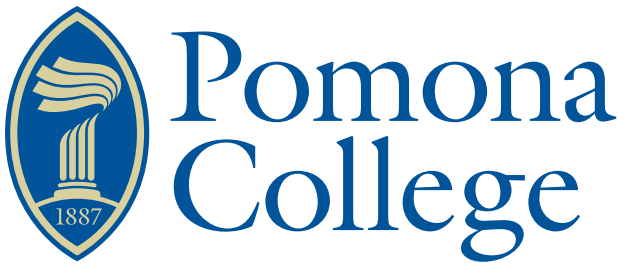


Senior Thesis in Mathematics

Analyzing Centrality in Complex Gene Networks

Author:
Dylan Quantz

Advisor:
Dr. Jo Hardin

Submitted to Pomona College in Partial Fulfillment
of the Degree of Bachelor of Arts

April 22, 2016

Abstract

Our research may help us understand what genes are most central in cancer patients. Our research is focused on trying to find the gene most connected in the gene network of children that have a form of brain cancer. We try and do this by creating a network of gene interactions. To aid us in solving our problem, we used a form of resampling, called bootstrapping. Bootstrapping is a powerful strategy when it comes to understanding gene interactions because our statistic of interest is a centrality measure and it does not have a known sampling distribution. Bootstrapping gives us a reliable estimate of the sampling distribution of the estimator of interest, and in our case the true gene interaction present in patients with these brain tumors. In this paper, we try to show which gene is most central in cancer patients using the process of bootstrapping.

Contents

Acknowledgement

I would first and foremost like to offer my deepest gratitude to my outstanding thesis advisor, Professor Jo Hardin, who without I wouldn't have been able to complete this endeavor. She provided great support whenever I ran into a snag in my R code, which was often, and offered timely feedback throughout the journey. Thank you, again!

Chapter 1

Introduction

The use of network analysis in biology, specifically genetics, could be critical in helping doctors and scientists understand more about debilitating diseases by understanding how the genes interact in affected people. One of the debilitating diseases that has received the most research has been cancer because it is incurable and extremely lethal. In 2016, there will be an estimated 1,685,210 new cancer cases diagnosed and 595,690 fatalities due to cancer throughout the United States of America [8] because of these facts, finding a cure would be a momentous step for future generations . Thus, we will be looking at the interactions of genes in humans both with tumors (tumor group) and without tumors (normal group). Through this analysis, we will hope to find some noteworthy gene that may increase the understanding of the disease mechanism.

A network, for our purposes, pertains to a complex interacting system that can be represented as a mathematical graph. For this case, a network is helpful because it can provide a representation of complex gene-gene interactions. Our goal is to find the most central or influential genes that interact in the tumor group of humans. In network analysis there are different indices that can be used to measure centrality which will help us analyze gene interactions in the tumor group. Centrality or importance in a network can be thought about different ways using different measures. Conceptually, centrality is fairly straightforward: we want to identify which genes are in the center of our gene interaction network.

For this analysis we will be using two main concepts of centrality to help us understand the data: closeness and betweenness. Closeness and betweenness are the two centrality indices that will help us understand what our data is conveying. For the sake of our study, closeness is a measure of the degree to which a gene is near all other genes in our network and betweenness is a measure of the degree to which one gene serves as a bridge to others. These indices serve as point estimates that tell us about the point estimates in our dataset but they don't give all the information about gene interactions. We can find out more information, if we know the variability of these point estimates. Knowing the variability

can then help us see which closeness is biggest.

In our case, the way we create variability within our network is through the process of bootstrapping. Each bootstrap sample we take will have a new network and corresponding centrality indices. All the bootstrapping networks taken together will provide us with a clearer picture of which genes keep appearing as most central in all these networks. If we are able to take a large number of bootstrap networks then we will be able to see if the genes we found most central in our sample network are actually important in the tumor group or just an anomaly. This process of increasing the variability of our networks by bootstrapping will allow us to be able to say something meaningful about gene interactions in humans.

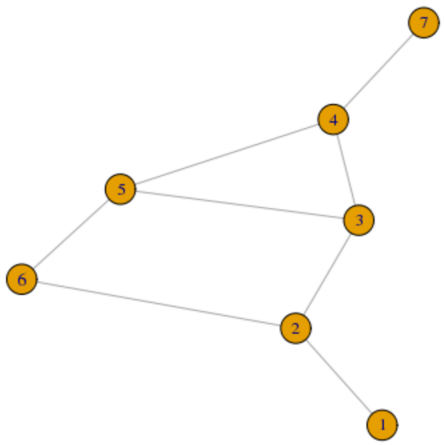
The sample we will be using contains 8150 genes in juvenile patients with tumors from pilocytic astrocytomas (PA), a form of pediatric brain cancer [9]. The first step we will take to is to find how correlated each gene is with all the other genes in the network. Once we have done this we will use the correlations to measure distances between all the genes, where the highly correlated genes will have the smallest distances between itself and other genes. The distances will then be used to help create the gene interaction network as they will serve as the edges between genes. Once we have a network of our 8150 genes and their corresponding centrality values, we can create bootstrap networks and get corresponding centrality values for the bootstrapped genes. These centrality measures will help us quantify how central each gene is in children with pilocytic astrocytomas.

Chapter 2 will include the background on the centrality indices and the practice of bootstrapping which will be used throughout the paper. Chapter 3 will consist of the results regarding which genes are most central in our gene network. Finally, Chapter 4 will be the conclusion which will sum up the paper.

Chapter 2

Mathematical Relationships

2.1 Networks



Let's use this simple network to understand the basics of what a network is. This graph consists of 7 nodes and has 8 edges connecting these nodes together. A node in our analysis will be a gene so we can picture each letter 1-7 corresponding to a certain gene and each edge is an interaction a gene shares with another.

Another key concept in network analysis is the idea of a connected network. A connected network is one in which every node can be reached by another node. The network above is clearly connected because every node can reach every other node using the connecting edges. It is very hard to analyze a network that is not connected so when we start talking about the network we will be analyzing it will be a connected network of genes.

Another idea that the network above can illustrate is the idea of a geodesic. A distance

between two nodes can be understood as the number of edges that connect nodes along a certain path. The shortest of these paths connecting two nodes is referred to as a geodesic. Geodesics are important when it comes to understanding how we calculate the centrality indices. Looking at the network above we can see that there are multiple paths that connect 5 and 2. Let's find the geodesic(s) connecting these two nodes. We can see that there is the path 5-6-3-2 that connects 5 and 2 which includes three edges. There is also path 5-3-2 which includes two edges. The last path that we can see is 5-6-2 which also includes two edges. We can see that out of the three paths connecting these two nodes the geodesic corresponds to the paths with two edges. Therefore, there are two geodesics connecting 5 and 2! The example shows that there can be multiple geodesics connecting two nodes.

The network is also an undirected network where each edge is bidirectional (information can travel both ways across the edge). The network we will be analyzing will also be undirected.

2.2 Centrality Metrics

There are many different indices that can be used to calculate centrality, including degree, betweenness, closeness, eigenvector, power, information, flow, and reach [6]. In our analysis, we will be using two of the most common centrality indices: closeness and betweenness. These indices will help us find which genes are the most central. The first is closeness and the second is betweenness. These two indices calculate centrality using different methods and understandings.

