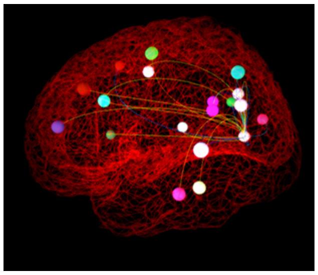
# Sex differences in the Human Connectome

Vivek Kulkarni<br>Jagat Sastry



November 28, 2012

"As long as our brain is a mystery, the universe, the reflection of the structure of the brain will also be a mystery." — Santiago Ramón y Cajal

# 1 Abstract

The human brain has long been an object of great scientific interest. We revel at the immense capabilities that our highly evolved brains possess and wonder at how the brain functions, how vision is interpreted, how consciousness arises etc, all of which neuroscience deals with to a great extent. Recent advances in neuro science and computer science have brought to the fore-front an exciting research area of Brain Networks. The fundamental idea giving rise to this area being that the brain can be thought of as composed of several simple elements that give rise to complex patterns like consciousness[6]. Thus the brain can be modeled as a network which admits the brain to network analysis. Over the years, network science has evolved to a great extent and is now in a position to analyze real world networks. Emergence of massive data, faster algorithms and the ubiquity of networks have contributed to this. We investigate sex differences in brain networks across males and females in this project. We outline some of the differences in brain networks across sexes. We also use some of these discriminative features for the related classification problem "*Can we classify a human connectome (brain network) to belong to one of the sexes ?*" and using a simple *decision tree* as well as a *support vector machine* for the above classification tasks. The rest of the document motivates the problem, describes our research methods and experiments and then presents our findings. We then conclude by interpreting our results and discussing future work.

# 2 Motivation

One of the over arching idea currently in brain research is the idea that it is crucial to study the connections in the brain to gain deeper insight into the functioning of the brain. This is an exciting research area resulting from the confluence of neuroscience and network science whch promises us great insight into the workings of the brain. Perhaps one of the most important projects , analogous to the Human Genome Project in 2005 is the *Human Connectome Project* that was kicked off in 2009. The human connectome project which aims to map the brain's connectivity across regions can help understand diseases like schizphrenia, Alzheimer's disease. It is to be noted that analyzing the human connectome is far more challenging in terms of scale ( it has more than a billion more connections than the letters in a genome)[7]. While the human connectome project is still an ongoing project , exciting initial results have been obtained by analyzing connectomes. A visualization Some important results include the small world property of brain networks, the presence of a rich club of hubs. Noting the larger goal outlined above, one of the research problems that seeks investigation is that of sex differences in brain networks and what they imply in a biological setting. We investigate this problem in our project. We define the problem precisely as follows:

1. What differences in brain network (connectome structure) do both the sexes exhibit ?

2. Would these discriminative features admit to classification of connectomes based on sexes ?

We now discuss related work related to sex differences in the next section.

## 3 Related Work

There are two main approaches to the above problem of identifying discriminative features of the connectome. The first approach would be to look for a subgraph structures (also called a signal sub-graph) which are discriminative and build a classifier based on this. This approach has been described in detail by Joshua Vogelstein et.al[1]. This model has been shown to perform better than other standard graph classification techniques like graph k-NN[1]. The second approach is to identify discriminative network measures (either global or local) and use standard machine learning techniques for classification. Julio.N.Duarte etal[2] analyzed connectome structure to help identify sex and kinship differences. They confirmed the small node nature of brain networks, and also outlined structural differences in brain networks in terms of network measures like communicability , edge betweenness centrality which improved the classification rate to around 93% accuracy(based on sex for their data sets). This was mostly done at a global scale(topological scale) with a data set of 303 individuals.

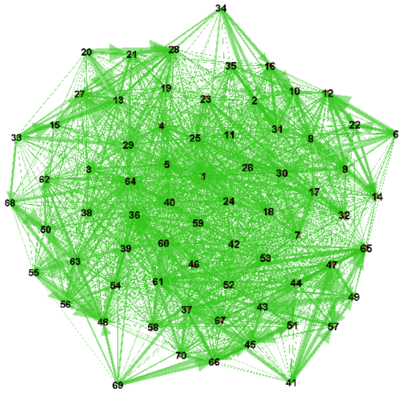
In our project, we are investigating structural differences on a different data set and also look at how both discriminative network measures can be used to classify connectomes according to sex.

