3.1, only four genes have confidence intervals that exclude zero.

Table 3.2. The set of statistically different genes v_i identified through the proposed confidence interval method in Algorithm 4 using non-parametric bootstrapping. Recall that $\hat{DC}_B(v_i)$ is the difference in Tumor and Normal log betweenness values of gene v_i from the original networks, and $\bar{\tilde{DC}}_B(v_i)$ and $\sigma_{\tilde{DC}_B(v_i)}$ are the mean and standard deviation of the difference in Tumor and Normal log betweenness values of gene v_i of the simulated networks, respectively.

Statistically Different Gene v_i	$\hat{DC}_B(v_i)$	$\bar{\tilde{DC}}_B(v_i)$	$\sigma_{\tilde{DC}_B(v_i)}$	**95% Confidence Interval**
Olig1	2.61471	1.9240	0.811	(0.993, 4.2367)
Sox8	1.81769	1.5084	0.861	(0.095, 3.5404)
Sp100	4.09251	2.9666	1.326	(1.440, 6.7451)
Thra	2.95650	2.1650	1.140	(0.676, 5.2370)

The sampling distribution of betweenness for the set of statistically different genes identified through Algorithm 4 using non-parametric bootstrapping are depicted in Figure 3.5. Although an obvious distinction between the Normal (left) and Tumor (right) betweenness distributions is their shape, note that the mean of the betweenness distributions in the Tumor network is larger than those of the Normal network. Additionally, the large scale of the Tumor betweenness, as compared to the Normal betweenness, should not be surprising, as the genes of interest were the ones which possessed a larger Tumor than Normal betweenness.

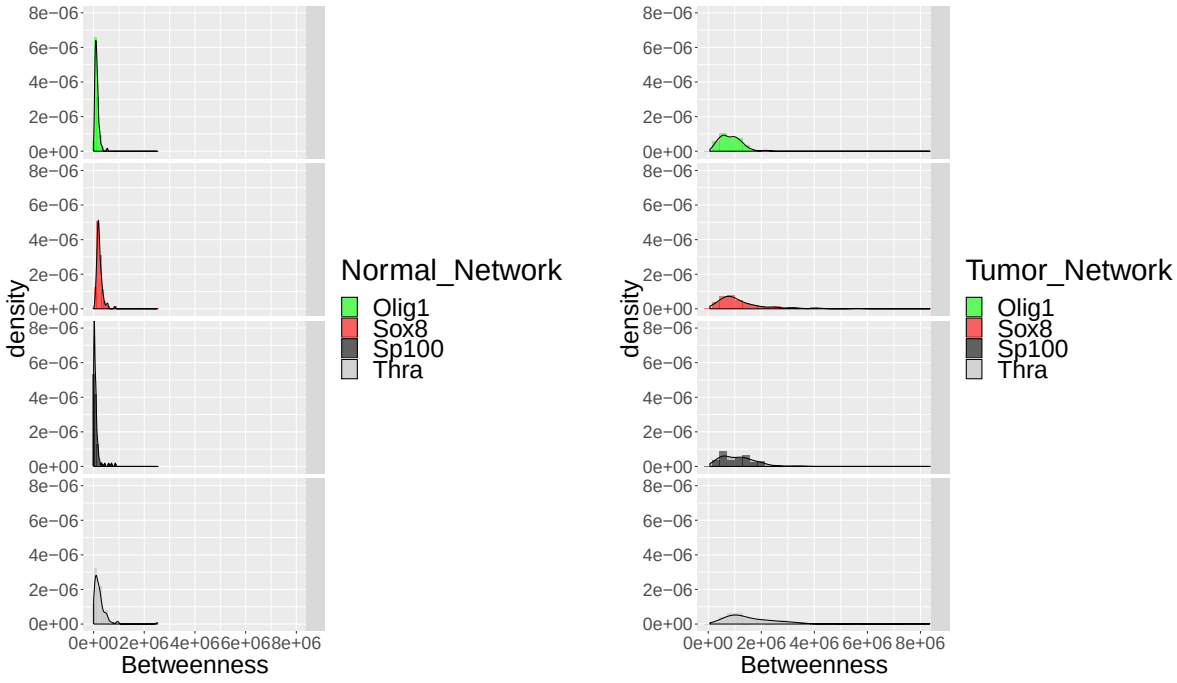

Figure 3.5. The sampling distribution of betweenness of the four statistically different genes in both the Normal (left) and Tumor (right) networks generated by non-parametric bootstrapping.

3.4.2 Perturbation with Random Noise Using the Standard Error of Correlation

By adding noise from a truncated normal distribution to the correlation value of each gene-gene pair, 100 simulated Normal and 100 simulated Tumor networks are generated. By applying the confidence interval procedure in Algorithm 4 to samples generated by adding noise to the correlation values, the highest achievable level of accuracy is 100%, as supported by Figure 3.4. The results of the random noise perturbation method are listed in Table 3.3. In particular, all of the five essentially different genes listed in Table 3.1 have confidence intervals that exclude zero.

Table 3.3. The set of statistically different genes identified through the proposed confidence interval method in Algorithm 4 by adding random noise from a truncated normal distribution to the correlation values.

Statistically Different Gene v_i	$\hat{DC}_B(v_i)$	$\bar{\bar{DC}}_B(v_i)$	$\sigma_{\tilde{DC}_B(v_i)}$	**95% Confidence Interval**
Cebpb	1.89417	1.9276	0.791	(0.3114, 3.477)
Olig1	2.61471	2.2570	0.961	(0.6929, 4.537)
Sox8	1.81769	2.2344	0.680	(0.4567, 3.179)
Sp100	4.09251	3.2332	0.795	(2.5019, 5.683)
Thra	2.95650	2.4985	0.659	(1.6385, 4.274)

The sampling distribution of betweenness for the set of statistically different genes identified through Algorithm 4 by adding random noise to the correlation values are depicted in Figure 3.6. The distinctions identified between the Normal and Tumor sampling distributions in Figure 3.5 are also identified between the distributions depicted in Figure 3.6. But unlike the sampling distributions determined by bootstrapping, there exists more variability in the distributions determined by adding random noise.

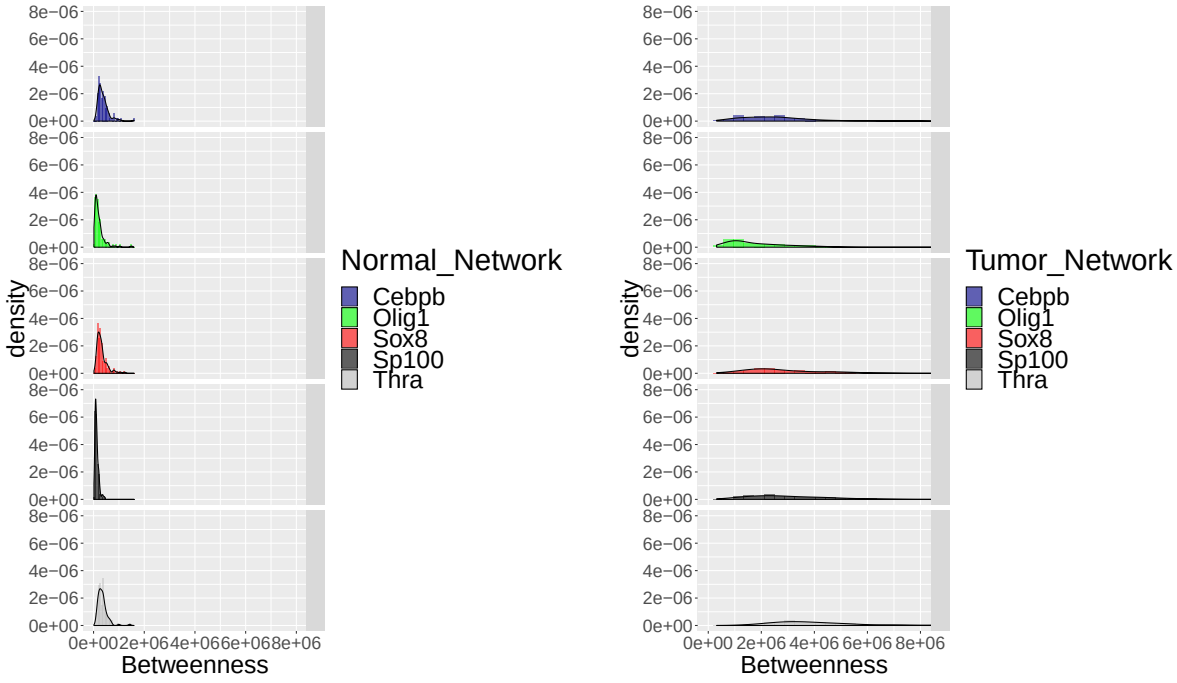

Figure 3.6. The sampling distribution of betweenness of the five statistically different genes in both the Normal (left) and Tumor (right) networks generated by adding random noise to the correlation values.

3.5 CONCLUSION

In the application, two separate weighted networks are generated based upon the glioma dataset, where the edge-weights are assigned using correlations between gene-gene measurements. The glioma dataset was selected because: (1) the availability of non-tumor healthy (i.e., normal) tissue samples, and (2) a previous study by Pan et al. (2018) which focused on low-grade gliomas from the murine dataset.

As highlighted in Chapter 2, the betweenness centrality measure can be used to identify a set of essentially different genes whose role substantially changes in the comparison of healthy to diseased states. But if a different sample of tissues had been collected, and the

methodology in Chapter 2 repeated, would the betweenness measure identify the same set of genes? To address this question concerning the variability of the betweenness statistic in the comparison of two networks, a theoretical framework was proposed to construct confidence intervals on the estimated betweenness difference measures. If an essentially different gene had a confidence interval that excluded zero, then the gene was determined to be statistically different across the Tumor and Normal networks.

Two separate edge-weight perturbation methods, non-parametric bootstrapping and the addition of random noise to correlation values, were used to simulate the distribution of the betweenness difference measure. Although both perturbation methods confirmed four of the five essentially different genes as statistically different, the fifth gene was identified only through the addition of random noise. While it is important to note that the proposed confidence interval procedure relies on a fixed structure of the given network, any relevant base network can be used. In this application, the base network is a gene regulatory network previously constructed by Margolin et al. (2006) from the human dataset. The results of the proposed framework, assessed with two separate edge-weight perturbation methods, suggest a general robustness of betweenness centrality when used as a method for identifying genes essential to the structure of diseased tissue. Yet, a question that will be addressed in Chapter 6 is whether the variability in the betweenness sampling distributions depicted in Figures 3.5 and 3.6 is a result of the glioma dataset, or more suggestive of measure's inherent variability.

In previous work, Pan et al. (2018) uses thresholding to find *essential* genes (i.e., genes that are *essentially different* in comparing networks derived from diseased samples as compared to networks from healthy samples) which were then experimentally validated. The proposed confidence interval procedure in Algorithm 4 provides an alternative way for

validating the set of essentially different genes and allows for computational validation in situations where experiments are not feasible. In particular, genes that are both essentially different and significantly different across the two networks are identified using the aforementioned perturbation methods. Furthermore, the proposed confidence interval methodology has applications to any fixed network structure with edge-weights that vary across conditions. In some situations, the genes whose centrality is repressed in the diseased group may need to be identified. This can be accomplished using the proposed methodology, with the roles of $C_{B_{\text{Tumor}}}$ and $C_{B_{\text{Normal}}}$ reversed. Finally, the official publication of the work detailed in this chapter may be found in Durón et al. (2018).

CHAPTER 4

THE STABLE BETWEENNESS CENTRALITY MEASURE

4.1 INTRODUCTION

As centrality measures are utilized in the analysis of biological networks to identify genes critical to the functioning of diseases, with some examples including Breitkreutz et al. (2012); Mistry et al. (2017); Ramadan et al. (2016); Zhang et al. (2013), it is important to consider the natural variability of these measures associated with experimental data. Recall the **betweenness** centrality of node $v_i \in V$:

$$C_B(v_i) \quad = \sum_{j \neq k \neq i} \frac{\sigma_{v_j,v_k}(v_i)}{\sigma_{v_j,v_k}}$$

where $v_j, v_k \in V$, σ_{v_j,v_k} is the total number of shortest paths from node v_j to node v_k, and $\sigma_{v_j,v_k}(v_i)$ is the number of those paths that pass through v_i.

A centrality measure is considered stable, according to Segarra and Ribeiro (2016), based upon its ability " to be robust to noise in the network data". According to their definition, how stable is the betweenness centrality? Would a small perturbation to the edge-weights in a network result in a "large" change to the betweenness centrality of a node?

Although the variability of the betweenness measure was briefly discussed in Section 3.2, to begin a more detailed examination of the measure's sensitivity to edge-weight perturbations, consider the weighted and undirected network $G = (V, E)$ in Figure 4.1 with nodes $V = \{1, 2, 3, 4, 5, 6\}$ and edges $E = \{e_{1,2}, e_{1,3}, e_{1,4}, e_{1,5}, e_{1,6}, e_{2,3}, e_{3,4}, e_{4,5}\}$ with corresponding edge-weights $\{1, 1-\epsilon, 1-\epsilon, 1, 1, 1, 1, 1\}$ for $\epsilon = 0.0001$. The betweenness centrality of each

node in network G is listed in Table 4.1. Because node 1 is part of the unique shortest path between node pairs $\{(2,4),(2,5),(2,6),(3,5),(3,6),(4,6),(5,6)\}$, its betweenness centrality is $C_B^G(1) = 7$.

To highlight the variability of the betweenness measure, assign edges $e_{1,3}$ and $e_{1,4}$ to have weights both equal to 1.0001. In this instance, instead of subtracting ϵ from 1 as in network G, $\epsilon = 0.0001$ is now added to 1. This perturbed network, G', depicted in Figure 4.2, is identical in structure to network G with the exception of the two increased weights on edges $e_{1,3}$ and $e_{1,4}$. The betweenness centrality of each node in network G' is listed in Table 4.2. Because node 1 is no longer an intermediate node on the shortest path between node pairs $\{(2,4),(3,5)\}$, its betweenness centrality decreases to $C_B^{G'}(1) = 5$.

If the betweenness centrality measure is used in the analysis of networks to identify drug-targets, the sensitivity of this centrality measure to small perturbations in edge-weights is concerning. This motivates the need for an alternative definition of the betweenness centrality to be a measure that is robust to edge-weight perturbations.

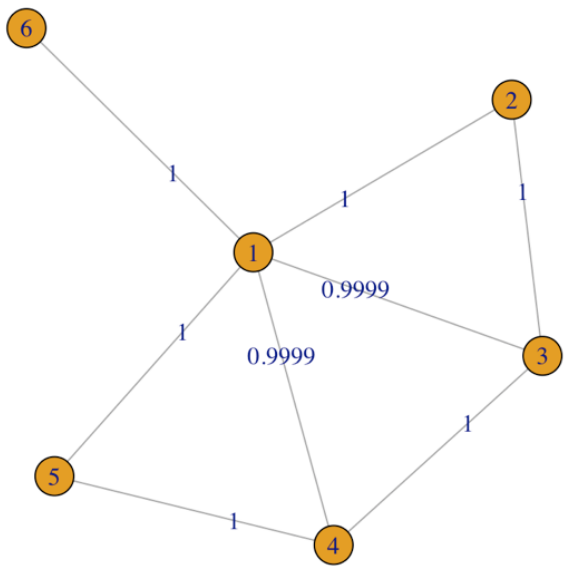

Figure 4.1. Let G be a weighted and undirected network with nodes $V = \{1, 2, 3, 4, 5, 6\}$ and edges $E = \{e_{1,2}, e_{1,3}, e_{1,4}, e_{1,5}, e_{1,6}, e_{2,3}, e_{3,4}, e_{4,5}\}$ with corresponding edge-weights $\{1, 0.9999, 0.9999, 1, 1, 1, 1, 1\}$.