2.2.1 Closeness

The closeness index focuses on which nodes are relatively close to all other nodes in the network. The index is measured based on the inverse of the sum of all the shortest paths between a node and all other nodes in the network, where shortest is the smallest number of edges it takes to get to all other nodes from the node of interest in an unweighted network [3]. The process of calculating closeness in a weighted network is a little different because the calculations are dependent on the distances between nodes. Instead of just adding up the total number of edges, we have to get the sum of all the distances it takes to get from one node to all the other nodes. This can be expressed mathematically in the following way:

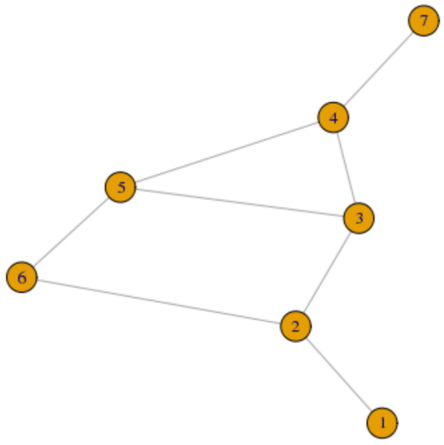
$$C_C(i) = \frac{n-1}{S(u)}$$

where n is the number of nodes, i is the node of interest, and S(u) or the distance function can be expressed as

$$S(u) = \sum_{j=1}^{n} d(i,j) \tag{2.1}$$

This distance function can be understood through the idea of shortest path distance. As we calculated before, distance in a network can be explained as the number of edges between two nodes. In an undirected network, the shortest path distance, d(i,j), is the number of edges in the shortest path between nodes i and j in a network [3]. All of these distances d(i,j) can be calculated among all the different vertices in the network and can be arranged in a matrix. These are compiled into a distance matrix. Thus, we can understand S(u) to be the sum of all entries (n) of a row/column of the distance matrix for a specific node. We can also understand this as the total distance of a node. Understanding this will help us understand the equation of calculating closeness.

Let's do an example to make the idea of calculating closeness far more clear. Let's again use a simple network to convey the ideas.



This is the same network from above and it has 7 nodes and 8 edges connecting the nodes. It is also an undirected, connected network. Let's calculate and compare the closeness indices for nodes (3) and (7). Let's start with node (3) and from the definition we want to know the shortest path between (3) and all the other nodes in the network. The distance between node (3) and (5), (2), (4) is one edge away. The distances between (3) and (7), (1), (6) are all two edges (eight total edges) away. The total distance it takes for (3) to get to all the other nodes in the network is the sum of total edges to get to all other nodes which in this case is $6 + 3 = 9$. If we were to have used the distance function and created a distance matrix to help us solve the problem, we would have gotten that $S(3) = 9$ as well. From the equation of calculating the closeness index, we can see that all

we need to know is n-1 to be able to solve for the closeness value of node (3) so n-1 will be equal to 6. Therefore our closeness value for node (3) is $\frac{2}{3}$.

Let's compare this to the closeness value for node (7), where again we want to know the shortest path between (7) and all the other nodes in the network. The distances between (7) and (4) is one edge away. The distance between (7) and (3), (5) is two edges (four total edges) away. The distance between (7) and (6), (2) is three edges (6 total edges) away. Last but not least, the distance between (7) and (1) is 4 edges away. The total distance it takes for (7) to get to all the other nodes in the network is going to be the sum of total edges to get to all the other nodes which in this case is $1 + 4 + 6 + 4 = 15$. Now all we need to know what the value of n-1 is and it will be the same as before so it is equal to 6. From the equation, we can see that the closeness value for node (7) is $\frac{2}{5}$.

Doing this example shows us that the higher the closeness value the closer it is to all the other nodes. This means that the node that corresponds to the highest closeness value will have the highest centrality in the network. We can clearly see from our picture that (3) is more central than (7) which is why it had a higher closeness value than (7).

The networks that we will be using in the analysis, however, will not be unweighted like this example but the process of calculating the closeness values will not change drastically. If we again use the same network, we can understand how we would calculate closeness for a weighted network. Let's look at calculating the closeness value for node (3) but this time each edge will be weighted. The edge connecting (4) and (7) will have a weight of .80, while the edge connecting (2) and (1) will have a weight of .70 and all the other edges will have a weight of .50. Now, just as in the unweighted network we will find the total distance it takes for node (3) to get to all the other nodes. This time instead of just summing the number of edges, we will be summing the corresponding weights which in this case is $.50 + .50 + .50 + 1 + 1.30 + 1.20 = 5.00$. Now the n-1 value will be the same as from the unweighted so it is 6. Thus the closeness value for node (3) in this weighted network will be $\frac{6}{5}$. We can compare the values of the closeness values from the unweighted network and the weighted network to see that there is a difference in values but the process of getting the values stays pretty much the same. Now that we understand how we are calculating closeness in our data, let's move on to see how we calculate betweenness.

2.2.2 Betweenness

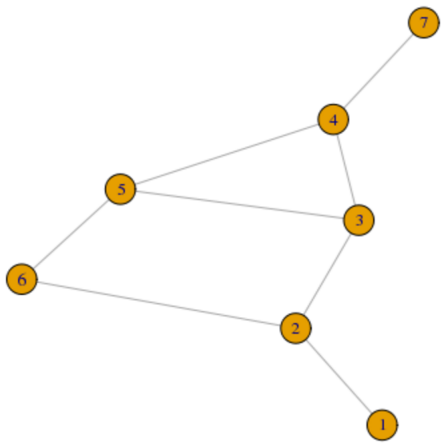
The betweenness index focuses on the relative importance of a node in the communication between pairs of nodes. The central nodes who facilitate the most interactions between other nodes are understood to be the most between. This means that if you were to take out the most central node between many other nodes than it could cause these latter nodes to communicate less or even not at all! Betweenness can be expressed in the following way:

$$C_B(i) = [\sum_{j<k} \frac{g_{jk}(i)}{g_{jk}}] \tag{2.2}$$

- where i is the node of interest
- g_{jk} = the number of geodesics, or shortest paths, from node j to k
- g_{jk}(i)= the number of geodesics, or shortest paths, between nodes j and k that go through node i
- $i \neq j \neq k$

Understanding the idea of betweenness is much harder to comprehend than closeness. The most intuitive way of defining betweenness is by considering that the information going from one node to another travels only through the shortest paths connecting these nodes, or in other words serves as a bridge between nodes.

Looking at an example will help us understand how one calculates betweenness.



At first glance of the network, we can see that there are 7 nodes and 8 edges connecting the nodes. The network is also undirected and connected which is ideal for this example. Let's calculate and compare the betweenness indices of nodes (1), (2) and (4) using the equation we defined above. Let's start with looking at node (1) and see what its betweenness index is. From the equation, we are looking for number of geodesics (shortest paths) that are between any two nodes that pass through node (1). But there are no two nodes that go through node (1) because it is on the outer edge of the graph and only connects to

one other node which means it cannot facilitate any interactions between any other nodes. Therefore, we would say that the betweenness index for node (1) is 0.

Let's look at node (4) and find out what its betweenness value is. We can see that node (7) passes through node (4) to get to all the other nodes in the network but because it is on the edge of the graph and only has one edge, we can see that the number of geodesics between node (7) and all the other nodes is going to be equal to the the total number of geodesics between node (7) and all the other nodes. Now looking at the betweenness equation, we can see that node (4) serves as the only link between node (7) and all the other nodes in the network which means that the betweenness value is equal to 5.

Lastly let's look at node (2). We can see that node (1) passes through node (2) to get to all the other nodes in the network but because it is on the edge of the graph and only has one edge, we can see that the number of geodesics between node (1) and all the other nodes is going to be equal to the the total number of geodesics between node (1) and all the other nodes. Looking at the equation given above we can see that this means that node (2) serves as a bridge between node (1) and five other nodes and node (1) must pass through node (2) to get to all other nodes so the betweenness value is equal to 5. There is one more node that has a geodesic that passes through node (2) and it connects node (3) and node (6) but there is another geodesic that connects node (3) to node (6) which travels through node (5) so the betweenness value is equal to $\frac{1}{2}$. Thus the total betweenness value is 5.5. We would use this same method to find out the betweenness values for the rest of the nodes and would find out that node (2) is the most between node in this network because it connects the most nodes.

Again let's look at how we would calculate betweenness for a weighted network. Let's look at the betweenness of node (3) and give the edges from (3) and (5) weight .80, from (5) and (6) weight .25, from (3) and (2) weight .15 and from (2) to (6) weight .60. Looking at the network now, we can see that there are no longer two geodesics that connect nodes (5) and (2) but only 1 because of the weighted edges. Thus, we can see that the betweenness value for (3) would now be 5 because the shortest path passing through node (6) [.85] is smaller than the shortest path passing through node (6) [.95]. Now we have a clear understanding of how we will calculate closeness indices for our data, let's move onto how we weighted our edges.