## 4 Experiments and Data Set

The data set consists of connectome data for 114 individuals( 50 of them being of 1 sex and 64 of the other, Mean age: 21 years). Each sample is a brain network on 70 nodes(where each node represents a particular brain region , and each weighted edge represents density of fibers between those regions(strength of their connections).Each sample is represented as a weighted undirected graph and is thus represented as a sparse strictly upper triangular matrix.Each sample or connectome is assigned a label (0 or 1) thus identifying what sex the connectome belongs to. The data set also contains labels for other interesting features like extraversion, neurotic-ism, agree-ability etc. These labels would be interesting for a regression task. For all the experiments, we denote the sex with label 0 as *Sex 0* and the sex with label 1 as *Sex 1* respectively. A sample network is shown below:

We now seek to identify discriminative network measures for sex classification. We now briefly outline our experimental method and research technique below

Figure 1: Sample Brain Network Visualized



## 4.1 Preprocessing the data set

Note that each connectome is represented as a weighted undirected graph(that is symmetric and hence strictly upper triangular). Since each weight of the edge represents the density of fibres between those regions, it is crucial that such connectome data be normalized between the range [0,1]. This is because the number of fibres detected by scanner varies from individual to individual. The authors in [2] also point out that there exists an inherent bias in tractography for a given cortical region that depends on the volume of the region, number of fibre crossings etc. However they also point out that there is no unique way of normalizing this data. They do however outline different normalization schemes(based purely on topological measures) which we briefly high-light below as it is crucial to understand the interpretation of a normalization scheme:

- *Global Normalization*: This essentially divides each edge weight by the total weight of all the edges in the connectome, effectively normalizing fibre count between each pair of regions by the total number of fibres.

$$w_{ij} = \frac{a_{ij}}{\sum_{ij} a_{ij}} \tag{1}$$

where $w_{ij}$ represents the normalized fibre count(edge weight) and $a_{ij}$ represents the raw edge weight between region $i$ and region $j$.However this scheme leads to biased weights because it does not account for the fact that some regions are expected to have higher number of fibres and also that if the area of the cortical region is larger , more fibres would be counted.

- *Geometric Mean based normalization*:This scheme divides the fibre count between each pair of edge by the geometric mean of the number of fibres leaving region $i$ or region $j$. This normalization is based on the assumption that it assumes that each pair of brain regions has the same total number of fibres. This normalization is claimed to work correctly globally and on a large sample sizes.

$$w_{ij} = \frac{a_{ij}}{\sqrt{\sum_i a_{ij} \sum_j a_{ij}}} \tag{2}$$

- *Row Mean Based Normalization:* This scheme just divides each edge's fibre count by the total number of fibres incident on that node. More formally:

  $$w_{ij} = \frac{a_{ij}}{\sum_j a_{ij}} \tag{3}$$

  The above normalization scheme has a very interesting interpretation: It can be viewed as the probablity of a connection between regions $i$ and regions $j$ given that there are $\sum_j a_{ij}$ fibres emanating from region $i$. It is worthwhile to note that this indeed provides us valuable information regarding the differences in connectivity between cortical regions. Even though a set of fibres leave a particular region $i$, only a subset of them are used for the connection to region $j$. This model also implies that $w_{ij} \neq w_{ji}$ thus making the resulting graph a weighted directed graph.

We decided to use the *Row Mean Based Normalization* as it captures valuable information about the connectivity differences and believe it models the brain more accurately.

Secondly, in order to reduce the effect of mean brain size differences between males and females, the authors highlighted that one must normalize the above by the maximum weight so that $max(w_{ij}) = 1$

## 4.2 Network Measures

We then calculated the following network measures(We used the Brain Connectivity Toolbox [5]to calculate the measures below):

1. Local Weighted Clustering Coefficient
2. Local Efficiency
3. Degree distribution
4. Edge Between-ness centrality
5. Participation Coefficient of each node

For all the node-based measures, we calculate the mean measure across all subjects of the class.