Table 4.1. The betweenness value of each node in the weighted and undirected network G in Figure 4.1.

Node v_i	$C_B^G(v_i)$
1	7.0
2	0.0
3	0.0
4	0.0
5	0.0
6	0.0

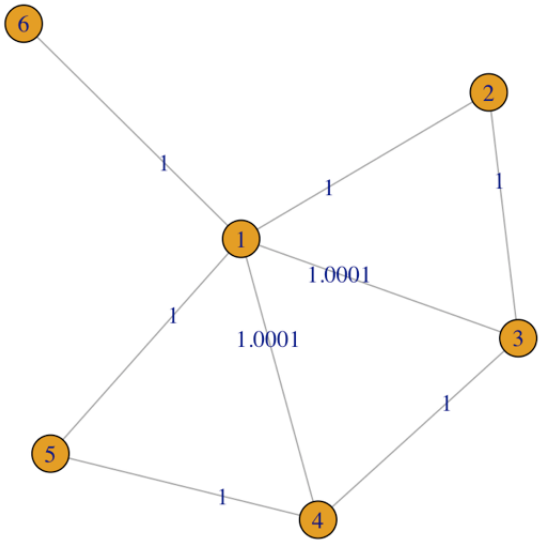

Figure 4.2. Let G' be an weighted and undirected network with nodes $V = \{1, 2, 3, 4, 5, 6\}$ and edges $E = \{e_{1,2}, e_{1,3}, e_{1,4}, e_{1,5}, e_{1,6}, e_{2,3}, e_{3,4}, e_{4,5}\}$ with corresponding edge-weights $\{1, 1.0001, 1.0001, 1, 1, 1, 1, 1\}$.

Table 4.2. The betweenness value of each node in the weighted and undirected network G' in Figure 4.2.

Node v_i	$C_B^{G'}(v_i)$
1	5.0
2	0.0
3	1.0
4	1.0
5	0.0
6	0.0

4.2 STABLE BETWEENNESS CENTRALITY MEASURE

4.2.1 Stability Definition

Define $\mathcal{X}$ as the space of networks on N nodes.

Definition. A centrality measure $C : V \to \mathbb{R}$ is **stable**, as defined by Segarra and Ribeiro (2016), if for every node set V and any two networks with N nodes, $M, M' \in \mathcal{X}$, there exists a finite, nonzero constant K such that for every node $v_i \in V$

$$|C^M(v_i) - C^{M'}(v_i)| \leq Kd(M, M') \tag{4.1}$$

where $C^M(v_i)$ and $C^{M'}(v_i)$ are the centrality values of node v_i in networks M and M', respectively, and $d : V^2 \times V^2 \to \mathbb{R}_+$ is the distance metric defined as

$$d(M, M') = \sum_{j,k \in |V|} |w_M(e_{j,k}) - w_{M'}(e_{j,k})| \tag{4.2}$$

for weights w_M and $w_{M'}$ on the edge $e_{j,k}$ between nodes v_j and v_k in networks M and M', respectively.

The stability definition in Equation (4.1) states that a centrality measure is stable if the difference in centrality values for a given node in two different networks is bounded by a constant K times the distance between these networks.

Recall network G and G' in Figures 4.1 and 4.2. The distance between networks G

and G' for node 1 equals 4ϵ:

$$\begin{aligned} d(G, G') &= \sum_{j,k \in |V|} |w_G(e_{j,k}) - w_{G'}(e_{j,k})| \\ &= 0 + 0.0002 + 0.0002 + 0 + 0 + 0 + 0 + 0 \\ &= 0.0004 \\ d(G, G') &= 4\epsilon \end{aligned}$$

whereas the difference in centrality values for node 1 equals 2:

$$|C_B^G(1) - C_B^{G'}(1)| = 7 - 5 = 2$$

For any constant K in Equation (4.1), there will always exist a small enough $\epsilon > 0$ such that the ratio

$$\frac{|C_B^G(1) - C_B^{G'}(1)|}{d(G, G')} = \frac{2}{4\epsilon}$$

will be larger than the proposed K. Therefore, a constant K does not exist and the betweenness centrality C_B, as defined in Equation (1.1), is not stable. This instability of C_B motivates the definition of the stable betweenness centrality measure, C_{SB}, as constructed by Segarra and Ribeiro (2016).

4.2.2 Definition of Stable Betweenness

Given an arbitrary network M of N nodes with weight function w_M and a node $v_i \in V$, define network M^{-i} to be identical in structure to network M, with the exception that the

edges containing node v_i are eliminated. Refer to Chapter 1 for the definitions of a shortest path $P_s(v_j, v_k)$ and its length $s(v_j, v_k)$.

Definition. The **stable betweenness** centrality of node $v_i \in V$, as defined by Segarra and Ribeiro (2016), is given by :

$$C_{SB}(v_i) \quad = \sum_{j \neq k \neq i} s_{M^{-i}}(v_j, v_k) - s_M(v_j, v_k) \tag{4.3}$$

where $v_i, v_j, v_k \in V$, and $s_M(v_j, v_k), s_{M^{-i}}(v_j, v_k)$ are the lengths of the shortest path(s) between nodes v_j and v_k in networks M, M^{-i} respectively.

While both the betweenness and stable betweenness centrality measures place importance on whether shortest paths contain a node, the stable betweenness measure also considers the impact of the node on the lengths of the shortest paths. In particular, the stable betweenness definition states that, if the shortest path between nodes v_j and v_k pass through node v_i, then removing node v_i would necessarily result in a longer "shortest" path between nodes v_j and v_k. But the power of the definition of this centrality measure C_{SB} is its stability, as defined in Equation (4.1), where K can be shown to be at most $2(N-1)(N-2)$. For a formal proof of the stability of C_{SB}, refer to Segarra and Ribeiro (2016).

4.3 APPLICATION OF STABLE BETWEENNESS CENTRALITY MEASURE

The length of all the shortest paths between each pair of nodes in the weighted and undirected network G from Figure 4.1 is below, where the ij-th entry in the symmetric

matrix corresponds to the length of the shortest path.

$$\begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0.0000 & 1.0000 & 0.9999 & 0.9999 & 1.0000 & 1.0000 \\ 1.0000 & 0.0000 & 1.0000 & 1.9999 & 2.0000 & 2.0000 \\ 0.9999 & 1.0000 & 0.0000 & 1.0000 & 1.9999 & 1.9999 \\ 0.9999 & 1.9999 & 1.0000 & 0.0000 & 1.0000 & 1.9999 \\ 1.0000 & 2.0000 & 1.9999 & 1.0000 & 0.0000 & 2.0000 \\ 1.0000 & 2.0000 & 1.9999 & 1.9999 & 2.0000 & 0.0000 \end{pmatrix} \end{array} \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array}$$

A new network G^{-1}, constructed by removing the edges containing node 1 from network G, is depicted in Figure 4.3. The length of all the shortest paths between each pair of nodes in network G^{-1} is provided in the symmetric matrix below:

$$\begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0 & \infty & \infty & \infty & \infty & \infty \\ \infty & 0 & 1 & 2 & 3 & \infty \\ \infty & 1 & 0 & 1 & 2 & \infty \\ \infty & 2 & 1 & 0 & 1 & \infty \\ \infty & 3 & 2 & 1 & 0 & \infty \\ \infty & \infty & \infty & \infty & \infty & \infty \end{pmatrix} \end{array} \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array}$$

The removal of the edges containing node 1 results in the disconnected network G^{-1},

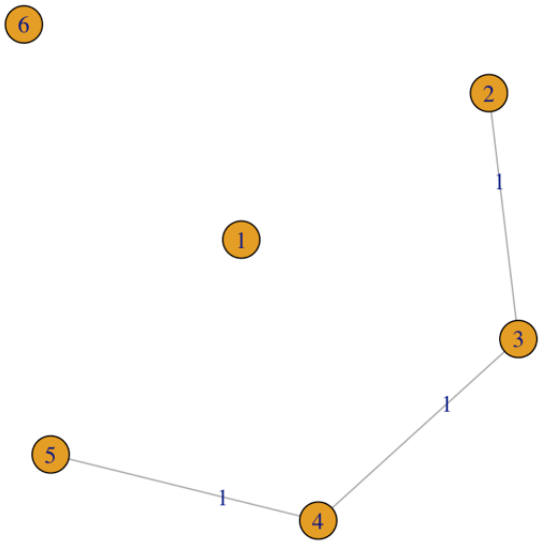

Figure 4.3. Let G^{-1} be the network with nodes $V = \{1, 2, 3, 4, 5, 6\}$ and edges $E = \{e_{2,3}, e_{3,4}, e_{4,5}\}$ with corresponding edge-weights $\{1, 1, 1\}$.

causing the lengths of shortest paths between certain node pairs to be infinitely long, an assignment that follows what is typically done in practice with shortest path calculation algorithms like Dijkstra (1959). The calculation of the stable betweenness centrality value for node 1 cannot be computed according to Equation (4.3) due to these infinity long paths, motivating a need to extend Segarra's stable betweenness centrality measure to one that is finitely defined for both connected and disconnected networks.

4.4 FINITE STABLE BETWEENNESS CENTRALITY MEASURE

Definition. The **finite stable betweenness** centrality of node $v_i \in V$ in a connected network M is given by:

$$C_{FSB}(v_i) = \sum_{j \neq i \neq k} s_{M^{-i}}(v_j, v_k) - s_M(v_j, v_k) \tag{4.4}$$

where

$$s_{M^{-i}}(v_j, v_k) \;=\; \begin{cases} s_{M^{-i}}(v_j, v_k), & \text{if } v_j \text{ and } v_k \text{ are connected in } M^{-i} \\ \sum_{j,k \in |V|} w_M(e_{j,k}), & \text{if } v_j \text{ and } v_k \text{ are disconnected in } M^{-i} \end{cases} \tag{4.5}$$

where $v_i, v_j, v_k \in V$, and $s_M(v_j, v_k), s_{M^{-i}}(v_j, v_k)$ are the lengths of the shortest path(s) between nodes v_j and v_k in networks M, M^{-i} respectively.

The finite stable betweenness definition follows the same idea of the stable betweenness definition (Segarra and Ribeiro, 2016) in that if the shortest path between nodes v_j and v_k pass through node v_i, then removing node v_i results in a longer "shortest" path between nodes v_j and v_k. But unlike the definition of Segarra and Ribeiro (2016), the finite stable betweenness measure is now defined for networks that become disconnected in the process of computing the stable betweenness. Provided that network M is connected, then following the definition of finite stable betweenness, should a path between two nodes in network M^{-i} not exist, thereby implying that network M^{-i} is disconnected, then the length of the "shortest" path equals the sum of all the edge-weights in network M. Furthermore, the finite stable betweenness centrality measure C_{FSB} is stable as defined in Equation (4.1), where K can be shown to be at most $2(N-1)(N-2)$.

4.5 PROOF OF STABILITY FOR FINITE STABLE BETWEENNESS CENTRALITY MEASURE

Following the proof by Segarra and Ribeiro (2016), given a node set V with $|V| = N$ nodes, a finite, nonzero constant K must be found such that for any two connected networks

$A = (V, E)$ and $B = (V, E)$ and weight functions w_A and w_B, respectively, for every node $v_i \in V$

$$|C^A_{FSB}(v_i) - C^B_{FSB}(v_i)| \quad \leq Kd(A, B) \tag{4.6}$$

To begin, suppose networks A^{-i} and B^{-i} are created by eliminating the edges containing node v_i in networks A and B, respectively. From Equation (4.4) in the definition of C_{FSB}:

$$|C^A_{FSB}(v_i) - C^B_{FSB}(v_i)| = \\ |\sum_{j\neq i\neq k} s_{A^{-i}}(v_j, v_k) - s_A(v_j, v_k) - \sum_{j\neq i\neq k} s_{B^{-i}}(v_j, v_k) - s_B(v_j, v_k)| \tag{4.7}$$

Re-arranging the terms and applying the triangle inequality, Equation (4.7) can be simplified as

$$\begin{aligned} |C^A_{FSB}(v_i) - C^B_{FSB}(v_i)| &= |\sum_{j\neq i\neq k} s_{A^{-i}}(v_j, v_k) - s_{B^{-i}}(v_j, v_k) + \sum_{j\neq i\neq k} s_B(v_j, v_k) - s_A(v_j, v_k)| \\ &\leq |\sum_{j\neq i\neq k} s_{A^{-i}}(v_j, v_k) - s_{B^{-i}}(v_j, v_k)| + |\sum_{j\neq i\neq k} s_B(v_j, v_k) - s_A(v_j, v_k)| \\ |C^A_{FSB}(v_i) - C^B_{FSB}(v_i)| &\leq \sum_{j\neq i\neq k} |s_{A^{-i}}(v_j, v_k) - s_{B^{-i}}(v_j, v_k)| + |s_B(v_j, v_k) - s_A(v_j, v_k)| \end{aligned} \tag{4.8}$$

Suppose that the shortest path between nodes v_j and v_k in network A is given by the path $P^A_s(v_j, v_k) = \{e_{j=j_0,j_1}, e_{j_1,j_2}, ..., e_{j_{n-1},j_n=k}\}$, while the shortest path between nodes v_j and v_k in network B is given by the path $P^B_s(v_j, v_k) = \{e_{j=k_0,k_1}, e_{k_1,k_2}, ..., e_{k_{m-1},k_m=k}\}$.