2.2.3 Centrality in terms of Correlation

The data that I am using to construct these network is based on correlation where correlation is calculated like we did in our introductory statistics class:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \tag{2.3}$$

Each node (gene) has a corresponding correlation, r, between itself and all other genes in

the network, which we know to be $-1 \leq r \leq 1$. The next step we took was to square each of the correlation values so that there were no negative values anymore. We then wanted to find the distance which was a measure based on corresponding distances between genes, i.e. the smaller the distance the more they interacted.

$$d(i,j) = 1 - r^2 \tag{2.4}$$

This distance was measured for each gene interaction. It is understood that a gene that has a short distance between itself and another gene is highly correlated and on the other side that if there is a large distance between two genes than the genes have very little correlation.

Using these distances, we are able to construct networks with edges whose length depends on the distance between the two genes. When we calculate the closeness and betweenness value, we use these edges. In the case of betweenness, the corresponding distances matter in finding the geodesics. In a weighted network there is usually only one gene serving as the strongest bridge connecting two other genes because of the corresponding distances. This is because one path will have a shortest distance so it will be considered the geodesic. Thus, in weighted networks the betweenness values are usually whole numbers. This contrasts how we calculate the closeness indices, however. These calculations are dependent on the corresponding distances between genes. When we calculate the closeness value for a certain gene we are not able to just count up the edges like in our example above. This time we have to add up the distances that it takes to get to every other gene from the gene of interest. This means that genes that have a small distance between itself and other genes will have a large closeness value which makes sense because it is more central to the network.

Understanding how we calculate the centrality indices using correlation in specific distance is critical to our gene network.

2.3 Bootstrapping

The bootstrap method introduced in Efron (1979) [2] is a very general resampling procedure for estimating the distributions of statistics based on independent observations. Bootstrapping is a very powerful technique because it can help us get around previous obstacles like having normal data or a large sample size. This method can be used for many different statistics even ones for whom the sampling distribution is not easy to characterize mathematically. The reason we bootstrap is to get a sampling distribution for a statistic. The process can be done using software that can handle the heavy computation necessary for all the resampling. Advancements in computing power of many desktops and laptops has made the process of bootstrapping much easier and in the process making it one of the

primary ways to do statistical inference on statistics with unknown sampling distributions.

The process of bootstrapping is an attempt to find the sampling distribution using only a single sample!

The first step in the bootstrapping process is resampling. Instead of many samples from the population, this method requires many resamples of the same size from the original sample with replacement. Sampling with replacement is the idea that we draw an object from the original sample but before we draw the next one, we replace the first one because if we sampled without replacement than we would have the same sample every single time. For each resample, we calculate the statistic of interest.

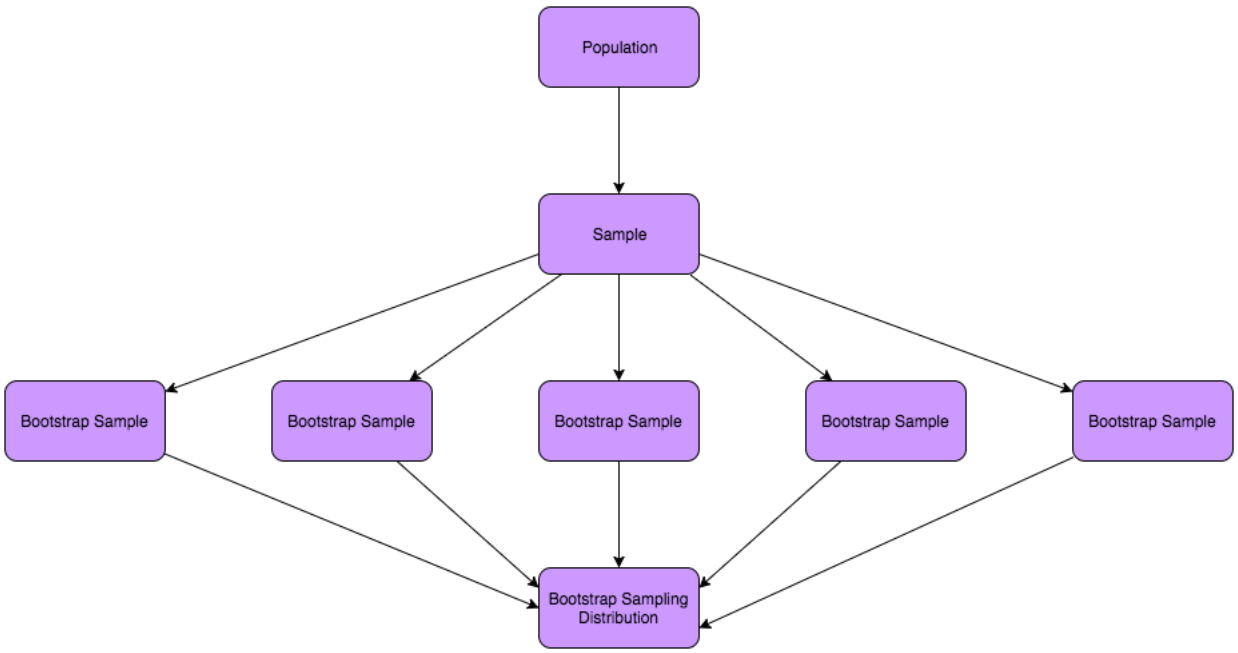
The second step in the process is creating the bootstrap distribution. The bootstrap distribution is created using the information from all the resamples and their corresponding statistics. We can use the bootstrap distribution to give us information about the sampling distribution of the statistic from the actual population.

The process of bootstrapping does not contain many steps, as you can see above but it can take a long time to create all the bootstrap samples because ideally you would want the largest value, which would be ∞, of bootstrap sample that you could to get the best idea of what the sampling distribution looks like.

There are some limitations to bootstrapping, however, that we must address as well. If the sampling distribution is not close to the true sampling distribution then our bootstrap will not provide a good estimate of the sampling distribution. Thus, we are assuming that when we take a bootstrap sample that our bootstrap procedure will work and produce a sampling distribution of a statistic that is similar to the true sampling distribution of that statistic.

The power in bootstrapping is not in constructing sampling distributions that we have a lot of theory about such as the mean or median but in understanding the spread of a statistic that we don't know much about, in our work our statistics are centrality measures of betweenness and closeness which we know very little about. This is because in most cases the bootstrap distribution has the same shape or close to as the sampling distribution if it meets some requirements. These requirements are that there is a random sample of data from the population, that the statistic converges to the parameter in large samples and that there is a big original sample.

Figure 2.1: A flowchart of the Bootstrapping Process



Consider the problem of estimating the variability of means by the bootstrap method. If we view the observations $x_1, x_2, \ldots, x_n$ as values from the sample, then we can create a sampling distribution of the mean using bootstrapping. This will help us calculate the variability of the sample mean (or $\bar{x}$).

In the bootstrapping method we know that we will randomly pick, with replacement, n values from our observations, $x_1, x_2, \ldots, x_n$, and we will do this many times to create a bootstrap distribution of $\bar{x_b}$.

Using this distribution $\bar{x_b}$ and the assumptions we make about bootstrapping, we assume that $\bar{x_b} \to \bar{x}$ if our sample is arbitrarily large and with this knowledge we get to learn about the variability and shape of the distribution of $\bar{x}$.

2.3.1 How we Bootstrap in the Network Context

We have talked about the idea of bootstrapping in the terms of sampling distributions but how does bootstrapping work in a network context. Our original data included a matrix of correlations, where each gene had a correlation with every other gene in the sample. Using this matrix of correlations, we created a distance matrix which helped us create an initial network. This network contained 8150 genes and edges corresponding to each of the distances. The process of creating bootstrap networks starts with the original observation

and we resampled along each column with replacement. This resampling created a new matrix of observations which we then found correlations for. We then turned this correlation matrix into a distance matrix which we then used to create a bootstrap network. For our research, we created 100 different bootstrap networks which we used to find the most central genes by using the closeness and betweenness index. The bootstrapping process helps us to get a good estimation of the centrality

Chapter 3

Results

3.1 Analysis of Data

The first step to understanding the subsequent results is looking at the data we are looking at. The data we have included gene interactions of 8150 genes in 58 people (49 people with pilocytic astrocytomas and 9 people without). We created subsets of this data and split them into normal and tumor samples. For our analysis, we only looked at the 49 people with pilocytic astrocytomas (tumor sample) which means an extension of this research would be to look at the 9 people without pilocytic astrocytomas (normal sample) and compare the differences. We used this tumor sample of the 49 patients to create a correlation and distance matrix that was 8150x8150 (each genes correlation with every other gene) which we used for all of our analysis. Now that we know a little more about the data that we are using, let's look at our results!