We then analyzed the data for differences in the mean measures across classes(sexes). To establish statistical significance of a difference, we decided to use a boot strapping approach. This approach is suited very well for our project as we have a small sample size and boot strapping allows us to test out

**Algorithm 1** Boot-strapping algorithm to establish statistical significance

Assume we have 2 independent sample sets (these correspond to samples of both the sexes)
Observed Sample Set 1 is of size $n$ :$\{x_{obs1}, x_{obs2}, x_{obs3} \dots x_{obsn}\}$ and has mean $\mu_{xobs}$
Observed Sample Set 2 of size $m$ : $\{y_{obs1}, y_{obs2}, y_{obs3} \dots y_{obsm}\}$ and has mean $\mu_{yobs}$
Observed Difference in the sample mean is $t^*_{obs} = \mu_{xobs} - \mu_{yobs}$
We need to see if the above difference is statistically significant at a pre-determined level of significance $\alpha$
*Hypothesis:*

- *Null Hypothesis($H_0$):* Both samples are from the same population
- Alternative hypothesis($H_1$): Both samples are from different population and $\mu_x > \mu_y$

1. Merge the 2 sample sets into 1 sample set of size $(m + n)$
2. Draw a boot strap sample , with replacement of size $(m + n)$ from the above merged set
3. Calculate the mean of the first $n$ observations and set it to $\mu_{x*}$
4. Calculate the mean of the remaining $m$ observations and set it to $\mu_{y*}$
5. Calculate the test statistic $t^* = \mu_{x*} - \mu_{y*}$
6. Repeat steps 2,3,4,5 $B$ times and obtain $B$ values of the test statistic.
7. The p-value is then given by:

$$p - value = \frac{NumberOfTimes(t* > t_{obs}*)}{B} \tag{4}$$

8. Reject the null hypothesis if $p - value < \alpha$

hypothesis by creating a large enough sample through repeated sampling. Secondly it has the added advantage that no assumption on the sample distribution is made.

We outline the boot strapping algorithm (in Algorithm 1[8]) :

# 5 Results

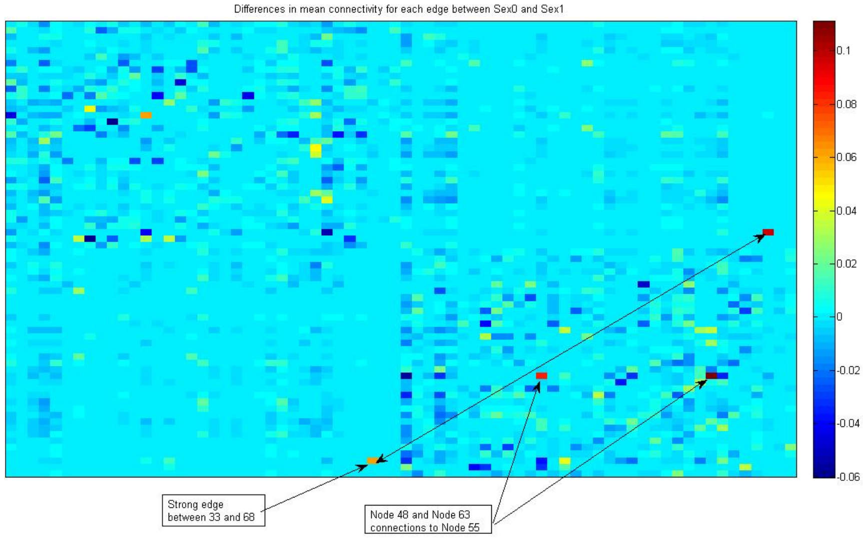
## 5.1 Analysis of the mean edge connectivity

We found the average (mean) weight of each edge for each class (by averaging over all subjects belonging to a class) to identify any edge weights differences among sexes. The heat map (Figure 1) shows the differences in the mean edge weights for each edge between Sex 0 and Sex 1.

We note the following differences:

1. We find strong connections from Node 48 and Node 63 to Node 55 ,in

Figure 2: Mean Edge Connectivity Differences between Sex 0 and Sex 1



one of the sexes.

2. We also note a particularly dominant edge between Node 33 and Node 68 in one of the sexes.

While we do not have labelings for the brain regions represented by the nodes, we speculate that the particularly dominant edge is between 2 hemispheres. This is based on the observation that the labelings given to connectome nodes (based on tractography)tend to be divided into 2 classes(based on hemisphere) perhaps Nodes 1-35 belong to the left hemisphere and Nodes 36 to Node 70 belong to the right hemisphere. Thus we hypothesize that one of the sexes has a particularly dominant edge across hemispheres which could be discriminative.

We will elaborate more on Node 55's role as we present other measures as well.