Without loss of generality, assume that the length of the shortest path in network A is larger than that in network B such that $s_A(v_j, v_k) \geq s_B(v_j, v_k)$. From the definition of the length of a shortest path provided in Chapter 1, consider the difference between the lengths of the shortest paths in networks A and B:

$$|s_A(v_j, v_k) - s_B(v_j, v_k)| = |\sum_{i=0}^{n-1} w_A(e_{j_i,j_{i+1}}) - \sum_{i=0}^{m-1} w_B(e_{k_i,k_{i+1}})|$$

Then the difference between the lengths of the shortest paths in networks A and B is bounded by the distance between those networks:

$$|s_A(v_j, v_k) - s_B(v_j, v_k)| \leq |\sum_{i=0}^{m-1} w_A(e_{k_i,k_{i+1}}) - \sum_{i=0}^{m-1} w_B(e_{k_i,k_{i+1}})| \tag{4.9}$$

Replacing the shortest path in network A with any other path, as is done in Equation (4.9), causes the length of the path between nodes v_j and v_k in network A to become larger. With a bit more algebra, the distance between the networks is shown to bound the distance between the shortest paths:

$$\begin{aligned}
|s_A(v_j, v_k) - s_B(v_j, v_k)| &\leq |\sum_{i=0}^{m-1} w_A(e_{k_i,k_{i+1}}) - \sum_{i=0}^{m-1} w_B(e_{k_i,k_{i+1}})| \\
&= |\sum_{i=0}^{m-1} w_A(e_{k_i,k_{i+1}}) - w_B(e_{k_i,k_{i+1}})| \\
&\leq \sum_{i=0}^{m-1} |w_A(e_{k_i,k_{i+1}}) - w_B(e_{k_i,k_{i+1}})| \\
&\leq \sum_{j,k \in |V|} |w_A(e_{j,k}) - w_B(e_{j,k})| \\
|s_A(v_j, v_k) - s_B(v_j, v_k)| &\leq d(A, B)
\end{aligned} \tag{4.10}$$

Now suppose that networks A^{-i} and B^{-i} are each connected networks such that a path exists between nodes v_j and v_k. Applying the result in Equation (4.10) to these networks, it must hold that

$$|s_{A^{-i}}(v_j, v_k) - s_{B^{-i}}(v_j, v_k)| \;\leq\; d(A^{-i}, B^{-i}) \tag{4.11}$$

Furthermore, by construction of networks A^{-i} and B^{-i}, then the distance between networks A and B is necessarily greater than the distance between networks A^{-i} and B^{-i}. To see this, consider the definition of the distance metric in Equation (4.2):

$$\begin{aligned} d(A,B) &= \sum_{j,k\in|V|} |w_A(e_{j,k}) - w_B(e_{j,k})| \\ &= \sum_{\substack{j,k\in|V| \\ j\neq i\neq k}} |w_A(e_{j,k}) - w_B(e_{j,k})| + \sum_{\substack{j,k\in|V| \\ i=j \text{ or } i=k}} |w_A(e_{j,k}) - w_B(e_{j,k})| \\ d(A,B) &= d(A^{-i}, B^{-i}) + \sum_{\substack{j,k\in|V| \\ i=j \text{ or } i=k}} |w_A(e_{j,k}) - w_B(e_{j,k})| \end{aligned}$$

which implies that

$$d(A^{-i}, B^{-i}) \;\leq\; d(A,B) \tag{4.12}$$

Upon substitution of Equation (4.11) into Equation (4.12), then

$$|s_{A^{-i}}(v_j, v_k) - s_{B^{-i}}(v_j, v_k)| \;\leq\; d(A,B)$$

Now suppose that nodes v_j and v_k are disconnected such that a path between the

node pair does not exist in networks A^{-i} and B^{-i}, respectively. From Equation (4.5) in the definition of C_{FSB}, then

$$\begin{aligned}
|s_{A^{-i}}(v_j, v_k) - s_{B^{-i}}(v_j, v_k)| &= |\sum_{j,k \in |V|} w_A(e_{j,k}) - \sum_{j,k \in |V|} w_B(e_{j,k})| \\
&= |\sum_{j,k \in |V|} w_A(e_{j,k}) - w_B(e_{j,k})| \\
&\leq \sum_{j,k \in |V|} |w_A(e_{j,k}) - w_B(e_{j,k})| \\
|s_{A^{-i}}(v_j, v_k) - s_{B^{-i}}(v_j, v_k)| &\leq d(A, B)
\end{aligned}$$

Thus, regardless of whether networks nodes v_j and v_k are connected or disconnected in networks A^{-i} and B^{-i}, it is always true that

$$|s_{A^{-i}}(v_j, v_k) - s_{B^{-i}}(v_j, v_k)| \leq d(A, B) \tag{4.13}$$

Upon substitution of Equations (4.10) and (4.13), Equation (4.8) can be simplified:

$$\begin{aligned}
|C^A_{FSB}(v_i) - C^B_{FSB}(v_i)| &\leq \sum_{j \neq i \neq k} |s_{A^{-i}}(v_j, v_k) - s_{B^{-i}}(v_j, v_k)| + |s_B(v_j, v_k) - s_A(v_j, v_k)| \\
&\leq \sum_{j \neq i \neq k} d(A, B) + d(A, B) \\
&= (N-1)(N-2)\left[d(A, B) + d(A, B)\right] \\
|C^A_{FSB}(v_i) - C^B_{FSB}(v_i)| &\leq 2(N-1)(N-2)d(A, B)
\end{aligned} \tag{4.14}$$

Therefore, setting $K = 2(N-1)(N-2)$ in Equation (4.14), the expression in Equation (4.1) is obtained, proving the stability of the finite stable betweenness centrality measure, C_{FSB} as defined in Equations (4.4) and (4.5).

4.6 APPLICATION OF FINITE STABLE BETWEENNESS CENTRALITY MEASURE

Recall the weighted, undirected network G in Figure 4.1. For each node in network G, the finite stable betweenness centrality value is calculated according to Equations (4.4) and (4.5), and is listed in Table 4.3. Using C_{FSB}, the finite stable betweenness of node 1 is $C^G_{FSB}(1) = 49.9992$.

Table 4.3. The finite stable betweenness value of each node in the weighted and undirected network G in Figure 4.1.

Node v_i	$C^G_{FSB}(v_i)$
1	49.9992
2	0
3	0
4	0
5	0
6	0

Recall that network G', depicted in Figure 4.2, is structurally identical to network G with the exception of two perturbed edge-weights. The C_{FSB} for each node in network G' is listed Table 4.4. In particular, note that $C^{G'}_{FSB}(1) = 50.0012$.

Table 4.4. The finite stable betweenness value of each node in the weighted and undirected network G' in Figure 4.2.

Node v_i	$C_{FSB}^{G'}(v_i)$
1	50.0012
2	0
3	0.0002
4	0.0002
5	0
6	0

Although perturbations of magnitude $\epsilon = 0.0001$ are made to two edge-weights, node 1 maintains a centrality value in network G' approximately equal to its value in network G. This result suggests that the finite stable betweenness centrality is stable, and a more robust measure to use in the network analysis and identification of essential nodes given the inherent variability in data. Furthermore, with the definition of finite stable betweenness C_{FSB}, a finite centrality value is guaranteed for each node in a network, regardless of whether the removal of edges containing a particular node results in a connected or disconnected network.

CHAPTER 5

EXPONENTIAL RANDOM GRAPH MODELS

5.1 INTRODUCTION

An exponential-family random graph model (ERGM) is a general class of models based in exponential family theory that specify the probability distribution that underlies a set of random graphs or networks (Handcock et al., 2015). The ERGM models aim to understand the structure of an observed network through the inclusion of a finite number of network statistics that are selected to best capture the structural features of the network of interest. Although Table 5.1 lists commonly used unweighted and weighted network statistics for undirected networks, additional statistics for directed and undirected networks are specified elsewhere (Denny et al., 2017; Morris et al., 2008).

Recall that the betweenness centrality captures some global structural properties of a network. Although more details are provided in Chapter 6, the framework proposed to determine a distribution of finite stable betweenness (refer to Section 4.4) involves incorporating the measure into the probability model as a network statistic. But in order to infuse the measure into the model and properly interpret its distribution, the theoretical underpinnings of exponential random graph models for unweighted and weighted networks are detailed in the subsequent sections, with particular attention towards the derivation of each probability model, and coefficient estimation process for weighted networks.

Table 5.1. Example of ERGM unweighted and weighted network statistics for undirected networks that may be included in probability models.

Network Statistic	**Unweighted Description**	**Weighted Description**
Edges	The number of edges in the network	The sum of the edge-weights in the network
Nodes	The number of nodes in the network	Not currently implemented as a weighted statistic
Isolates	The number of nodes in the network with no edges	Not currently implemented as a weighted statistic
Triangle	The number of 3-cycles in the network defined as any edge set $\{e_{i,j}, e_{j,k}, e_{k,i}\}$	The sum of the edge-weight product for edge set $\{e_{i,j}, e_{j,k}, e_{k,i}\}$
2-Star	The number of nodes in the network with 2 edges	The sum of the edge-weight product for nodes with 2 edges
k-Star	The number of nodes in the network with k edges	Not currently implemented as a weighted statistic
k-Cycle	The number of k-cycles in the network	Not currently implemented as a weighted statistic
k-Degree	The number of nodes in the network with degree k	Not currently implemented as a weighted statistic

5.2 PROBABILITY MODEL FOR UNWEIGHTED NETWORK

Define the sample space $\mathcal{X}$ to be the set of all possible unweighted networks with N nodes. Let X be a random variable representing an unweighted network with N nodes. To estimate the probability model for random variable X as a function of the observed network x, define $\vec{S}(x) = [S_1(x), S_2(x), ..., S_n(x)]^T$ as a vector of network statistics, where each statistic can be any function on the observed network. The network statistics vector quantifies the number of local configurations within an unweighted network, as in Table 5.1.

Assume that there exists a vector of coefficient estimates $\vec{\theta} = [\theta_1, \theta_2, ..., \theta_n]^T \in \mathbb{R}^n$, such that

$$\log\left(P(X = x; \vec{\theta})\right) \propto \theta_1 S_1(x) + \theta_2 S_2(x) + ... + \theta_n S_n(x) = \quad \vec{\theta}^T \vec{S}(x)$$

Exponentiating both sides and dividing by a normalizing constant $\kappa(\vec{\theta}) = \sum_{z \in \mathcal{X}} \exp\{\vec{\theta}^T \vec{S}(z)\}$, the probability model for an weighted network is derived as:

$$P(X = x; \vec{\theta}) \quad = \quad \frac{\exp\{\vec{\theta}^T \vec{S}(x)\}}{\sum_{z \in \mathcal{X}} \exp\{\vec{\theta}^T \vec{S}(z)\}}, \quad x \in \{0, 1\}^{\binom{N}{2}} \tag{5.1}$$

Equation (5.1) is the general form of the ERGM and describes a probability distribution on $\mathcal{X}$, the space of all unweighted networks with N nodes. Because Equation (5.1) is based upon a vector of network configurations, $\vec{S}(x)$, the likelihood of observing each particular network depends upon the presence or absence of the particular network configurations.

5.3 PROBABILITY MODEL FOR WEIGHTED NETWORKS

The standard definition of an ERGM requires the edges of the observed network to be unweighted, thereby denoting the presence or absence of an edge. As a result, ERGMs, in the standard model definition, are unable to model weighted networks.

Two models have been proposed to model weighted networks of N nodes using the ERGM framework. While each model extends the methodology of ERGMs to weighted networks, the method of Krivitsky (2012) concerns networks with integer-valued edge-weights, while the networks in the approach taken by Desmarais and Cranmer (2012) have edge-weights that are continuous-valued. Although an overview of the derivation of the probability model proposed by Desmarais and Cranmer (2012) is provided, the estimation process

and the sampling of networks with weights in [0,1] using the Metropolis-Hastings algorithm are specifically detailed, as the framework proposed to determine a distribution of the betweenness centrality (provided in Section 6.4) utilizes networks solely from this space.

Following the work of Desmarais and Cranmer (2012), the probability distribution for GERGMs is specified by a joint probability density function $f_Y(y;\vec{\theta})$ that involves two steps: (1) specifying a joint distribution that captures the network statistics of interest on a restricted valued network X, and (2) transforming X onto the sample space of the observed network Y. Requiring networks to have weights in [0,1] in the first step of the derivation is needed in order to use the Gibbs sampling algorithm, which is one method for approximating the normalizing constant. In order to sample networks using the Gibbs sampling algorithm, networks are drawn from the conditional distribution through the inverse of the cumulative distribution function method (i.e., the inverse CDF). Additional details of the Gibbs sampling algorithm and its utilization in the estimation process for networks with continuous-valued edges are provided by Desmarais and Cranmer (2012).

Define the space of networks as $\mathcal{X} \in \{[0,1]^M\}$ to be the set of all possible networks with N nodes and M edges with weights in $[0,1]$. Denote x as the restricted valued network which has the same nodes as the observed weighted network y, but whose edge-weights are bounded and continuous between zero and one, $x_{j,k} \in [0,1]$. Adapting the ERGM probability model in Equation (5.1) to handle the continuous-valued edges, the probability density function for the restricted valued network X becomes

$$f_X(x;\vec{\theta}) \;=\; \frac{\exp\{\vec{\theta}^T\vec{S}(x)\}}{\int_{z\in\mathcal{X}}\exp\{\vec{\theta}^T\vec{S}(z)\}dz}, \quad x \in [0,1]^M \tag{5.2}$$

where $\vec{\theta} : \mathbb{R}^n$ is the vector of coefficients, and $\vec{S} : [0,1]^M \to \mathbb{R}^n$ is the vector of network statistics.

To transform the distribution of the restricted valued network X onto the sample space of the continuous-valued network Y, a parameterized, one-to-one, monotonically non-decreasing transformation $T^{-1}(\cdot)$ is applied to the edge $e_{j,k}$ between each pair of distinct nodes $v_j, v_k \in V$, such that

$$Y_{j,k} \;=\; T_{j,k}^{-1}(X_{j,k}; \lambda_{j,k}) \tag{5.3}$$

for unknown transformation coefficient $\vec{\lambda} \in \mathbb{R}^M$. Note that each edge $e_{j,k}$ of network y, denoted $y_{j,k}$, is now defined as a parameterized transformation of the same edge $x_{j,k}$ in the restricted valued network x.

Since the transformation $Y_{j,k} = T_{j,k}^{-1}(X; \lambda)$ is one-to-one for every node pair $v_j, v_k \in V$, the inverse function exists and is defined as $X_{j,k} = T_{j,k}(Y; \vec{\beta})$ for some transformation coefficient $\vec{\beta} \in \mathbb{R}^M$.

Denote J as the Jacobian of the inverse function, $X = T(Y, \vec{\beta})$:

$$J \;=\; \begin{bmatrix} \frac{\partial X_{1,1}}{\partial Y_{1,1}} & \cdots & \frac{\partial X_{1,1}}{\partial Y_{j,j}} & \cdots & \frac{\partial X_{1,1}}{\partial Y_{N,N}} \\ \frac{\partial X_{1,2}}{\partial Y_{1,1}} & \cdots & \frac{\partial X_{1,2}}{\partial Y_{j,j}} & \cdots & \frac{\partial X_{1,2}}{\partial Y_{N,N}} \\ \vdots & \ddots & \vdots & \cdots & \vdots \\ \frac{\partial X_{j,j}}{\partial Y_{1,1}} & \cdots & \frac{\partial X_{j,j}}{\partial Y_{j,j}} & \cdots & \frac{\partial X_{j,j}}{\partial Y_{N,N}} \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ \frac{\partial X_{N,N}}{\partial Y_{1,1}} & \cdots & \frac{\partial X_{N,N}}{\partial Y_{j,j}} & \cdots & \frac{\partial X_{N,N}}{\partial Y_{N,N}} \end{bmatrix} = \begin{bmatrix} \frac{\partial T_{1,1}(Y;\vec{\beta})}{\partial Y_{1,1}} & \cdots & \frac{\partial T_{1,1}(Y;\vec{\beta})}{\partial Y_{j,j}} & \cdots & \frac{\partial T_{1,1}(Y;\vec{\beta})}{\partial Y_{N,N}} \\ \frac{\partial T_{1,2}(Y;\vec{\beta})}{\partial Y_{1,1}} & \cdots & \frac{\partial T_{1,2}(Y;\vec{\beta})}{\partial Y_{j,j}} & \cdots & \frac{\partial T_{1,2}(Y;\vec{\beta})}{\partial Y_{N,N}} \\ \vdots & \ddots & \vdots & \cdots & \vdots \\ \frac{\partial T_{j,j}(Y;\vec{\beta})}{\partial Y_{1,1}} & \cdots & \frac{\partial T_{j,j}(Y;\vec{\beta})}{\partial Y_{j,j}} & \cdots & \frac{\partial T_{j,j}(Y;\vec{\beta})}{\partial Y_{N,N}} \\ \vdots & \cdots & \vdots & \ddots & \vdots \\ \frac{\partial T_{N,N}(Y;\vec{\beta})}{\partial Y_{1,1}} & \cdots & \frac{\partial T_{N,N}(Y;\vec{\beta})}{\partial Y_{j,j}} & \cdots & \frac{\partial T_{N,N}(Y;\vec{\beta})}{\partial Y_{N,N}} \end{bmatrix}$$

Because T^{-1} is a monotone, non-decreasing, and one-to-one transformation, the joint

probability density function of Y can be expressed as

$$f_Y(y;\vec{\theta},\vec{\beta}) \;=\; f_X(x;\vec{\theta})\cdot|J| = \quad f_X\left(T(y;\vec{\beta});\vec{\theta}\right)\cdot|J| \tag{5.4}$$

where $|J|$ is the absolute value of the determinant of J.