3.2 Calculating Centrality Indices

3.2.1 Closeness

Looking at just the rankings of the top genes according to the betweenness and closeness calculations provides a lot of information. We can look at the closeness and betweenness values for each gene to see how the centrality of the gene differs across different bootstrap samples. Let's first look at how much the closeness values differed across all bootstrap samples.

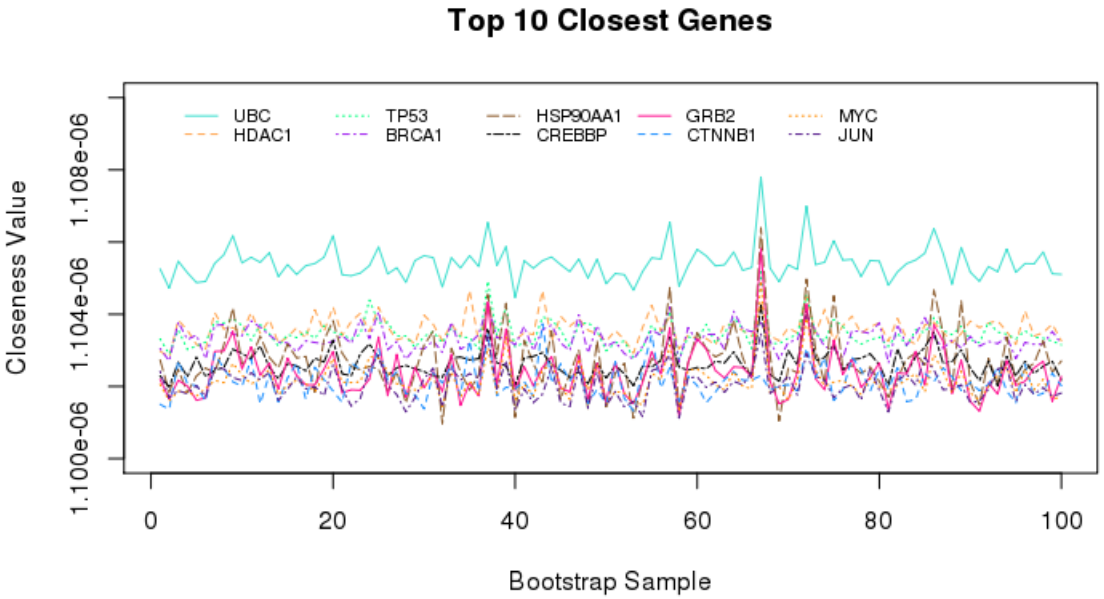


Figure 3.1: Line graph of the top 10 genes and their corresponding closeness values for each bootstrap sample

Looking at the Figure 3.1, the main conclusion that can be seen is that UBC is the closest gene in each of the bootstrap networks. This graph also shows how much greater the UBC closeness value is as compared to all the other top 10 genes, which can be seen by how much greater the UBC closeness value is than the second highest for every sample. The rest of the top 10 genes are more intertwined and not many conclusions can be reached with respect to the significance of the ranked order by looking at the other nine closeness values. One other interesting piece of information that can be seen by looking at the graph is that the closeness values are extremely miniscule with values such as $1.04x10^{-6}$! This, however, makes intuitive sense because of the calculation of the closeness value for a gene of interest. The closeness value has the sum of all the distances in the denominator so as the networks become more complex (more nodes and/or more edges) the distances increase. This results in the closeness values getting smaller and smaller. The most interesting part of the analysis is that these extremely small values do not seem to effect the top ranked results. Thus, in our data collection and analysis we are not scared of the small values for closeness.

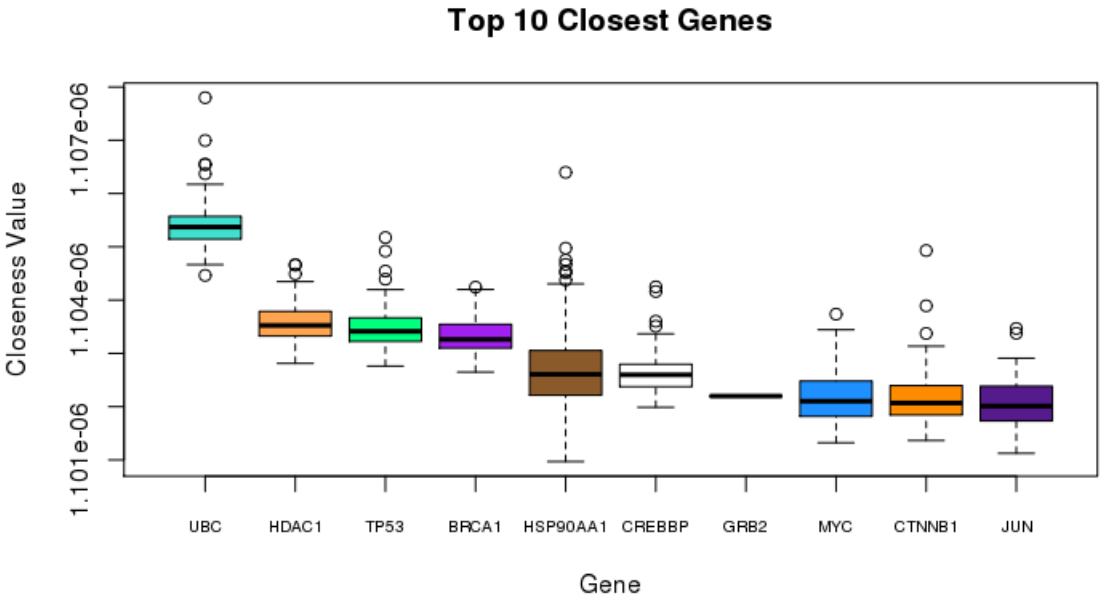


Figure 3.2: Boxplot for each of the top 10 closest genes

Looking at Figure 3.2, we can gather more information about the rest of the top 10 genes not including UBC. We can clearly see that the threesome of HDAC1, TP53 and BRCA1 are, on average, higher than the rest. We can see that by analyzing the black bars in each of the boxplots. These black bars tell us the median closeness value for each gene. The black bars of these three genes are pretty close to each other; HSP90AA1 begins a discontinuous drop of the average closeness value for the rest of the genes. Another characteristic we can see by looking at the Figure 3.2 is the variability in the estimates of the closeness values by looking at the whiskers or lines extending from the box. Looking at the variability of HSP90AA1 we can see that its variability is pretty large with its whiskers extending far more than the other genes. This variability is contrasted by looking at GRB2 whose variability is extremely small with its whiskers not even visible. The variability outside of these genes, appear to be quite similar which show that the closeness values seem pretty stable overall.

3.2.2 Betweenness

Now let's shift our attention from closeness to betweenness. We will again look at a line graph and boxplots of betweenness values for different genes.