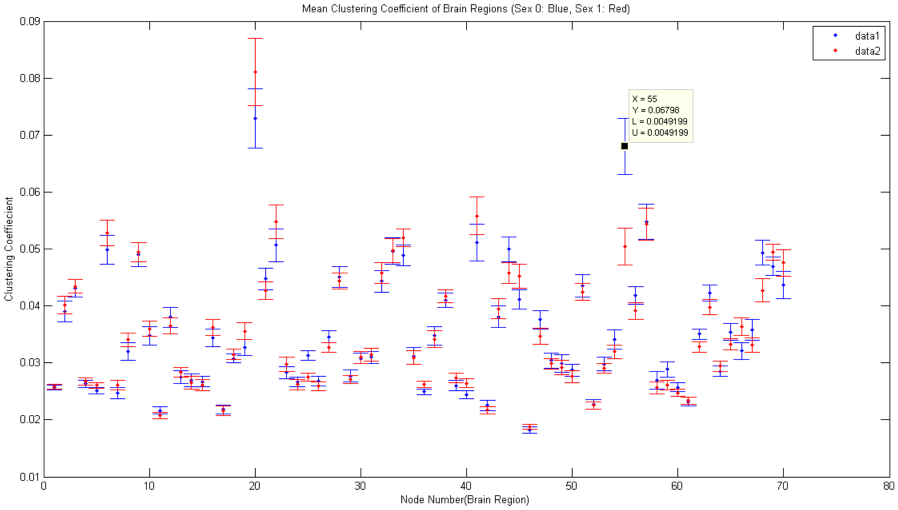
## 5.2 Analysis of the Mean Clustering Coefficient and Local Efficiency

We analyzed the mean clustering coefficient of each node and present our findings (across sexes) below(Figure 2).

We note that mean clustering coefficient of Node 55 in Sex 0 is higher than that in Sex 1. We hypthesize that this difference is statistically siginificant.

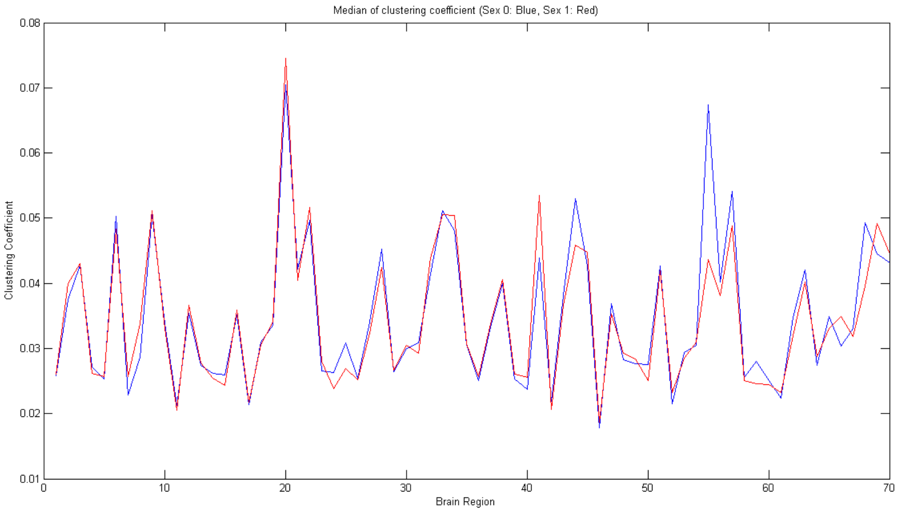
In order to rule out the effects of outliers (as the mean is influenced by outliers) we also looked the median(instead of the mean).We again note that indeed

Figure 3: Mean local clustering coefficient



Node 55's clustering coefficient is higher in Sex 0 than in Sex 1 bolstering our hypothesis.

Figure 4: Median of Clustering Coefficient



The observed sample difference is *0.0175*

To establish statistical significance of this difference, we used the boot-strapping procedure with a significance level of $\alpha = 0.05$ and $B = 3000$. We show a histogram of the test-value's obtained on one run of boot-strapping algorithm. We note that the number of times the test statistic was greater than the observed sample mean is very small. The p-values we obtained are provided as well. We also note that the distribution looks almost normal which is inline with our intuition as we expect that the distribution of the differences in sample means to be normally distributed.