As the transformation is defined for each distinct node pair $v_j, v_k \in V$, the Jacobian J can be simplified to a diagonal matrix,

$$J \;=\; \begin{bmatrix} \frac{\partial T_{1,1}(Y;\vec{\beta})}{\partial Y_{1,1}} & \dots & 0 & \dots & 0 \\ 0 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \dots & \vdots \\ 0 & \dots & \frac{\partial T_{j,j}(Y;\vec{\beta})}{\partial Y_{j,j}} & \dots & 0 \\ \vdots & \dots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & \frac{\partial T_{N,N}(Y;\vec{\beta})}{\partial Y_{N,N}} \end{bmatrix} \tag{5.5}$$

so that its determinant is the product of its diagonal entries,

$$|J| \;=\; \prod_{j=1}^{N} \frac{\partial T_{j,j}(y;\vec{\beta})}{\partial Y_{j,j}} \tag{5.6}$$

Therefore, using Equations (5.2), (5.4), and (5.6), the probability model for the GERGM of a weighted network with continuous-valued edges can be written as

$$f_Y(y;\vec{\theta},\vec{\beta}) \;=\; \frac{\exp\{\vec{\theta}^T\vec{S}(T(y;\vec{\beta}))\}}{\int_{z\in\mathcal{X}}\exp\{\vec{\theta}^T\vec{S}(z)\}dz}\prod_j t_j(y;\vec{\beta}) \tag{5.7}$$

where $t_j(y;\vec{\beta}) = \frac{\partial T_{j,j}(y;\vec{\beta})}{\partial y_{j,j}}$. The vector of network statistics $\vec{S}(\cdot)$ in Equation (5.7) is now specified on a transformation of the network, rather than on the observed network.

As the framework proposed to determine a distribution on finite stable betweenness (discussed in Chapter 6) concerns the sample space of all networks of N nodes and M edges with weights in the interval [0,1], the estimation process of the probability model

$$f_X(x;\vec{\theta}) = \frac{\exp\{\vec{\theta}^T\vec{S}(x)\}}{\int_{z\in\mathcal{X}}\exp\{\vec{\theta}^T\vec{S}(z)\}dz}, \quad x\in[0,1]^M$$

is discussed below.

5.3.1 Maximum Likelihood Estimation

Given the set of network statistics $\vec{S}(\cdot)$ and an observed network x_{obs}, the value of the unknown coefficient $\hat{\vec{\theta}}$ that maximizes the likelihood of x_{obs} can be determined. From the probability model $f_X(x_{\text{obs}};\vec{\theta})$ in Equation (5.2), define the likelihood function $L(\vec{\theta}|x_{\text{obs}})$ as

$$L(\vec{\theta}|x_{\text{obs}}) = \frac{\exp\{\vec{\theta}^T\vec{S}(x_{\text{obs}})\}}{\int_{z\in\mathcal{X}}\exp\{\vec{\theta}^T\vec{S}(z)\}dz}$$

Since the logarithm is a monotonically-increasing function of its argument, the value of the function that maximizes the log-likelihood is the same as the value that maximizes the likelihood. To simplify the expression for L, it is preferable to estimate the coefficient that maximizes the likelihood by maximizing the log-likelihood function,

$$\ell(\vec{\theta}|x_{\text{obs}}) = \log(L(\vec{\theta}|x_{\text{obs}})) = \vec{\theta}^T\vec{S}(x_{\text{obs}}) - \log\kappa(\vec{\theta}) \tag{5.8}$$

where $\kappa(\vec{\theta}) = \int_{z \in \mathcal{X}} \exp\{\vec{\theta}^T \vec{S}(z)\} dz$.

Direct maximization of $\ell(\vec{\theta}|x_{\text{obs}})$ provided in Equation (5.8) is computationally difficult due to the intractability of the normalizing constant $\kappa(\vec{\theta})$. A way around the direct maximization of $\ell(\vec{\theta}|x_{\text{obs}})$ is to, instead, fix an arbitrary coefficient $\tilde{\vec{\theta}} \in \mathbb{R}^n$ and maximize the difference of log-likelihoods $\ell(\vec{\theta}|x_{\text{obs}}) - \ell(\tilde{\vec{\theta}}|x_{\text{obs}})$:

$$\ell(\vec{\theta}|x_{\text{obs}}) - \ell(\tilde{\vec{\theta}}|x_{\text{obs}}) \quad = \quad \left(\vec{\theta} - \tilde{\vec{\theta}}\right)^T \vec{S}(x_{\text{obs}}) - \log\left(\frac{\kappa(\vec{\theta})}{\kappa(\tilde{\vec{\theta}})}\right) \tag{5.9}$$

Because the difference of log-likelihoods $\ell(\vec{\theta}|x_{\text{obs}}) - \ell(\tilde{\vec{\theta}}|x_{\text{obs}})$ is equivalent to the difference of the log-likelihood $\ell(\vec{\theta}|x_{\text{obs}})$ minus a constant, the direct maximization of Equation (5.9) yields the same maximum likelihood estimate of $\vec{\theta}$, denoted as $\hat{\vec{\theta}}$, that results from the maximization of Equation (5.8).

The ratio of normalizing constants may be re-expressed as an expected value,

$$\begin{aligned}
\frac{\kappa(\vec{\theta})}{\kappa(\tilde{\vec{\theta}})} &= \frac{\int_{z \in \mathcal{X}} \exp\{\vec{\theta}^T \vec{S}(z)\} dz}{\kappa(\tilde{\vec{\theta}})} \\
&= \frac{\int_{z \in \mathcal{X}} \exp\{\left(\vec{\theta} - \tilde{\vec{\theta}}\right)^T \vec{S}(z)\} \exp\{\tilde{\vec{\theta}}^T \vec{S}(z)\} dz}{\kappa(\tilde{\vec{\theta}})} \\
&= \int_{z \in \mathcal{X}} \exp\{\left(\vec{\theta} - \tilde{\vec{\theta}}\right)^T \vec{S}(z)\} \left(\frac{\exp\{\tilde{\vec{\theta}}^T \vec{S}(z)\}}{\kappa(\tilde{\vec{\theta}})}\right) dz \\
\frac{\kappa(\vec{\theta})}{\kappa(\tilde{\vec{\theta}})} &= E_{\tilde{\vec{\theta}}}\left[\exp\{\left(\vec{\theta} - \tilde{\vec{\theta}}\right)^T \vec{S}(X)\}\right]
\end{aligned} \tag{5.10}$$

and substitution of Equation (5.10) into Equation (5.9) allows the difference of log-likelihoods

$\ell(\vec{\theta}|x_{\text{obs}}) - \ell(\tilde{\vec{\theta}}|x_{\text{obs}})$ to be re-expressed:

$$\ell(\vec{\theta}|x_{\text{obs}}) - \ell(\tilde{\vec{\theta}}|x_{\text{obs}}) \;=\; \left(\vec{\theta} - \tilde{\vec{\theta}}\right)^T \vec{S}(x_{\text{obs}}) - \log\left(E_{\tilde{\vec{\theta}}}\left[\exp\{\left(\vec{\theta} - \tilde{\vec{\theta}}\right)^T \vec{S}(X)\}\right]\right) \quad (5.11)$$

In order to maximize Equation (5.11), a Markov Chain Monte Carlo (MCMC) method is used to approximate $E_{\tilde{\vec{\theta}}}\left[\exp\{\left(\vec{\theta} - \tilde{\vec{\theta}}\right)^T \vec{S}(X)\}\right]$. A MCMC procedure, such as the Metropolis-Hastings algorithm, generates a sample of K networks $x_1, x_2, ..., x_K$ from the probability distribution $f_X(x; \tilde{\vec{\theta}})$ such that

$$E_{\tilde{\vec{\theta}}}\left[\exp\{\left(\vec{\theta} - \tilde{\vec{\theta}}\right)^T \vec{S}(X)\}\right] \;\approx\; \frac{1}{K}\sum_{k=1}^{K} \exp\{\left(\vec{\theta} - \tilde{\vec{\theta}}\right)^T \vec{S}(x_k)\} \quad (5.12)$$

Exploiting the Law of Large Numbers to approximate the expected value, the maximum likelihood estimate $\hat{\vec{\theta}}$ can be determined from the maximization of $\ell(\vec{\theta}|x_{\text{obs}}) - \ell(\tilde{\vec{\theta}}|x_{\text{obs}})$,

$$\hat{\vec{\theta}} \;=\; \text{argmax}_{\vec{\theta}}\left[\left(\vec{\theta} - \tilde{\vec{\theta}}\right)^T \vec{S}(x_{\text{obs}}) - \log\left(\frac{1}{K}\sum_{k=1}^{K} \exp\{\left(\vec{\theta} - \tilde{\vec{\theta}}\right)^T \vec{S}(x_k)\}\right)\right] \quad (5.13)$$

given an arbitrary coefficient $\tilde{\vec{\theta}}$.

Stepping Algorithm

To obtain an accurate approximation of the log-likelihood function, the MCMC sampling algorithm must sample networks from a region where the probability mass is concentrated (Handcock et al., 2003). Because the mass is concentrated to a small region, the MCMC sampling algorithm must begin with a coefficient very close to the MLE $\hat{\vec{\theta}}$. In order to find the best estimate for $\hat{\vec{\theta}}$ such that $\tilde{\vec{\theta}}$ is "close" to $\hat{\vec{\theta}}$ (Geyer, 1992), Hummel (2012) intro-

duces a systematic method for moving $\tilde{\vec{\theta}}$ closer to $\hat{\vec{\theta}}$ to obtain a more accurate approximation of the log-likelihood function.

In exponential random graph models, the MLE, if it exists, is the coefficient $\hat{\vec{\theta}}$ such that $E_{\vec{\theta}}\left[\vec{S}(X)\right] = \vec{S}(x_{\text{obs}})$ (Hummel, 2012). As a result, the proposed algorithm of Hummel (2012) suggests taking steps toward the vector of observed network statistics $\vec{S}(x_{\text{obs}})$ by assigning an intermediate point $\hat{\zeta}_t$ the role of $\vec{S}(x_{\text{obs}})$, thereby allowing the search for $\tilde{\vec{\theta}}$ to reside in a region where the approximation of the log-likelihood function is reasonably accurate.

The stepping algorithm proposed by Hummel (2012) is detailed in Algorithm 5. If $\hat{\zeta}_t$ is in the interior of the convex hull of $\vec{S}(x_1), ..., \vec{S}(x_K)$ for two consecutive iterations, the "stepping" portion of Algorithm 5 has converged, and the coefficient $\tilde{\vec{\theta}}$ is set to the converged value, $\tilde{\vec{\theta}} = \vec{\theta}_{t+1}$.

Iterative MCMC-MLE Algorithm

Recall that the best estimate for $\hat{\vec{\theta}}$ occurs when $\tilde{\vec{\theta}}$ is "close" to $\hat{\vec{\theta}}$. The stepping algorithm in Algorithm 5 provides the coefficient $\tilde{\vec{\theta}}$ used to initialize the MCMC-MLE algorithm that estimates the MLE $\hat{\vec{\theta}}$. An estimate of $\vec{\theta}_{(r+1)}$ is obtained by iterating through two steps in Algorithm 6, where a hill-climbing algorithm is used in the maximization of Step 4.

Metropolis-Hastings Algorithm

The Metropolis-Hastings sampling algorithm, a Markov Chain Monte Carlo method, is used to generate a sample of networks required in Step 2 of Algorithm 5 and Step 3 of

Algorithm 5 Estimation of $\tilde{\vec{\theta}}$ by stepping algorithm

1: Set the iteration number t equal to 0 and select an initial coefficient $\vec{\theta}_0$, which is often taken to be the maximum pseudolikelihood estimate.

2: Use the Metropolis-Hastings in Algorithm 7 to generate a sample of K networks $x_1, x_2, ..., x_K$ from probability density function $f_X(x; \vec{\theta}_t)$ defined as

$$f_X(x; \vec{\theta}_t) \quad = \quad \frac{\exp\{\vec{\theta}_t^T \vec{S}(x)\}}{\int_{z \in \mathcal{X}} \exp\{\vec{\theta}_t^T \vec{S}(z)\} dz}, \quad x \in [0,1]^M$$

3: Calculate the sample mean $\bar{\zeta}_t = \frac{1}{K} \sum_{k=1}^{K} \vec{S}(x_k)$.

4: For some $\gamma_t \in (0, 1]$, define $\hat{\zeta}_t$ as

$$\hat{\zeta}_t \quad = \quad \gamma_t \vec{S}(x_{\text{obs}}) + (1 - \gamma_t)\bar{\zeta}_t$$

5: Update $\vec{\theta}_{t+1}$ via

$$\vec{\theta}_{t+1} \quad = \quad \text{argmax}_{\vec{\theta}} \left[\left(\vec{\theta} - \vec{\theta}_t \right)^T \hat{\zeta}_t - \log \left(\frac{1}{K} \sum_{k=1}^{K} \exp\{ \left(\vec{\theta} - \vec{\theta}_t \right)^T \vec{S}(x_k) \} \right) \right]$$

using a hill-climbing algorithm.

6: Increment t and return to Step 2.

Algorithm 6 Estimation of $\hat{\vec{\theta}}$ by iterative MCMC-MLE.

1: Initialize $\vec{\theta}_{(0)}$
2: **while** $\triangle\left(\vec{\theta}_{(r+1)}, \vec{\theta}_{(r)}\right) <$ tolerance **do**
3: Use the Metropolis-Hastings in Algorithm 7 to generate sample of networks $x_1, x_2, ..., x_K$ from probability density function $f_X(x; \vec{\theta}_{(r)})$ defined as

$$f_X(x; \vec{\theta}_{(r)}) = \frac{\exp\{\vec{\theta}_{(r)}^T \vec{S}(x)\}}{\int_{z \in \mathcal{X}} \exp\{\vec{\theta}_{(r)}^T \vec{S}(z)\} dz}, \quad x \in [0,1]^M$$

4: Update $\vec{\theta}_{(r+1)}$ via

$$\vec{\theta}_{(r+1)} = \text{argmax}_{\vec{\theta}} \left(\left(\vec{\theta} - \vec{\theta}_{(r)}\right)^T \vec{S}(y) - \log\left(\frac{1}{K} \sum_{k=1}^{K} \exp\left(\left(\vec{\theta} - \vec{\theta}_{(r)}\right)^T \vec{S}(x_k) \right) \right) \right)$$

using a hill-climbing algorithm.
5: **end while**

Algorithm 6. To implement the sampling algorithm, three aspects need specification: the target function, proposal function, and acceptance probability.

The target function is the desired stationary distribution for the Markov Chain, which in this setting, is the GERGM distribution,

$$f_X(x; \vec{\theta}) = \frac{\exp\{\vec{\theta}^T \vec{S}(x)\}}{\int_{z \in \mathcal{X}} \exp\{\vec{\theta}^T \vec{S}(z)\} dz}, \quad x \in [0,1]^M$$

The algorithm generates the next network, x_{p+1}, in the Markov Chain by nominating a proposal network, x^*_{p+1}, based upon the previously accepted network x_p. Formally, define the proposal distribution as the truncated normal distribution with mean dependent upon the jk-th edge-weight in network x_p, denoted $x_{j,k;p}$, and fixed variance σ^2 as $q\left(x^*_{j,k;p+1} | x_{j,k;p}\right) = TN\left(x_{j,k;p}, \sigma^2, 0, 1\right)$. A proposal network x^*_{p+1} is generated after a new weight has been proposed for each edge. The algorithm produces an irreducible Markov Chain among all

networks with N nodes and M edges with weights in [0,1], as each proposal network is accessible with positive transition probability.