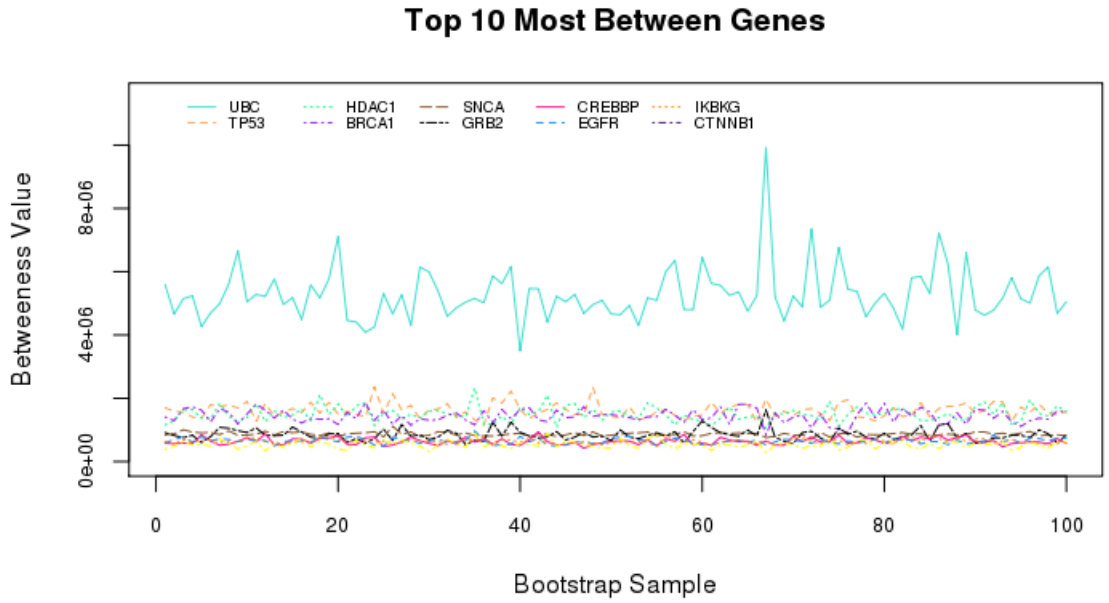


Figure 3.3: Line graph of the top 10 genes and their corresponding betweenness values for each bootstrap sample

We can see from looking at Figure 3.3 above that UBC has a higher betweenness value for every single bootstrap network but the magnitude of the gap is what is surprising. UBC has a betweenness value which is at least four times greater than the next highest gene! It is again very hard to glean any information from any of the other top genes because the UBC values are just so much larger than them. Another feature that can be seen is that the betweenness values are much larger in magnitude than the closeness values.

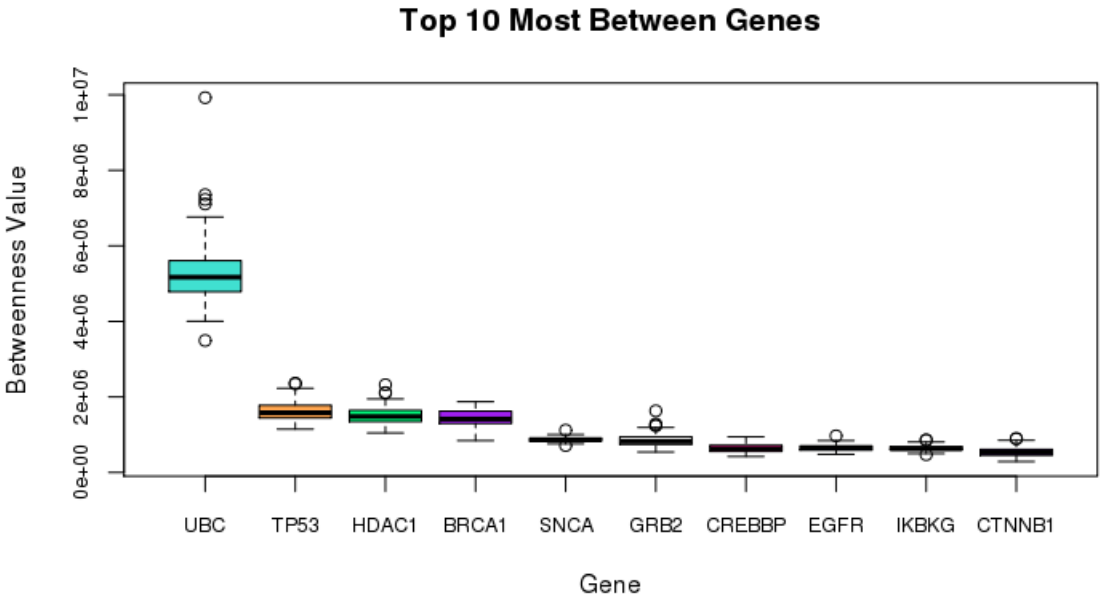


Figure 3.4: Boxplot for each of the top 10 most between genes

However, this time when we look Figure 3.4 for the top between genes, we are not able to really gather information to differentiate the rest of the top 10 genes. Again we can clearly see how much larger the betweenness values are for UBC than all the others. None of the other top 10 genes have a higher betweenness value than the lowest UBC value recorded! Therefore to try and get more information about the other genes, let's plot them without UBC.

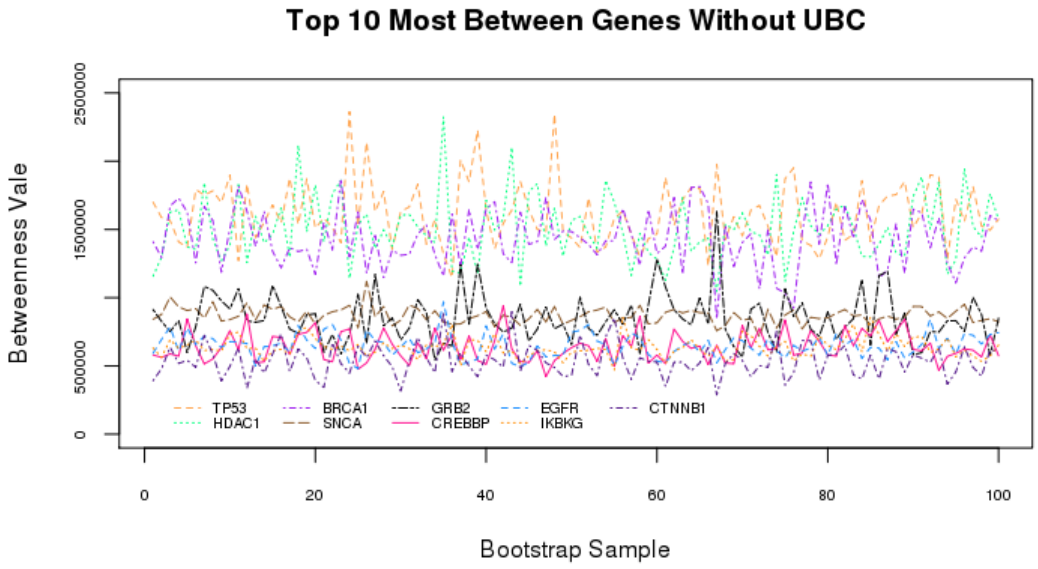


Figure 3.5: Line graph of the top 10 genes without UBC and their corresponding betweenness values for each bootstrap sample

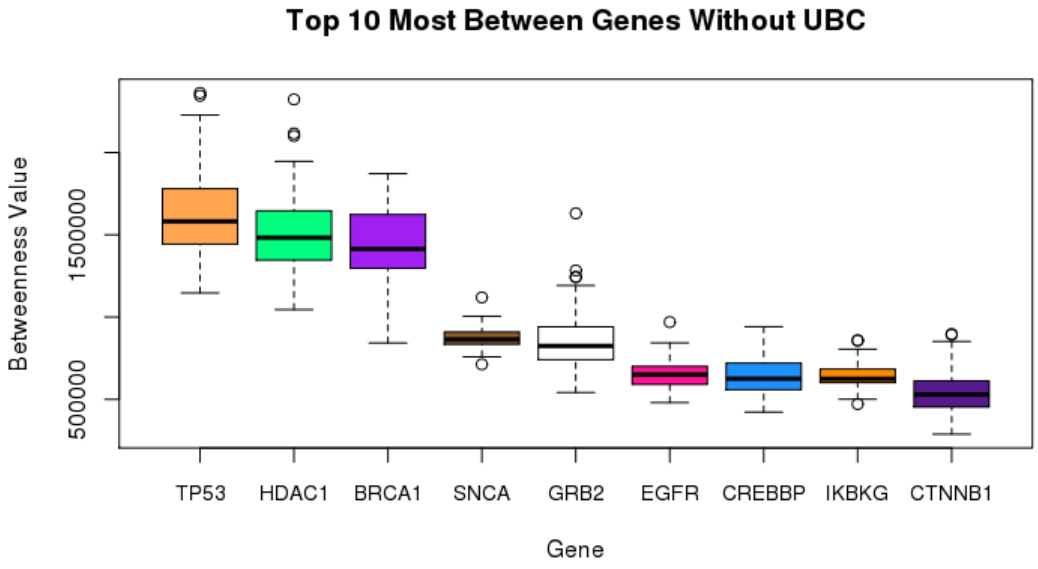


Figure 3.6: Boxplot for each of the top 10 most between genes without UBC

Now looking at the Figure 3.5 and Figure 3.6, we can start seeing some differences between the betweenness values of the rest of the genes minus UBC. We can start to make out a clear gap between the fourth most between gene and the rest of the other genes. The limits of the y-axis are about a quarter as big now than they were in Figure 3.3 and Figure 3.4 with UBC, which is why we are finally able to say something about the other genes. The same genes are in 2-4 for when looking at both closeness and betweenness which is quite encouraging.

3.3 How to Visualize Closeness and Betweenness

We have seen the definition of closeness and betweenness but what does it mean for a gene to be more central than another gene. Let's look at some visualizations to help us understand.

3.3.1 Closeness

Let's try to figure out what it means for UBC to be the closest gene in our network of 8150 genes. We are measuring the distance based on how correlated the genes are. A large correlation therefore means a short distance. To try and show that visually, I was not concerned with the first layer of genes that UBC was highly correlated with but the second layer.

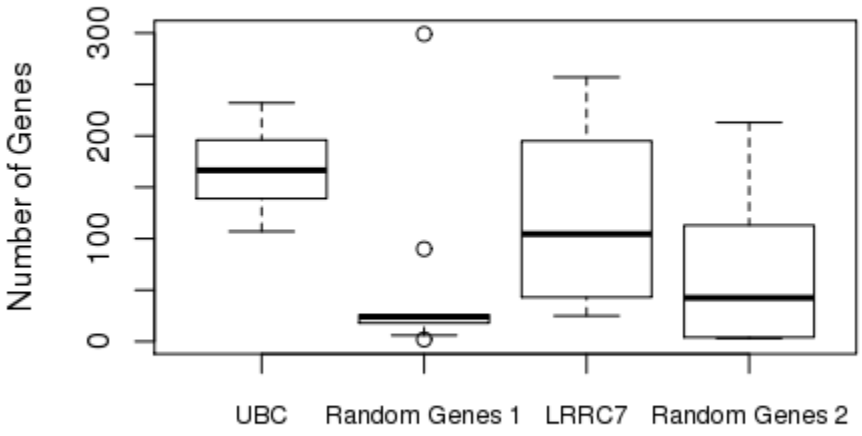
I will do this by following the following algorithm:

1. Set the gene of interest (either UBC or LRRC7)
2. Find top 10 most correlated genes with gene of interest (again either UBC or LRRC7)
3. Find all the genes with a correlation greater than .50 for the genes in the top 10 list from step (2)
4. Count number of genes from step (3) for each of the top 10 genes from step (2)

Looking at Figure 3.7, we can see that the number of genes secondarily highly correlated with UBC had a median value of around 160. I contrasted this boxplot with a boxplot in which I chose random genes (its process starts at step (3)) to see how many genes they were correlated with at a cutoff of .50. This boxplot shows that the median value appears to be around 30 for the random genes (Random Genes1). This comparison shows how close UBC is to a second layer of genes compared to the initial (first layer) correlations of genes

with randomly selected nodes, with the second layer of genes being far more enumerative based on the cutoff of .50. This second layer for UBC contains far more genes on average than the random genes which means UBC is not only more correlated on the first layer but more importantly the second one too.

Figure 3.7: Boxplots for Highly Correlated Genes with the Second Layers of both UBC and a Random Gene (LRRC7) contrasted with the Boxplots of First Layer Correlations of Random Genes



We extend this analysis to look at the top 10 most correlated genes with a random gene that follows the same algorithm above as UBC, which in this case turns out to be LRRC7, and then I followed the same process of finding out how many genes that the 10 most correlated with LRRC7 are highly correlated with at at a cutoff of .50. Looking at Figure 3.7, we can see that the genes secondarily highly correlated with LRRC7 had a median value a little greater than 100. I then again contrasted this boxplot with a boxplot in which I again chose random genes (its process starts at step (3) again) to see how many genes they were correlated with at a cutoff of .50. This boxplot has a median value of around 50 for the random genes (Random Genes2). I followed the same process to get both lists of Random Genes1 and Random Genes2. Each lists of Random Genes is made up of different genes and looking at the boxplots in Figure 3.7 we can see that their is far more variability in Random Genes2. This shows the natural variability in the correlations of genes! All of

this shows that the second layer of UBC has more genes on average that are closer to UBC than the second layer of a random gene. The closer second layer of UBC shows what it means for it to be the closest gene in the network in that the distances between it and all other genes especially in the first two layers is far less than for other genes.

3.3.2 Betweenness

Now let's figure out what it means for a gene to be the most between gene in a gene network. First we will look at Figure 3.8, a network of the top 30 most between genes. I created this network by setting a lower bound on the betweenness value to limit the number of genes in the simplified network. I wanted to limit the number of genes in the network because if there are far too many it becomes extremely convoluted and you can't glean any information from it. The bigger the size of the node the larger the betweenness value is. As we can see in Figure 3.8 UBC is the largest node and is very connected to many other genes. The edges of the network are pretty condensed around UBC.

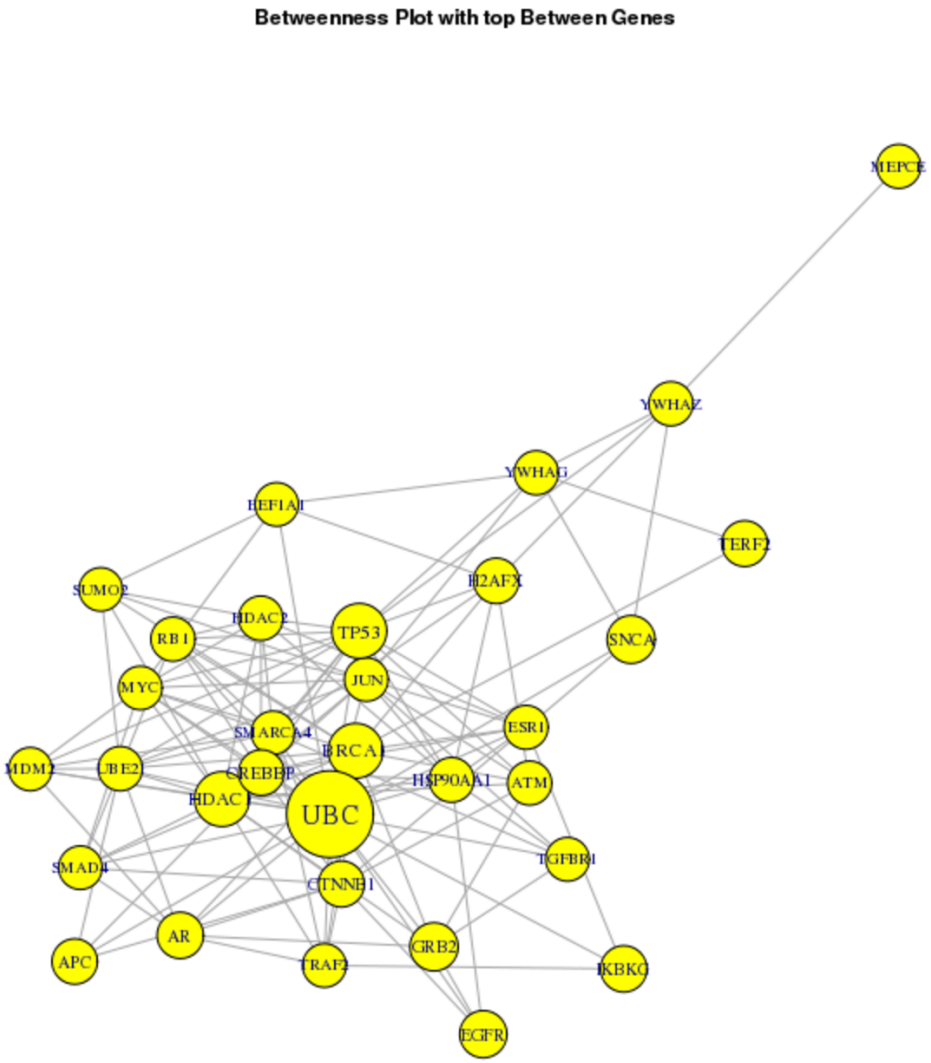


Figure 3.8: Simplified Network of the 30 Most Between Genes

Now if we look at Figure 3.9 without the top 4 most between genes (UBC, HDAC1, TP53 and BRCA1), we can see how much simpler the network is. As we can see without UBC in the network, the density of edges is far less and the most between genes are not centered in the middle like they were before. This is emblematic of how critical UBC is to the network, it provides a quicker path for many genes to interact.