Once a proposal network x^*_{p+1} is generated, the Metropolis-Hastings algorithm will either accept the proposal network into the sample, $x_{p+1} = x^*_{p+1}$, with probability $\alpha\left(x_p, x^*_{p+1}\right)$, or reject it and re-accept the previously accepted network into the sample, $x_{p+1} = x_p$. The acceptance probability $\alpha\left(x_p, x^*_{p+1}\right)$ ensures proper sampling of the target distribution by including the probability that transitions from the proposal function may not be symmetric. The acceptance probability $\alpha\left(x_p, x^*_{p+1}\right)$ is defined as:

$$\alpha(x_p, x^*_{p+1}) \quad = \quad \min\left(1, \frac{f_X(x^*_{p+1}; \vec{\theta}) q\left(x_p | x^*_{p+1}\right)}{f_X(x_p; \vec{\theta}) q\left(x^*_{p+1} | x_p\right)}\right)$$

In summary, the Metropolis-Hastings algorithm for GERGMs is listed in Algorithm 7.

Algorithm 7 The MCMC procedure of Metropolis-Hastings for GERGMs.

Require: Of K networks, p networks have already been accepted into the sample.

1: Generate a proposal network x^*_{p+1} by changing the weight $x_{j,k;p}$, according to

$$q\left(x^*_{j,k;p+1}|x_{j,k;p}\right) \quad = \quad TN\left(x_{j,k;p}, \sigma^2, 0, 1\right)$$

independently across all edges.

2: Set the $(p+1)$-st sample

$$x_{p+1} \quad = \quad \begin{cases} x^*_{p+1} \text{ with probability } \alpha\left(x_p, x^*_{p+1}\right) \\ x_p \text{ with probability } 1 - \alpha\left(x_p, x^*_{p+1}\right) \end{cases}$$

where

$$\alpha(u, v) \quad = \quad \min\left(1, \frac{f_X(v;\vec{\theta})}{f_X(u;\vec{\theta})} \prod_{j,k} \frac{q\left(u_{j,k}|v_{j,k}\right)}{q\left(v_{j,k}|u_{j,k}\right)}\right)$$

for probability density function $f_X(x;\vec{\theta})$ defined as

$$f_X(x;\vec{\theta}) \quad = \quad \frac{\exp\{\vec{\theta}^T \vec{S}(x)\}}{\int_{z\in\mathcal{X}} \exp\{\vec{\theta}^T \vec{S}(z)\} dz}, \quad x \in [0,1]^M$$

CHAPTER 6

PROBABILITY DISTRIBUTIONS

6.1 INTRODUCTION

While there exist many centrality measures to identify influential nodes within biological networks, their distributions in a random network remain open-ended. As a step towards developing a framework to determine the distribution of betweenness centrality, Chapter 2 highlighted the use of the centrality measure in the identification of *Etv5* as an essential gene when comparing networks constructed from diseased and healthy optic glioma samples. Motivated by the use of centrality measures to identify genes whose role changes as tissue transitions from a healthy to diseased state, Chapter 3 investigated the variability of betweenness to identify these essentially different genes through two separate edge-weight perturbations. Yet, as suggested through the sampling distributions in Figures 3.5 and 3.6 and formally proven in Chapter 4, the betweenness measure, in its standard definition, is sensitive to edge-weight perturbations, as an arbitrarily small change to its weights produces large changes in its values. As a result, a finite stable betweenness measure was defined as a measure robust to edge-weight perturbations and whose distribution will be investigated in an effort to better understand its variability.

Recall that a **parameter** is defined as a numerical quantity that characterizes the population from which data can be obtained, while a **statistic** is defined as a numerical quantity that characterizes a sample from the population. Although more details are provided in Section 6.3.2, the proposed framework to build a distribution of finite stable betweenness involves the utilization of the generalized exponential random model (refer to Section 5.3) to

generate a sample of networks from an underlying probability distribution. After calculating the finite stable betweenness centrality of a particular node from each network in the sample, the resulting values are used to construct a sampling distribution for the finite stable betweenness statistic.

6.2 SAMPLING DISTRIBUTION

A sampling distribution of a statistic is defined as the distribution of the values of the statistic that are computed on all possible samples taken from the population. Knowing the sampling distribution may provide information about the reproducibility and accuracy of the statistic through standard errors (i.e., how much to expect estimates to vary between experiments), bias (i.e., the expected difference between the estimate and true value) and confidence intervals (i.e., how close the estimate is to the true value) (Kulesa et al., 2015). As examples, consider Figures 3.5 and 3.6 which depict the sampling distribution of the betweenness statistic for each statistically different gene identified in the glioma dataset through two separate edge-weight perturbation methods.

Although the sampling distribution of a statistic may sometimes be calculated analytically, a theoretical framework to predict the sampling distribution may be difficult to develop. In such cases, the bootstrap method builds the sampling distribution empirically and approximates its shape by simulating replicated samples based upon the observed data (Kulesa et al., 2015). In Durón et al. (2018), the bootstrap method was employed to generate an empirical sampling distribution in order to analyze the variability of the betweenness centrality measure in the identification of essential genes. The bootstrap method is summarized in Section 3.3.4.

The generalized exponential random graph model (GERGM) provides a mathematical model to approximate the sampling distribution of the finite stable betweenness statistic. Yet, because the variability of the sampling distribution is driven by the variability within the data, a brief overview of how the inherent variability of the data can influence the sampling distribution is subsequently provided before detailing the proposed GERGM framework (refer to Section 6.3.2) used to model one possible distribution of the finite stable betweenness measure.

6.2.1 Variability in the Data

Inferences about a population are based upon an evaluation of repeated samples taken from the population and are driven by the variability that exists within the dataset. Recall that the topological representation of a network is a collection of nodes and edges. As discussed in Section 3.3.2, the edge-weights in a weighted network are functions of the correlations between the measurements made on each gene, or node. While the use and analysis of weighted, correlation-based networks is becoming increasingly prevalent in biological applications (Yates and Mukhopadhyay, 2013), different weighting schemes can result in different analyses of the network (Ghosh et al., 2014). Therefore, although the inherent variability in the dataset drives the variability present in the network, the variability in the data also influences the variability of the distribution of the sample statistic.

6.3 THE PROPOSED GERGM FRAMEWORK

6.3.1 Overview of GERGMs

As discussed in Section 5.3, the generalized exponential random graph model (GERGM) represents a general class of models based in exponential-family theory that specifies the probability distribution for a set of weighted networks. Within the GERGM framework, a maximum-likelihood estimate for the coefficient vector $\vec{\theta}$ of a specified model is obtained for a given dataset, where additional networks can then be simulated from the probability distribution implied by the specified model.

With the sample space of networks $\mathcal{X} \in \{[0,1]^M\}$ representing the set of all possible weighted networks with N nodes and M edges with weights in $[0,1]$, recall the GERGM probability density function of a weighted network $x \in \mathcal{X}$:

$$f_X(x;\vec{\theta}) \quad = \quad K \exp\{\vec{\theta}^T \vec{S}(x)\}, \quad x \in [0,1]^M \tag{6.1}$$

where $K = \int_{z \in \mathcal{X}} \exp\{\vec{\theta}^T \vec{S}(z)\} dz$ is a normalizing constant, and $\vec{S}(x)$ is a vector of network statistics computed on the network x with the same number of elements as the coefficient vector $\vec{\theta}$ (Desmarais and Cranmer, 2012).

Distribution of Networks

Once the coefficients of the GERGM model are estimated, the model is completely specified and defines a probability distribution on a collection of networks of identical size. In particular, because the maximum likelihood estimate provides the model with coefficient specifications that create the highest likelihood of the observed network, the specified dis-

tribution of networks is centered around the observed network statistics. Therefore, if the model is a good fit to the network derived from the observed data, then networks drawn from the GERGM distribution are likely to resemble the observed network.

As an example, consider two networks that are identical with respect to the values of the network statistics vector $\vec{S}(x)$. The likelihoods of the two networks are identical because, as specified by the GERGM model, every network that exhibits the same value of network statistics that are included in the model specification is equally likely to be observed.

6.3.2 Proposed Distributions of the Finite Stable Betweenness Statistic

Because the betweenness centrality captures some global structural properties of a network, the proposed framework to determine a distribution of finite stable betweenness (whose definition is provided in Section 4.4) involves incorporating the measure into the GERGM probability model as a network statistic. Examples of possible specifications for the finite stable betweenness measure as a network statistic, and thus possible distributions of the measure, are highlighted below.

To determine one possible model that explicitly specifies the distribution of the finite stable betweenness centrality, define the network statistic $S(x)$ to equal $C_{FSB}(v_i)$, the finite stable betweenness value of node v_i using the definition provided in Equations (4.4) and (4.5). By letting $S(x) = C_{FSB}(v_i)$, the GERGM density function becomes

$$f_X(x;\theta) \quad \propto \quad \exp\{\theta \cdot C_{FSB}(v_i)\} \tag{6.2}$$

Thus, the distribution of $\hat{C}_{FSB}(v_i)$, the finite stable betweenness statistic of node v_i, follows

an exponential distribution with rate equal to $-\theta$, provided $\theta < 0$.

As another example, let the network statistic $S(x)$ be the square of the difference between the observed and target finite stable betweenness value of node v_i, $(C_{FSB}(v_i) - C_{FSB}(v_i)^*)^2$. Upon letting $S(x) = (C_{FSB}(v_i) - C_{FSB}(v_i)^*)^2$, the GERGM specification becomes

$$f_X(x;\theta) \quad \propto \quad \exp\{\theta\,(C_{FSB}(v_i) - C_{FSB}(v_i)^*)^2\} \tag{6.3}$$

In this construction, the statistic $\hat{C}_{FSB}(v_i)$ follows a normal distribution with mean and variance equal to $C_{FSB}(v_i)^*$ and $-\frac{1}{2\theta}$, respectively, provided $\theta < 0$.

Finally, let the network statistics vector $\vec{S}(x)$ be a combination of the finite stable betweenness statistic and other weighted network statistics, such as two-stars and transitive triads as depicted in Figure 6.1. Additional weighted network statistics that may be included in the GERGM probability model are listed in Table 5.1.

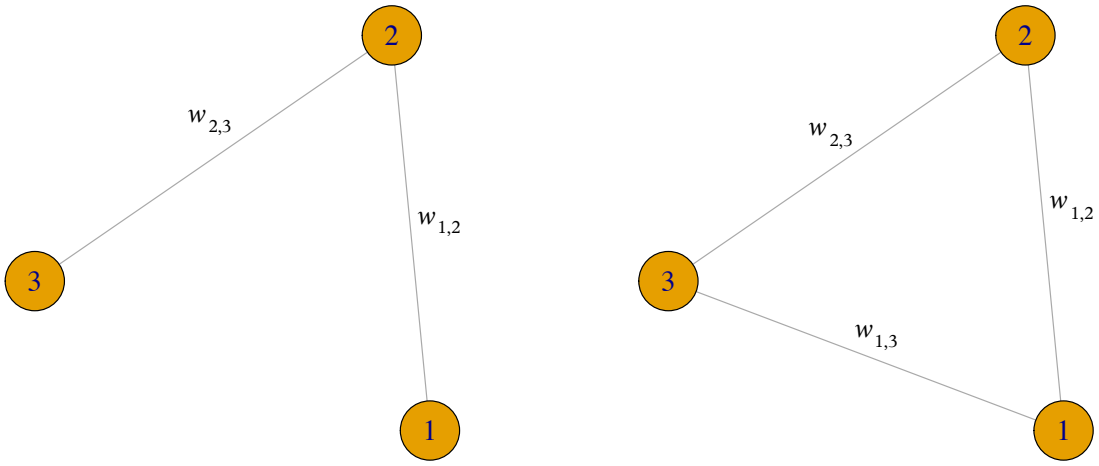

Figure 6.1. Two-stars (left) and transitive triads (right), also referred to as triangles, are two possible weighted network configurations that may be included in the network statistic vector $\vec{S}(x)$.

By letting $\vec{S}(x) = (C_{FSB}(v_i), 2Stars, Ttriad)$, the GERGM specification becomes

$$f_X(x;\vec{\theta}) \quad \propto \quad \exp\{\vec{\theta}^T \vec{S}(x)\} \tag{6.4}$$

from which the distribution of $\hat{C}_{FSB}(v_i)$ is derived. Note that the distribution of $\hat{C}_{FSB}(v_i)$ depends upon the definition of the vector of network statistics $\vec{S}(x) = (C_{FSB}(v_i), 2Stars, Ttriad)$ for $\vec{\theta} = (\theta_1, \theta_2, \theta_3)$.

6.4 APPLICATION

The GERGM provides a description of the distribution of networks from which a distribution of finite stable betweenness can be determined. In particular, and as demonstrated through the examples in Section 6.3.2, the GERGM model allows the distribution of the finite stable betweenness statistic to be flexible, yet dependent on the definition of the vec-

tor of network statistics $\vec{S}(x)$. Yet, the variability of the empirical sampling distributions in Figures 3.5 and 3.6 motivate an important question: Given the variability in the data, is the variability of finite stable betweenness explained by the GERGM model?

6.4.1 The Noisy Structure Model

To address the question above, the distribution of finite stable betweenness determined by the GERGM model can be compared to the sampling distribution determined by a model that generates variability in a realistic manner. One natural way to determine such a distribution is to generate variability by perturbing the edge-weights of a network in a small, yet intelligent, fashion. A similar approach was taken in Chapter 3 to examine the variability of the betweenness measure in the identification of genes essential to the structure of a diseased state.

Previous work by Hardin et al. (2013) provides an algorithm to produce variability in a reasonable, yet controlled fashion. In particular, the proposed algorithm generates $\hat{\Sigma}$, a sample network from the known-correlation structure of a population network Σ by perturbing the known structure with noise. Using the sample of networks generated according to this model, an empirical sampling distribution of the network statistic(s) can be determined and used to assess the ability of the GERGM model to capture the variability of the distribution of finite stable betweenness.

According to the algorithm of Hardin et al. (2013), a simulated N-node network $\hat{\Sigma}$ is defined as:

$$\hat{\Sigma} \;=\; \Sigma + \epsilon \left(U^T U - I \right) \tag{6.5}$$

where $U = (\vec{u}_1, \vec{u}_2, ..., \vec{u}_N)$ is a matrix of unit vectors generated from a high-dimensional noise space, ϵ is the maximum entry-wise random noise, and I is the identity matrix. Note ϵ is a function of ρ and τ, and is defined in the algorithm.

The Population Network

Utilizing the algorithm provided by Hardin et al. (2013) with correlation $\rho = 0.9$ and step-size $\tau = \frac{2}{45}$ such that $\epsilon = \frac{1}{15}$, the known-correlation structure of a 20-node network is given below:

$$\Sigma = \begin{pmatrix} 1 & \alpha_2 & \alpha_3 & \alpha_4 & \dots & \alpha_{20} \\ \alpha_2 & 1 & \alpha_2 & \alpha_3 & \dots & \alpha_{19} \\ \alpha_3 & \alpha_2 & 1 & \alpha_2 & \dots & \alpha_{18} \\ \alpha_4 & \alpha_3 & \alpha_2 & 1 & \dots & \alpha_{17} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_{20} & \alpha_{19} & \alpha_{18} & \alpha_{17} & \dots & 1 \end{pmatrix} \tag{6.6}$$

where $\alpha_i = \rho - \tau(i-2)$ for $2 \leq i \leq 20$.

As weighted networks based upon gene expression data are often constructed by thresholding correlation among the gene measurements (Langfelder and Horvath, 2008), a similar approach is taken to generate a network from the known structure provided in Equation (6.6). The "true" population network, as presented in Figure 6.2, is constructed by first thresholding the node-pair correlation values, and then assigning edge-weights equal to one minus the absolute value of the remaining node-pair correlations. Refer to Section 3.3.2 for additional details on the construction of such edge-weights.

Thresholds are applied to the correlation values in order to reduce the density (i.e.,

the actual number of edges in proportion to the maximum possible number of edges) of the population network. As a result, nodes 10 and 11 in the population network in Figure 6.2 share the largest finite stable betweenness value of 1. The finite stable betweenness value of each node in the population network is listed in Table 6.1.

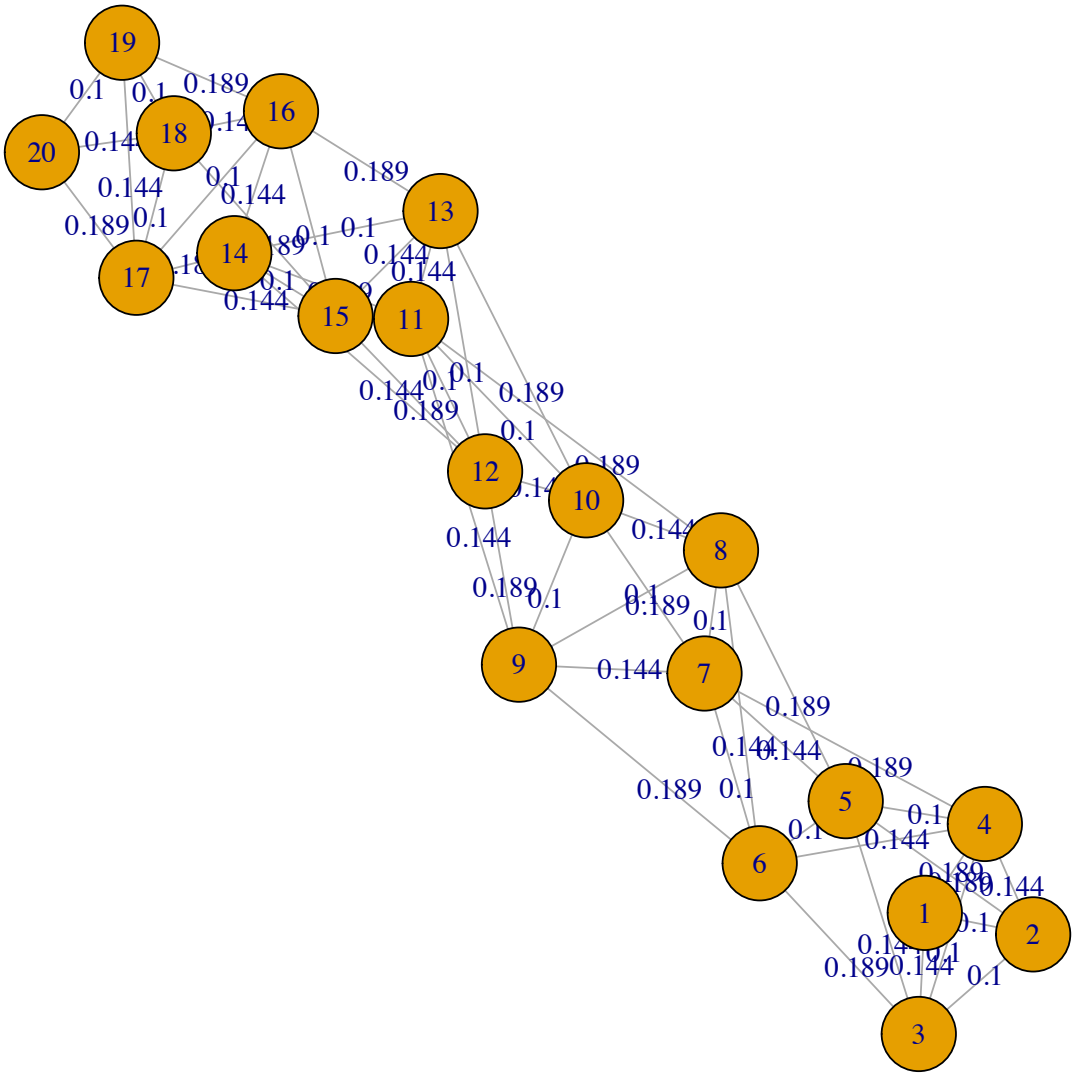

Figure 6.2. The population network: a 20-node weighted network with 54 edges generated with a known-correlation structure following the work provided by Hardin et al. (2013).

Table 6.1. The finite stable betweenness parameter value of each node in the weighted and undirected population network in Figure 6.2.

Node v_i	$C_{FSB}(v_i)$
1	0.0
2	0.0
3	0.0
4	0.556
5	0.556
6	0.444
7	0.889
8	0.889
9	0.667
10	1.0
11	1.0
12	0.667
13	0.889
14	0.889
15	0.444
16	0.556
17	0.556
18	0.0
19	0.0
20	0.0

6.4.2 The Proposed GERGM Models

To determine a distribution of finite stable betweenness of one of the essential nodes in the population network, say node 10, two separate GERGM models are proposed based upon different combinations of network statistics, and are specified as:

Model	**Network Statistics**	**Metropolis-Hastings SD**
1	Finite stable betweenness of node 10	0.005
2A	Finite stable betweenness of node 10 and two-stars	0.005
2B	Finite stable betweenness of node 10 and two-stars	0.01

where the finite stable betweenness statistic, denoted $FSB(10)$, is defined as

$$FSB(10) \quad = (C_{FSB}(10) - 1)^2$$

with $C_{FSB}(10)^* = 1$ as provided from the finite stable betweenness parameters given in Table 6.1; and the two-stars statistic, denoted $2Stars$, is defined the sum of the product of all existing edge-weights $w_{i,j}, w_{j,k}$. The two-star configuration is depicted by the left structure in Figure 6.1.

The GERGM Model 1 has the probability density function

$$f_X(x;\theta_1) \quad \propto \quad \exp\{\theta_1 \cdot FSB(10)\} \tag{6.7}$$

while GERGM Models 2A and 2B have the probability model

$$f_X(x;\theta_1,\theta_2) \quad \propto \quad \exp\{\theta_1 \cdot FSB(10) + \theta_2 \cdot 2Stars\} \tag{6.8}$$

where θ_1 and θ_2 are the corresponding coefficients to the network statistics $FSB(10)$ and $2Stars$, respectively. Although GERGM Model 2A and 2B have the same probability model, the difference lies in how the sampling in the proposal distribution is generated by the Metropolis-Hastings algorithm (refer to Algorithm 6).

6.4.3 The GERGM Estimation Results

The results in Table 6.2 detail the coefficient estimates that best describe the population network in Figure 6.2 according to the specified GERGM model. Recall that $FSB(10)$ denotes the finite stable betweenness network statistic of node 10, while $2Stars$ denotes the sum of the product of all existing edge-weights $w_{i,j}, w_{j,k}$.

Table 6.2. The coefficient estimates for each proposed GERGM model are detailed below.

Network Statistics	**Metropolis-Hastings SD**	$\hat{\theta}_1, \hat{\theta}_2$ **Estimate**	**Standard Error Estimate**
$FSB(10)$	0.005	-2047.096, NA	2866.749, NA
$FSB(10)$, $2Stars$	0.005	-424.733, -17.639	654.158, 3.065
$FSB(10)$, $2Stars$	0.01	-186.981, -19.897	269.767, 4.381

Examination of the MLEs

To better understand the distribution of networks produced by the GERGM models, and thus obtain a better understanding of the distribution of finite stable betweenness de-

termined by each model, the maximum likelihood estimates $\hat{\vec{\theta}}$ should be considered. Recall the estimates $\hat{\vec{\theta}}$ in Table 6.2 are produced by the MCMC estimation process detailed in Algorithm 6. As an example, consider the probability density function of GERGM Model 2B in Equation (6.8) where $\vec{\theta} = (\theta_1, \theta_2)$.

Although more details are provided in Section 5.3.1, the maximum likelihood estimates (MLE) $\hat{\vec{\theta}}$ are determined by maximizing the difference in log-likelihoods due to the computational intractability of the normalizing constant present in the log-likelihood function. Given the probability model in Equation (6.8), the difference of log-likelihoods can be defined as:

$$\ell(\vec{\theta}|x_{\text{obs}}) - \ell(\tilde{\vec{\theta}}|x_{\text{obs}}) \quad = \quad \left(\vec{\theta} - \tilde{\vec{\theta}}\right)^T \vec{S}(x_{\text{obs}}) - \log\left(\frac{1}{K}\sum_{k=1}^{K} \exp\{\left(\vec{\theta} - \tilde{\vec{\theta}}\right)^T \vec{S}(x_k)\}\right) \quad (6.9)$$

where $\tilde{\vec{\theta}} = (\tilde{\theta}_1, \tilde{\theta}_2)$ is an initial coefficient vector, $\vec{S}(x_{\text{obs}})$ is the vector of statistics $FSB(10)$ and $2Stars$ for the observed network, and $\vec{S}(x_k)$ is the network statistics vector corresponding to the statistic values of the k-th network simulated from the distribution $f_X(x; \tilde{\vec{\theta}})$.

To obtain an accurate approximation of the log-likelihood function, and thus accurate MLEs $\hat{\vec{\theta}} = (\hat{\theta}_1, \hat{\theta}_2)$, the MCMC sampling algorithm must sample networks from a region where the probability mass in Equation (6.8) is concentrated. Because the mass is concentrated to a small region, the MCMC sampling algorithm must begin with a coefficient very close to $\hat{\vec{\theta}}$. The stepping algorithm, as detailed in Section 5.3.1, is used to determine an appropriate value of $\tilde{\vec{\theta}}$ which is, in turn, used to initialize the sampling algorithm.

Setting $\vec{S}(x_{\text{obs}}) = (0, 5.017)$ and $\tilde{\vec{\theta}} = (-18.525, -17.973)$ as provided by the stepping algorithm, and simulating a sample of 5000 networks from the distribution $f_X(x; \tilde{\vec{\theta}})$ specified

as

$$f_X(x;\tilde{\vec{\theta}}) \;\propto\; \exp\{-18.525 \cdot FSB(10) - 17.973 \cdot 2Stars\} \tag{6.10}$$

the difference of log-likelihoods for GERGM Model 2B in Equation (6.9) can be re-expressed as

$$\begin{aligned}\ell(\theta_1, \theta_2 | x_{\text{obs}}) &= (\theta_1 + 18.525) \cdot 0 + (\theta_2 + 17.973) \cdot 5.017 \\ &\quad - \log\left(\frac{1}{5000}\sum_{k=1}^{5000} \exp\{(\theta_1 + 18.525) \cdot FSB_k(10) + (\theta_2 + 17.973) \cdot 2Stars_k\}\right)\end{aligned} \tag{6.11}$$

where $FSB_k(10)$ and $2Stars_k$ are the $FSB(10)$ and $2Stars$ values for network x_k simulated from the probability distribution $f_X(x;\tilde{\vec{\theta}})$ provided in Equation (6.10).

In Figure 6.3, the difference in log-likelihoods in Equation (6.11) has its maximum at (-200, -19.9, 1.298), suggesting that the highest likelihood of the observed network occurs when $\theta_1 = -200$ and $\theta_2 = -19.9$. Recall the maximum likelihood estimates $\hat{\vec{\theta}} = (\hat{\theta}_1, \hat{\theta}_2) = (-186.981, -19.897)$ in Table 6.2. The maximum of the difference in log-likelihoods in Equation (6.11) confirms the MLE for the $2Stars$ statistic and thus, the clear existence of a maximum in the $2Stars$ direction. But the plot in Figure 6.3 suggests a flat curvature of the difference in log-likelihoods in the direction of $FSB(10)$, and thus, a lack of precision in the estimate of the $FSB(10)$ statistic.

As discussed in Section 6.3.2, the variability of the proposed distribution of $\hat{C}_{FSB}(v_i)$ is dependent upon θ_1, the coefficient of the $FSB(10)$ statistic. The flat curvature in the

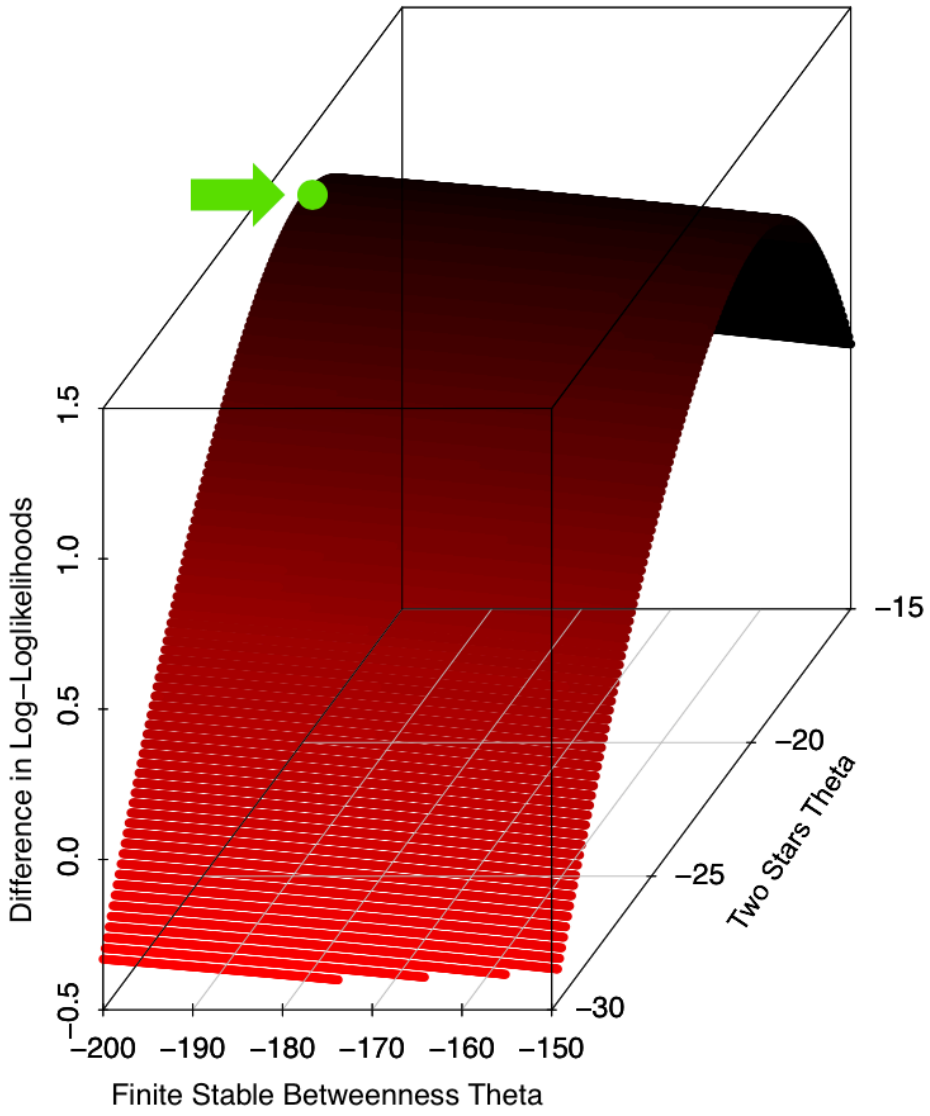

Figure 6.3. The difference in log-likelihoods in Equation (6.11), where a standard deviation of 0.01 was specified in the Metropolis-Hastings sampling algorithm. The maximum occurs at (-200, -19.9, 1.298) and is colored in green. A green arrow has been added to the plot to provide assistance in its identification.

direction of the $FSB(10)$ statistic suggests that the GERGM model is flexible in creating variability of the distribution of $\hat{C}_{FSB}(v_i)$, as multiple yet equivalent coefficients for the $FSB(10)$ statistic produce similar maximizations of the likelihood of observing the network.