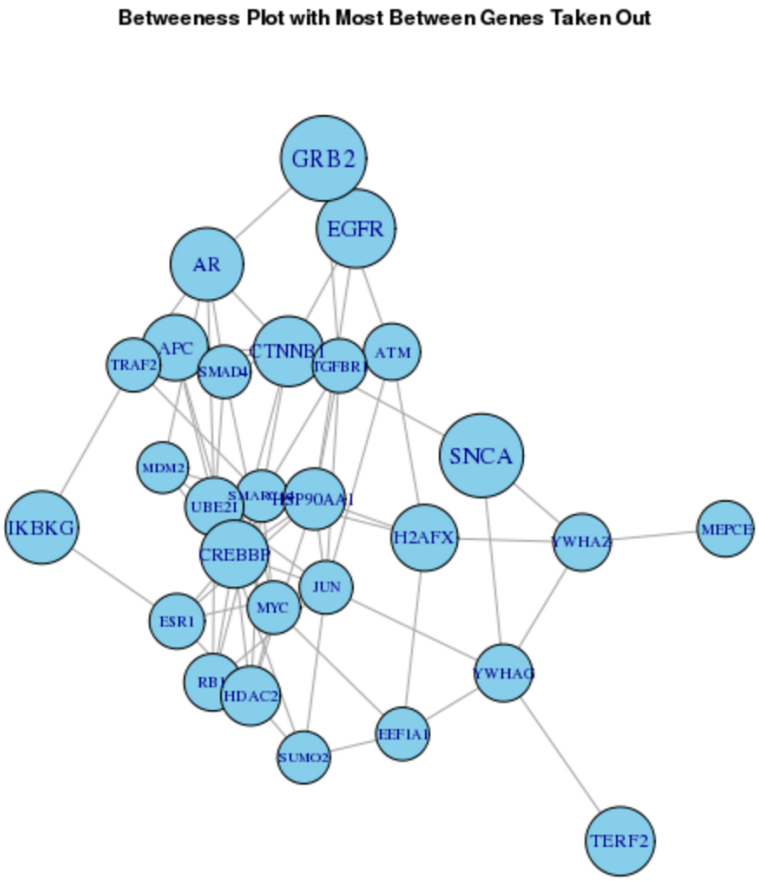


Figure 3.9: Simplified Network with HDAC1, BRCA1, UBC and TP53 removed

All of the genes are still connected in Figure 3.9 without UBC but just compare HDAC2 between the two networks. Look at Figures 3.10 and 3.11 to see HDAC2 highlighted. There are many edges connecting HDAC2 to other genes in the full network but if you look at the simplified network there are far fewer edges connecting it. This is a great example of what the idea of betweenness is and how it is represented!

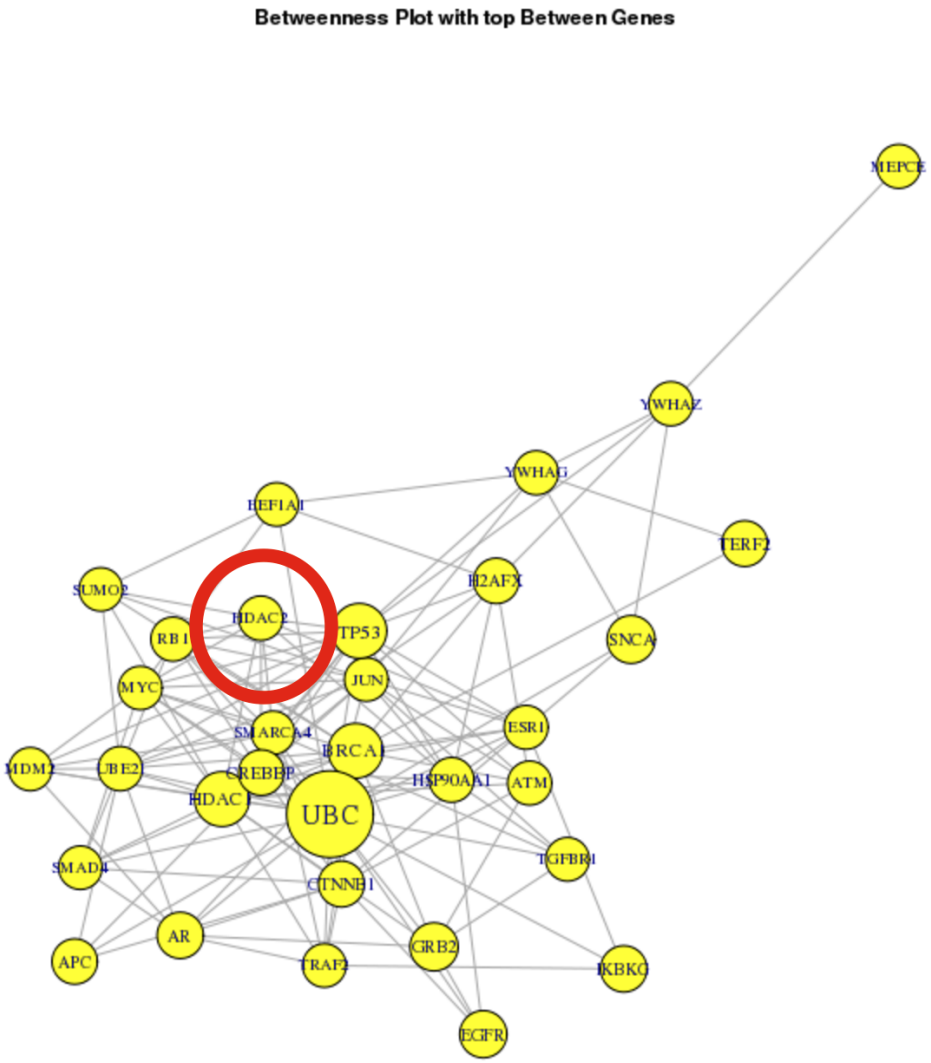


Figure 3.10: Simplified Network of the 30 Most Between Genes with HDAC2 Highlighted

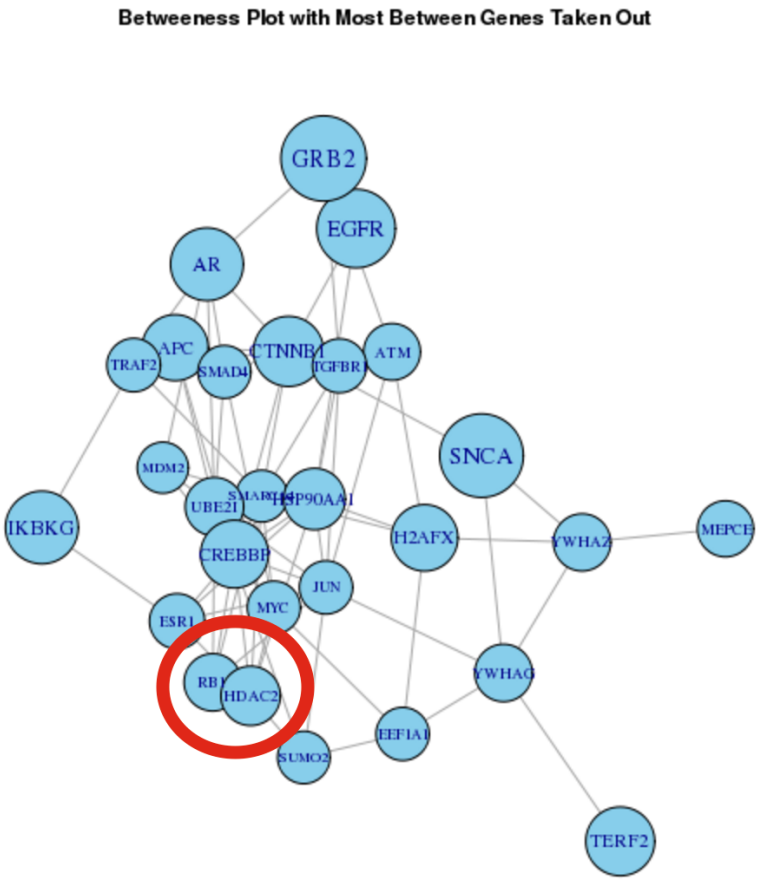


Figure 3.11: Simplified Network with HDAC1, BRCA1, UBC and TP53 removed and with HDAC2 Highlighted