6.4.4 The Sampling Distribution of Finite Stable Betweenness

Using the maximum likelihood estimates provided in Table 6.2, 5000 networks are simulated from the probability functions corresponding to each of the two proposed GERGM models. The simulation is accomplished using the Metropolis-Hastings sampling algorithm detailed in Algorithm 7, which constrains each simulated network to have a topology identical to the population network in Figure 6.2. Once a sample of networks is generated from the corresponding probability density function defined by each GERGM model, the finite stable betweenness value of node 10 is calculated in each of the 5000 simulated networks. Figure 6.4 displays the sampling distribution of the finite stable betweenness as determined by each GERGM model, generated from three separate samples of finite stable betweenness values of node 10.

To assess whether the variability of finite stable betweenness is explained by the GERGM model, the distribution of the centrality measure determined by the GERGM model is compared to the sampling distribution generated by the model discussed in Section 6.4.1 that produces variability in a realistic yet intelligent manner. Using the framework provided by Hardin et al. (2013), 5000 networks are simulated whose topology is identical to that of the population network in Figure 6.2. The finite stable betweenness value of node 10 is calculated from each of the simulated networks, from which an empirical distribution is

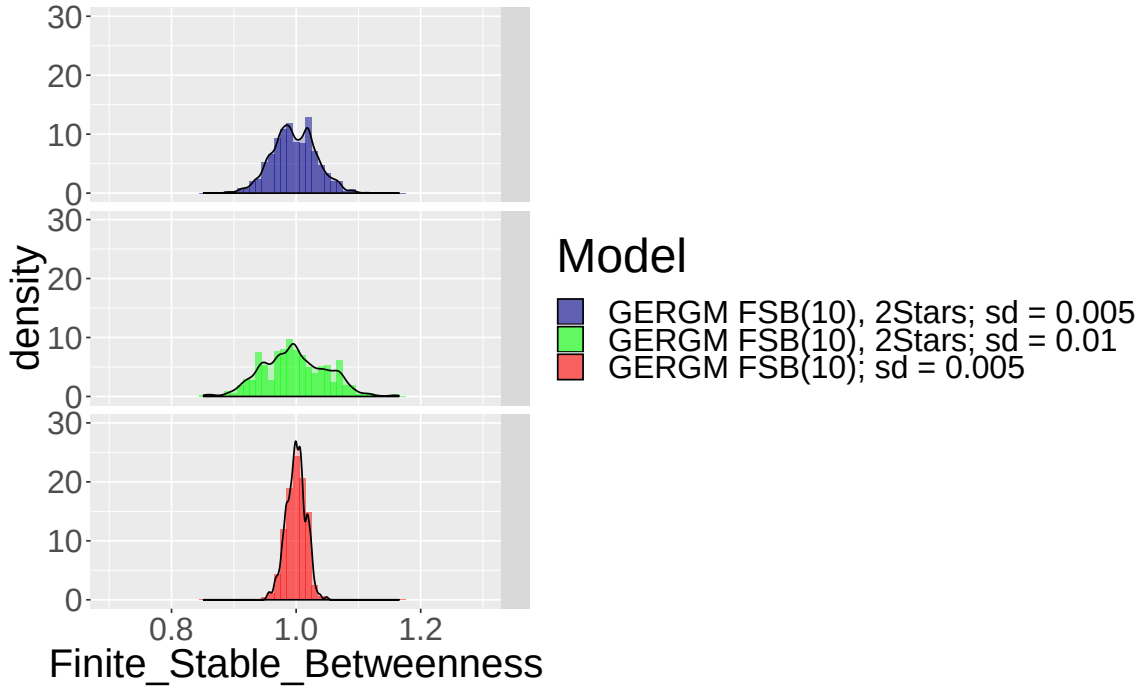

Figure 6.4. Histograms of each sampling distribution for the finite stable betweenness of node 10 as determined by each of the two proposed GERGM models. Although each sampling distribution has approximately the same mean of a finite stable betweenness value of 1, the distributions determined by the two GERGM models have a notably different standard error.

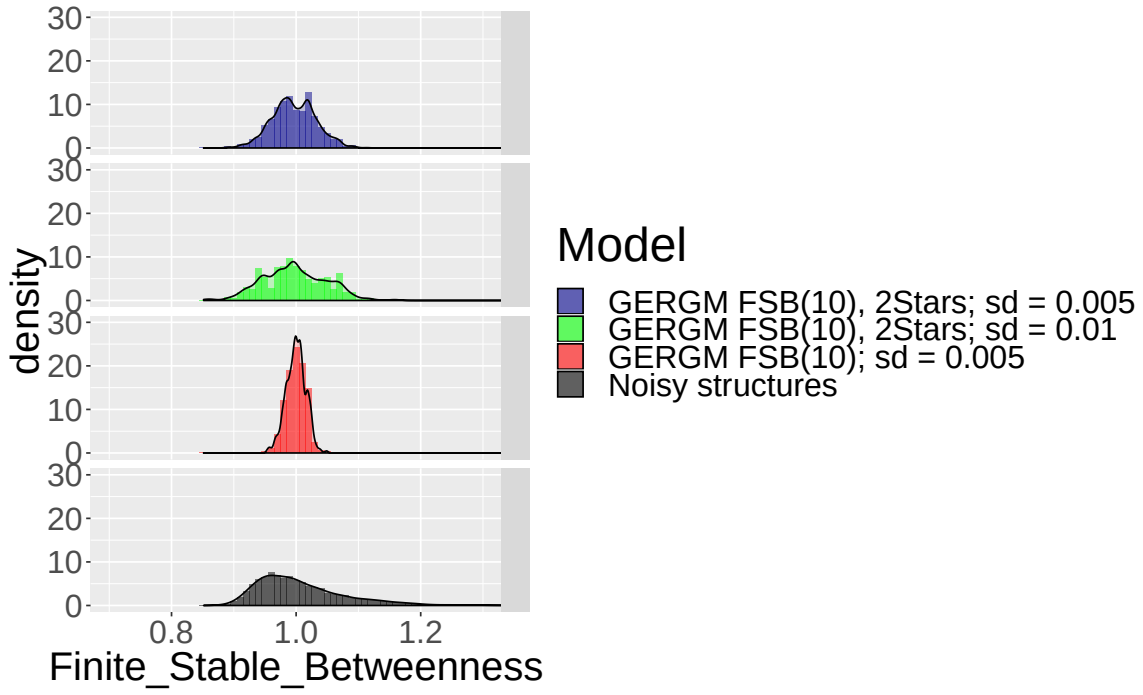

Figure 6.5. The sampling distribution of the finite stable betweenness centrality determined by the two GERGM models (A - C) and the noisy structure model (D) provided by Hardin et al. (2013).

generated.

The sampling distributions of finite stable betweenness determined by the two GERGM models and the noisy structure model (Section 6.4.1) are displayed in Figure 6.5. Additionally, the mean and standard error of the sampling distributions of finite stable betweenness determined by each GERGM model and the noisy structure model are detailed in Table 6.3.

Table 6.3. The mean and standard error of the sampling distributions of finite stable betweenness of node 10 generated by each model.

Model	**Mean**	**Standard Error**	**-$\frac{1}{2\hat{\theta}_1}$**
$FSB(10)$; sd = 0.005	0.999	0.0002	0.0002
$FSB(10)$, $2Stars$; sd = 0.005	0.997	0.001	0.001
$FSB(10)$, $2Stars$; sd = 0.01	0.997	0.002	0.002
noisy structure	1.006	0.003	NA

6.5 DISCUSSION

Consider the plots in Figure 6.5 (A - C) which depict the sampling distributions determined by GERGM Models 2A and 2B. By construction of the statistic $FSB(10)$, the sampling distribution of finite stable betweenness is expected to follow a normal distribution, conditional on the value of the $2Stars$ statistic, with mean equal to the true finite stable betweenness value of node 10 and variance equal to $-\frac{1}{2\theta_1}$, provided $\theta_1 < 0$. Upon substitution of the estimated value of θ_1 from each GERGM model provided in Table 6.2 into the proposed variance equation $-\frac{1}{2\theta_1}$, the estimated variance of each sampling distribution determined by the GERGM model is confirmed by its proposed variance (last column in Table 6.3). Yet although the mathematical model of the finite stable betweenness distribution determined by the GERGM model describes the variability evidenced in Figure 6.5, the model is unable to capture the natural variability present within the data. This distinction is apparent by comparing the distributions determined by the GERGM models (Figure 6.5 A - C) with the distribution (Figure 6.5 D) determined by the noisy structure model (refer

to Section 6.4.1). But as suggestive of the curvature of the difference in log-likelihoods, a particular value of θ_1 can be set in the GERGM model to match the variability produced by the realistic noise model without substantially decreasing the likelihood of observing the network.

Yet, the discrepancy between the sampling distributions suggests the question: How is the step size generated by the Metropolis-Hastings sampling in the derivation of each GERGM model connected to the variability of the sampling distribution? To address this question, consider the histograms of the noise generated by each GERGM model (A - C) and the noisy structure model (D) in Figure 6.6, where noise is defined as the difference in edge-weights of each simulated network in the sample with that of the population network. As seen in Figure 6.6, more noise is generated by the Metropolis-Hastings sampling algorithm in the GERGM models than is generated by the noisy structure model. Yet, recall the Metropolis-Hastings algorithm detailed in Algorithm 7, in which each network in the sample is produced by perturbing the edge-weights of the previously accepted network. As a result of the sampling algorithm, as more networks are accepted into the sample, the difference in edge-weights between each simulated network and the population network will increase. Put another way, networks accepted towards the end of the sample will differ more in their edge-weight comparison with the population network as opposed to the networks accepted earlier in the sample. Therefore, the difference in edge-weights between each successive network in the sample and the population network will propagate as more and more networks are accepted into the sample.

Consider the histograms of noise in Figure 6.7, where noise is now defined as the difference in edge-weights of each successive, yet distinct network in the sample of 5000 networks,

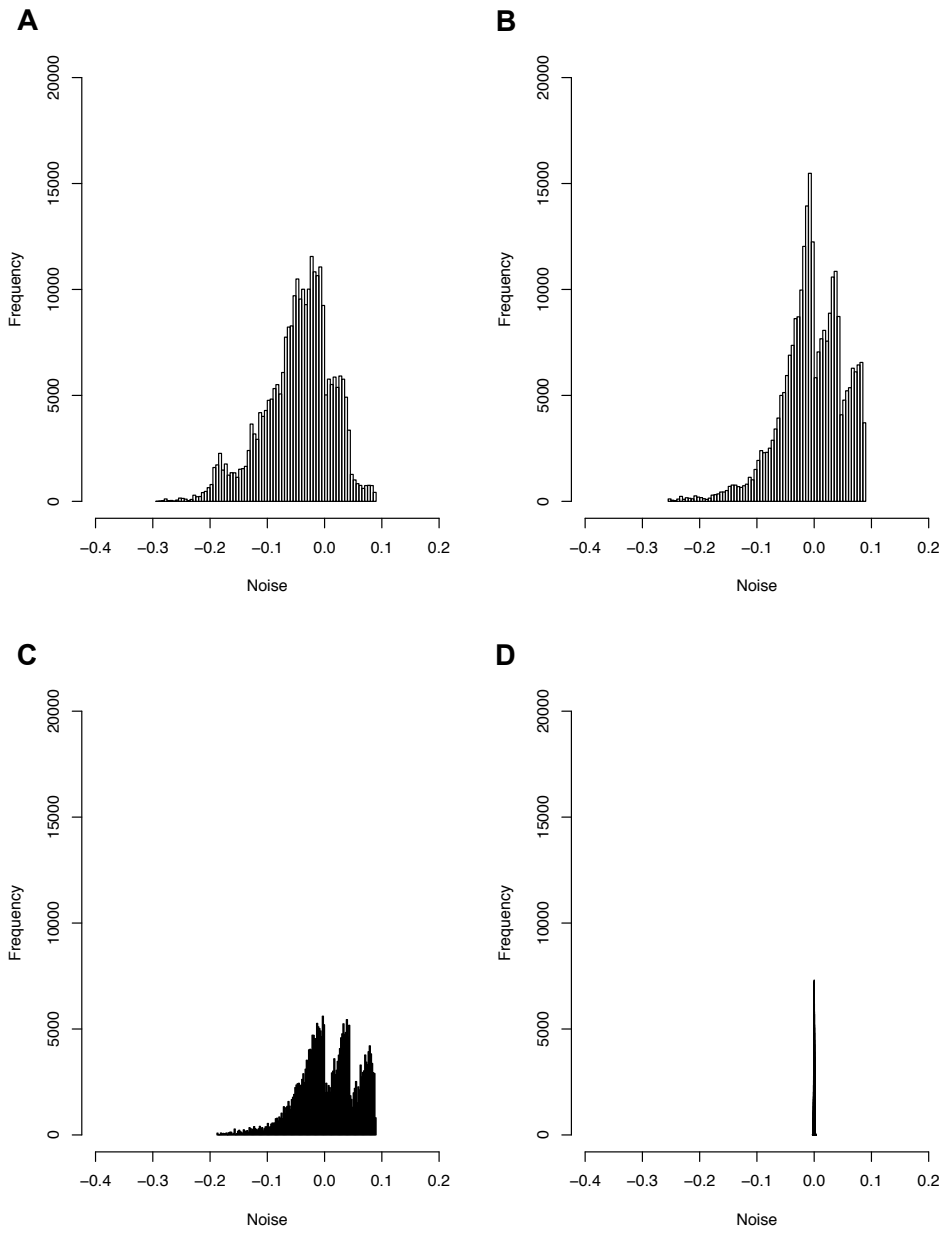

Figure 6.6. Histograms of the noise, defined as the difference in edge-weights of each of the 5000 simulated networks and the population network, generated by GERGM Model 2A with standard deviation 0.005 (A), GERGM Model 2B with standard deviation 0.005 and 0.01, respectively (B - C), and the noisy structure model (D).

for each proposed GERGM model. Because the proposal distribution is a truncated normal distribution in the Metropolis-Hastings algorithm, each noise distribution, as expected, follows such a distribution. Although a sample of networks was generated according to the distribution specified by each GERGM model, only 9.56%, 14.4%, and 5.84% of each of the 5000 simulated networks are distinct networks. The small percentages suggest: (1) the distribution of networks generated by each proposed GERGM model is restrictive, as only a small proportion of possible networks exhibited network statistics similar to those of the population network, and (2) the sampling algorithm may not be thoroughly exploring the sample space, as implied by the jagged peaks in Figure 6.6 (A - C). The appearance of the jagged peaks is a result of the sampling algorithm repeatedly rejecting the proposed network and re-accepting the most recently accepted network, further indicating that networks are being sampled in a region where the probability mass is not concentrated.