Betweenness can be seen in these networks by how much more readily genes interact with UBC in the network. It is pretty clear that UBC facilitates many interactions between many other genes which can be visualized by the large number of edges present with it in the network that make it easier for these genes to interact. This is in stark contrast to the network without UBC which has far fewer edges in it which then means that information is not able to be transferred as effectively or efficiently as before. All of the genes are still able to interact with each other but without UBC these interactions become harder which in essence is the idea of what it means for a gene to have high betweenness.

3.4 Ranking the Top Genes

We understand from the definition of closeness and betweenness that the two calculations may result in different genes considered most central in analysis of our gene network. We created a list of the top 10 most central genes according to our two indices by taking the average rank of the gene through each of the bootstrap samples which in our case is 100 bootstrap samples.

Top 10 Central Genes		
Rank	According to Closeness	According to Betweenness
1	UBC	UBC
2	HDAC1	TP53
3	TP53	HDAC1
4	BRCA1	BRCA1
5	HSP90AA1	SNCA
6	CREBBP	GRB2
7	GRB2	EGFR
8	MYC	CREBBP
9	CTNNB1	IKBKG
10	JUN	CTNNB1

Table 3.1: Table of top 10 most central genes according to closeness and betweenness where matching colors correspond to the same gene present under both calculations

We can see by looking at the table above how the top 10 genes overlap between both centrality measures. The first aspect that becomes obvious is that the top four genes appear to be the same regardless of the centrality measure used. In the table, we can see that UBC, which is in cyan, ranks first in both closeness and betweenness. The pairing of TP53 and HDAC1 is 2 or 3 in the rankings and BRCA1 is ranked 4 in both of the rankings. There is also more overlap in the rankings with 7 out of 10 genes appearing in both top 10 lists.

3.5 Understanding the Human Genome

We keep talking about genes and how they interact, but let's talk a step back and understand what these genes do. There are around 24,000 genes in the human body so it is very hard to know what each gene's purpose is. The genes we care about are the 'cancer genes,' the ones that are associated with cancer such as BRCA1. We will care about the genes most active in patients with pilocytic astrocytomas in our case. Thus, we look at the top 10 genes according to the closeness index and also the betweenness index. Understanding these genes will not be as hard of a task and could provide valuable information concerning the gene networks in patients with pilocytic astrocytomas.

The top gene (most close and most between!) from our analysis is UBC (Ubiquitin C). The ubiquitin gene is involved in almost every cellular process which can be seen through the latin root *ubiqui* which translate to 'everywhere.' Ubiquitin C is one of the four genes encoding for Ubiquitin in the human genome. UBC is known to be critical in the regulation of various cellular processes. These processes include protein degradation, protein trafficking, cell cycle regulation, and DNA repair. These genes make up to 5% of the total proteins in human cells. There is an increase in concentration of UBC when people deal with different types of cancer. This increased amount of UBC is thought to be due to the fact that UBC is essential for the growth of cancer cells. [4] This makes intuitive sense from our analysis above that UBC is first under both centrality indices because this gene is constant throughout the body and makes up a large portion of the total genes in the human body!

The next most significant gene is HDAC1 (Histone Deacetylase). HDAC1 are key enzymes regulating very important cell processes such as cell-cycle progression and apoptosis which is the programmed cell death of a certain cell. Mutations of the genes that encode HDAC1 can have pretty profound results and have been linked to tumor growth because 'they induce the aberrant transcription of key genes regulating important cellular functions such as cell proliferation, cell-cycle regulation and apoptosis.' [7]

The next gene is TP53 (Tumor Protein P53). TP53 encodes a tumor suppressor protein which helps stop the growth of cellular problems. This encoded gene targets potential

tumors and other threatening genes and induces cell death or cell repair. Mutations in this gene have been linked to cancer. This mutation leads to a predisposition to have certain cancers including brain tumors (e.g. pilocytic astrocytomas). [5]

The last gene of note is BRCA1 (Breast Cancer 1, Early Onset). BRCA1 is another gene that produces tumor suppressor proteins. These proteins help repair damaged cells which means that they are critical in maintaining the stability of genetic materials. Mutations of this gene cause them to not function properly, the repairing of damaged cells is not fixed which can lead to cancer. BRCA1 mutations increase the risk of breast cancers as well as many other cancers. [1]

Looking at these genes who rank 2-4 in centrality rankings for both indices makes us feel confident about our results because all three of these genes have been seen to be prevalent in people with cancer (i.e. cancer genes). Understanding the rest of the top 10 genes and their purpose in the body would be an easy way to expand the research and see if more cancer genes pop up.

Chapter 4

Conclusion

Given a massive gene network with 8150 genes from children with pilocytic astrocytomas, the bootstrapping method provided an effective way to approximate what genes were most central in the network. Bootstrapping was critical when it came to getting the best approximation to the two centrality measures that we used. These centrality indices helped to shed some light into gene interactions and what genes are most central in these children. Through our extensive data collection, we saw many genes that have been associated with cancer rank in the top tier of most central genes in our gene networks (i.e. BRCA1, HDAC1 and UBC). Having these genes show up helps quell questions about the validity of the results we were getting because these genes have been studied for years in connection with cancer. Understanding these genes and the other most central genes is something that must be done. But still these results provide a very interesting look into gene interaction networks.

This data and research can and should be expanded so that more conclusions can be reached and gene interactions especially in people with cancer. It would be advantageous to increase the size of bootstrap networks from 100 to a much larger number so that the true variability can be seen and a network of best fit can be found. Also, the background on genes and their purposes could be expanded with additional understanding of the human genome. Another logical way to expand the research would be to look the normal group and compare the top genes when conducting that data to see if fewer so called cancer genes show up. All of this would help the data become more powerful and also more informative so that more knowledge can be gleaned about understanding gene interactions in patients with cancer.

Bibliography

[1] Brca1 and brca2: Cancer risk and genetic testing. http://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet#q1, April 2015.

[2] Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, January 1979.

[3] Ernesto Estrada. *The Structure of Complex Networks: Theory and Applications*. Oxford University Press, 2011.

[4] Lucia Radici et al. Ubiquitin c gene: Structure, function, and transcriptional regulation. *Advances in Bioscience and Biotechnology*, pages 1057–1062, December 2013.

[5] Magali Olivier et al. Tp53 mutations in human cancers: Origins, consequences, and clinical use. *Cold Spring Harbor Perspectives in Biology*, January 2010.

[6] Thomas W. Valente et al. How correlated are network centrality measures? http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2875682/, January 2008.

[7] Santiago Ropero and Manel Esteller. The role of histone deacetylases (hdacs) in human cancer. *Molecular Oncology*, 1(1):19–25, June 2007.

[8] American Cancer Society. Cancer facts & figures 2016. http://www.cancer.org/research/cancerfactsstatistics/cancerfactsfigures2016/index, 2016.

[9] H Zhao, W Cai, S Su, D Zhi, J Lu, and S Liu. Screening genes crucial for pediatric astrocytoma using weighted gene expression network analysis combined with methylation data analysis. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE44971, October 2014.