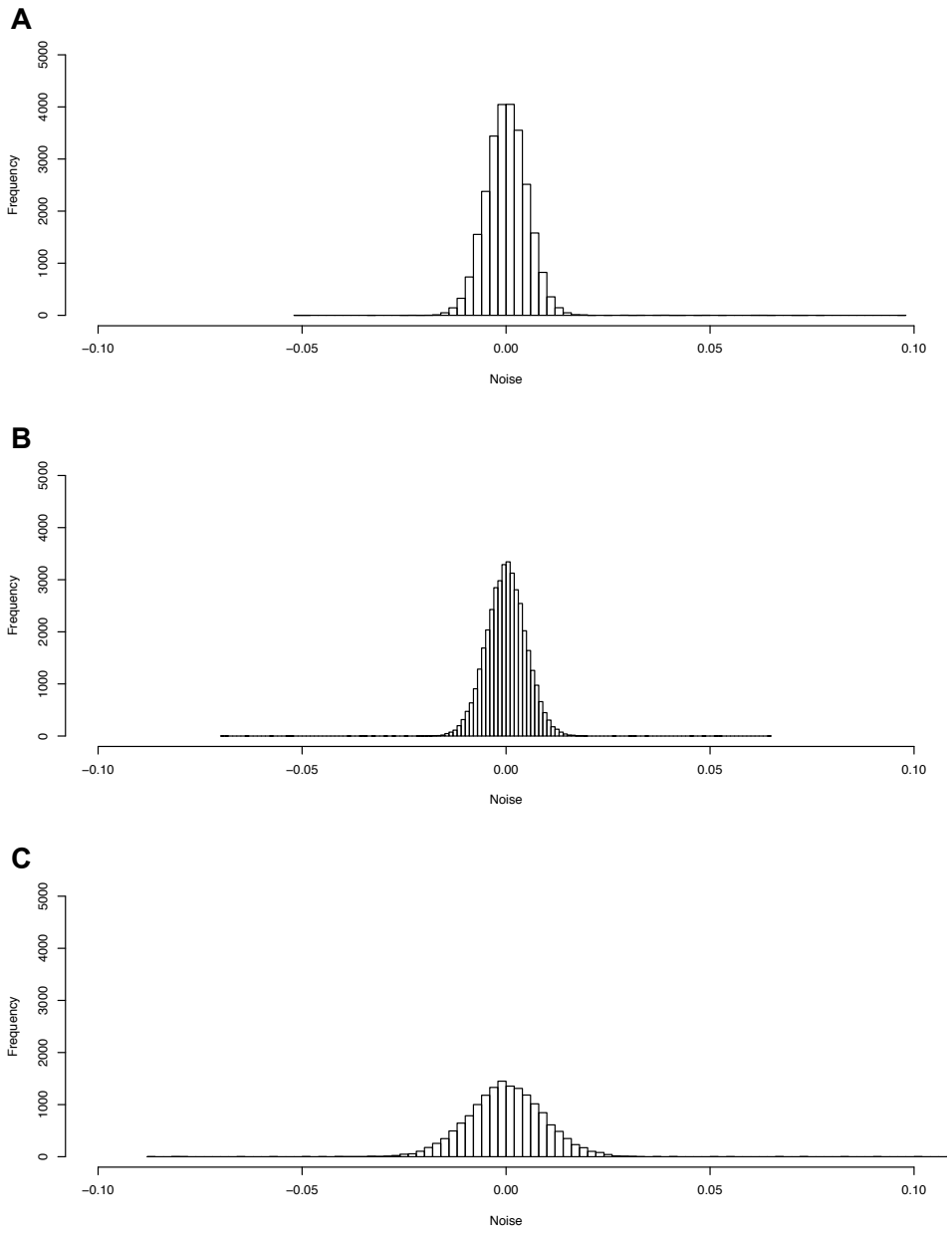

Figure 6.7. Histograms of the noise, defined as the difference in edge-weights of each pair of distinct networks in the sample of 5000 simulated networks, generated by GERGM Model 2A with standard deviation 0.005 (A), and GERGM Model 2B with standard deviation 0.005 and 0.01, respectively (B - C).

CHAPTER 7

CONTRIBUTIONS

7.1 CONCLUSIONS

Centrality measures have been utilized in the analysis of biological networks, as the measures provide rankings of influential nodes within the network and thus, a better idea of the properties, features, and sub-networks that contribute to a network's biological complexity (Breitkreutz et al., 2012; Mistry et al., 2017; Ramadan et al., 2016; Zhang et al., 2013). Yet not much thought has been published on the inherent variability of these measures. Put another way, had different tissue samples been collected and the methodology repeated in Chapter 2, would the betweenness centrality have identified the same set of essential genes? To address this question, Chapter 3 examined the variability of the betweenness centrality measure to identify genes whose role changes as tissue transitions from a healthy to diseased state through two separate edge-weight perturbation methods. Yet, although the betweenness measure was shown to be robust in the identification of structurally important genes, the large range of betweenness values that resulted from edge-weight perturbations was suggestive of instability. After formally proving that the betweenness centrality is an unstable measure, in the sense that an arbitrarily small change to edge-weights causes large fluctuations in its value, the finite stable betweenness measure was defined in Chapter 4.

To determine a distribution of the finite stable betweenness measure, and thus obtain a better understanding of its variability, a framework involving generalized exponential random graph models (GERGM) was proposed in Chapter 6. In particular, the proposed framework rests upon the assumption provided by the GERGM probability model in that

the structure of the observed network may be explained by network configurations. Because the betweenness centrality captures some global structural properties of a network (Abbasi et al., 2012; Alahakoon et al., 2011), the proposed framework to determine a distribution of finite stable betweenness focused on incorporating the measure into GERGM probability model as a network configuration. Various models to determine the distribution of the measure were proposed and compared to the distribution of finite stable betweenness of a small 20-node network constructed from a known structure.

Based upon the results of the application presented in Section 6.4, GERGM models are flexible in creating distributions of finite stable betweenness. However, the initial comparison with one possible realistic noise mechanism (refer to Section 6.4.1) that adds noise to edge-weights was not in line with the theoretical structure provided by the GERGM model. Although none of the GERGM models determined distributions identical to that from the realistic noise model in regards to shape, the proposed framework still may be useful. As previously noted, the distribution of the finite stable betweenness is dependent upon the definition of the statistic provided in the vector of network statistics $\vec{S}(x)$. If the statistic $S(x)$, for example, is modeled as the squared difference between a target and observed finite stable betweenness value, then the finite stable betweenness follows a normal distribution. Depending on the application, the statistic could be constructed in a variety of ways, and the variance of the sampling algorithm specified, so as to model the sampling distribution determined by other techniques, such as the bootstrap method.

7.2 FUTURE WORK

A more comprehensive investigation into both the flexibility of the GERGM and different mechanisms, such as the standard deviation used in the sampling algorithm, that might impact the finite stable betweenness sampling distribution are next steps in developing the GERGM framework to determine distributions of centrality measures. Additionally, although some work has already been done in the context of exponential random graph models (Handcock et al., 2003; Kim et al., 2016), a thorough investigation into the degeneracy of GERGMs, in which the estimation process fails to converge upon maximum likelihood estimates, will also be conducted.

7.3 SUPPLEMENTARY MATERIAL

The code used to generate the distributions of the finite stable betweenness in generalized exponential random graph models is provided by Durón (2019), and is based upon the code written by Denny (2018). Additionally, all figures were produced using software provided by RStudio Team (2016).

REFERENCES

Abbasi, A., Hossain, L., and Leydesdorff, L. (2012). Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks. *Journal of Informetrics*, 6(3):403–412.

Alahakoon, T., Tripathi, R., Kourtellis, N., Simha, R., and Iamnitchi, A. (2011). K-path centrality: A new centrality measure in social networks. In *Proceedings of the 4th Workshop on Social Network Systems*, pages 1–6. ACM.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106.

Bajenaru, M. L., Garbow, J. R., Perry, A., Hernandez, M. R., and Gutmann, D. H. (2005). Natural history of neurofibromatosis 1–associated optic nerve glioma in mice. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 57(1):119–127.

Bajenaru, M. L., Hernandez, M. R., Perry, A., Zhu, Y., Parada, L. F., Garbow, J. R., and Gutmann, D. H. (2003). Optic nerve glioma in mice requires astrocyte Nf1 gene inactivation and Nf1 brain heterozygosity. *Cancer Research*, 63(24):8573–8577.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics Reports*, 424(4–5):175–308.

Breitkreutz, D., Hlatky, L., Rietman, E., and Tuszynski, J. (2012). Molecular signaling

network complexity is correlated with cancer patient survivability. *Proceedings of the National Academy of Sciences*, 109(23):9209–9212.

Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., Mustonen, V., Gonzalez-Perez, A., Pearson, J., Sander, C., et al. (2015). Pathway and network analysis of cancer genomes. *Nature Methods*, 12(7):615–21.

Daginakatte, G. C., Gianino, S. M., Zhao, N. W., Parsadanian, A. S., and Gutmann, D. H. (2008). Increased c-Jun-NH2-kinase signaling in neurofibromatosis-1 heterozygous microglia drives microglia activation and promotes optic glioma proliferation. *Cancer Research*, 68(24):10358–10366.

Daginakatte, G. C. and Gutmann, D. H. (2007). Neurofibromatosis-1 (Nf1) heterozygous brain microglia elaborate paracrine factors that promote Nf1-deficient astrocyte and glioma growth. *Human Molecular Genetics*, 16(9):1098–1112.

Denny, M., Wilson, J., Cranmer, S., Desmarais, B., and Bhamidi, S. (2017). GERGM: Estimation and fit diagnostics for generalized exponential random graph models. *R package version 0.11*, 2.

Denny, M. J. (2018). GERGM: Estimation and fit diagnostics for generalized exponential random graph models. `https://github.com/matthewjdenny/GERGM`.

Desmarais, B. A. and Cranmer, S. J. (2012). Statistical inference for valued-edge networks: The generalized exponential random graph model. *PLoS ONE*, 7(1):e30136.

Dijkstra, E. W. (1959). A Note on Two Problems in Connexion with Graphs. *Numerische Mathematik*, 1(1):269–271.

Durón, C. (2019). Finite Stable Betweenness. `https://github.com/cduron1/STBTWN_GERGM/`.

Durón, C., Pan, Y., Gutmann, D. H., Hardin, J., and Radunskaya, A. (2018). Variability of Betweenness Centrality and Its Effect on Identifying Essential Genes. *Bulletin of Mathematical Biology*, pages 1–19.

Epskamp, S., Borsboom, D., and Fried, E. I. (2017). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, pages 1–18.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

Geyer, C. J. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society*, pages 657–699.

Ghosh, S., Baloni, P., Vishveshwara, S., and Chandra, N. (2014). Weighting schemes in metabolic graphs for identifying biochemical routes. *Systems and Synthetic Biology*, 8(1):47–57.

Greene, C. S., Krishnan, A., Wong, A. K., Ricciotti, E., Zelaya, R. A., Himmelstein, D. S., Zhang, R., Hartmann, B. M., Zaslavsky, E., Sealfon, S. C., et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics*, 47(6):569.

Handcock, M. S., Butts, C. T., Hunter, D. R., Goodreau, S. M., de Moll, S. B., Krivitsky, P. N., and Morris, M. (2015). Temporal Exponential Random Graph Models (TERGMs) for dynamic network modeling in statnet.

Handcock, M. S., Robins, G., Snijders, T., Moody, J., and Besag, J. (2003). Assessing degeneracy in statistical models of social networks. Technical report, Citeseer.

Hardin, J., Garcia, S. R., and Golan, D. (2013). A method for generating realistic correlation matrices. *The Annals of Applied Statistics*, pages 1733–1762.

Hegedus, B., Banerjee, D., Yeh, T.-H., Rothermich, S., Perry, A., Rubin, J. B., Garbow, J. R., and Gutmann, D. H. (2008). Preclinical cancer therapy in a mouse model of neurofibromatosis-1 optic glioma. *Cancer Research*, 68(5):1520–1528.

Hummel, R. (2012). Improving Simulation-Based Algorithms for Fitting ERGMs. *Journal of Computational and Graphical Statistics*, 21(4):920–939.

Kaul, A., Toonen, J. A., Cimino, P. J., Gianino, S. M., and Gutmann, D. H. (2014). Akt-or MEK-mediated mTOR inhibition suppresses Nf1 optic glioma growth. *Neuro-Oncology*, 17(6):843–853.

Kim, Y., Antenangeli, L., and Kirkland, J. (2016). Measurement Error and Attenuation Bias in Exponential Random Graph Models. *Statistics, Politics and Policy*, 7(1-2):29–54.

Krivitsky, P. N. (2012). Exponential-Family Random Graph Models for Valued Networks. *Electronic Journal of Statistics*, 6:1100.

Kulesa, A., Krzywinski, M., Blainey, P., and Altman, N. (2015). Sampling distributions and the bootstrap: The bootstrap can be used to assess uncertainty of sample estimates. *Nature Methods*, 12(6):477.

Lachmann, A., Giorgi, F. M., Lopez, G., and Califano, A. (2016). ARACNe-AP: gene

network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics*, 32(14):2233–2235.

Langfelder, P. and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559.

Listernick, R., Ferner, R. E., Liu, G. T., and Gutmann, D. H. (2007). Optic pathway gliomas in neurofibromatosis-1: controversies and recommendations. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 61(3):189–198.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biology*, 15:550.

Madhavan, S., Zenklusen, J.-C., Kotliarov, Y., Sahni, H., Fine, H. A., and Buetow, K. (2009). Rembrandt: helping personalized medicine become a reality through integrative translational research. *Molecular Cancer Research*, 7(2):157–167.

Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC Bioinformatics*, volume 7, page S7.

Mistry, D., Wise, R. P., and Dickerson, J. A. (2017). DiffSLC: A graph centrality method to detect essential proteins of a protein-protein interaction network. *PLoS ONE*, 12(11):e0187091.

Morris, M., Handcock, M. S., and Hunter, D. R. (2008). Specification of Exponential-

Family Random Graph Models: Terms and Computational Aspects. *Journal of Statistical Software*, 24(4):1548.

Narang, V., Ramli, M. A., Singhal, A., Kumar, P., de Libero, G., Poidinger, M., and Monterola, C. (2015). Automated identification of Core Regulatory genes in Human Gene Regulatory Networks. *PLoS Computational Biology*, 11(9):e1004504.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

Pan, Y., Bush, E. C., Toonen, J. A., Ma, Y., Solga, A. C., Sims, P. A., and Gutmann, D. H. (2017). Whole tumor RNA-sequencing and deconvolution reveal a clinically-prognostic PTEN/PI3K-regulated glioma transcriptional signature. *Oncotarget*, 8(32):52474.

Pan, Y., Durón, C., Bush, E. C., Ma, Y., Sims, P. A., Gutmann, D. H., Radunskaya, A., and Hardin, J. (2018). Graph complexity analysis identifies an ETV5 tumor-specific network in human and murine low-grade glioma. *PLoS ONE*, 13(5):e0190001.

Ramadan, E., Alinsaif, S., and Hassan, M. R. (2016). Network topology measures for identifying disease-gene association in breast cancer. *BMC Bioinformatics*, 17(7):274.

RStudio Team (2016). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 1.0.153 edition.

Segarra, S. and Ribeiro, A. (2016). Stability and continuity of centrality measures in weighted graphs. *IEEE Transactions on Signal Processing*, 64(3):543–555.

Wang, B., Pourshafeie, A., Zitnik, M., Zhu, J., Bustamante, C. D., Batzoglou, S., and

Leskovec, J. (2018). Network Enhancement: a general method to denoise weighted biological networks. *Nature Communications*, 9(1):3108.

West, J., Bianconi, G., Severini, S., and Teschendorff, A. E. (2012). Differential network entropy reveals cancer system hallmarks. *Scientific Reports*, 2:802.

Yates, P. D. and Mukhopadhyay, N. D. (2013). An inferential framework for biological network hypothesis tests. *BMC Bioinformatics*, 14(1):94.

Zhang, B. and Horvath, S. (2005). A General Framework for Weighted Gene Co-expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology*, 4(1).

Zhang, X., Xu, J., and Xiao, W. (2013). A new method for the discovery of essential proteins. *PLoS ONE*, 8(3):e58763.