Figure 5: Histogram of test statistic on Boot strapping

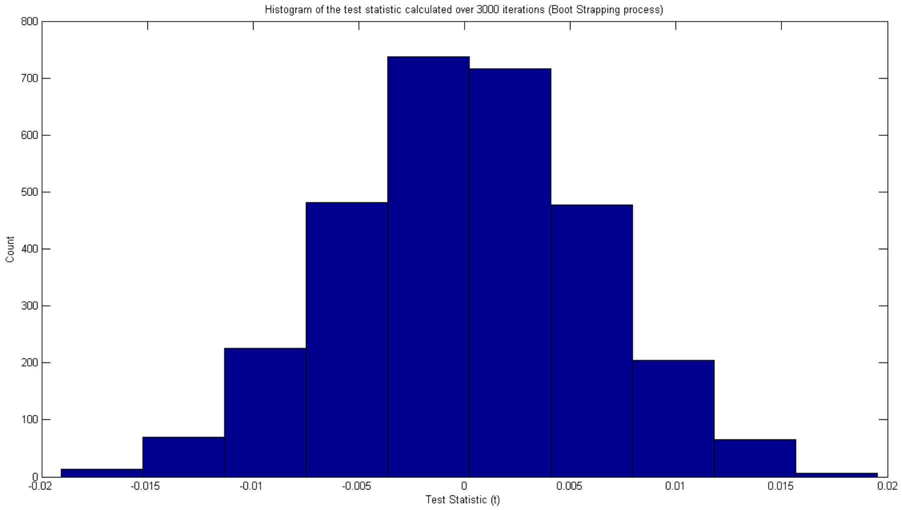


*Table 1: The p-values obtained by the boot-strapping procedure*

| *Run* | *p-value* |
|---|---|
| 1 | 1.0000e-03 |
| 2 | 0.0013 |
| 3 | 6.6667e-04 |
| 4 | 0.0013 |
| 5 | 0.0020 |

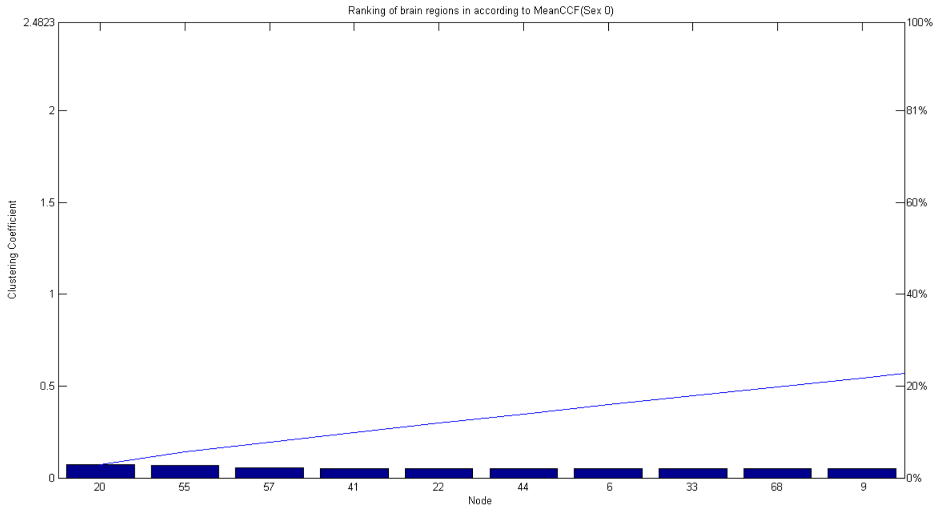
Thus we establish statistical significance in the difference between mean clustering coefficient of Node 55 between the sexes at a significance level of 0.05.

To gain more insight, we ranked the brain regions according to their mean clustering coefficients for both males and females, in the form of a Pareto Chart.

A Pareto Chart represents the ranking of nodes according to the clustering coefficient in decreasing order. The straight line depicts the cumulative total of the values.

We clearly observe the difference in the rankings of the nodes across the sexes(especially that of Node 55). We believe it would be interesting to under-

Figure 6: Pareto Chart for the Clustering Coefficient of Node 55 (Sex 0)



stand if these differences in rankings can be explained biologically and if they have any biological significance.

It would also be useful to look at the Cumulative Distribution Function of Node 55's clustering coefficient across sexes.The individual histograms (which are also represented by the CDF) are shown as well.

The CDF shows that the clustering coefficients are in general lower in Sex 1 than in Sex 0. About 40% of the subjects in Sex 0 have a clustering coefficient less than 0.06 while about 70% of the subjects in Sex 1 have a clustering coefficient less than 0.06

We also note that Node 55 also shows higher efficiency in Sex 0 than in Sex 1. It is also to be noted that there are other regions that also manifest differences although we highighted only the largest ones. The efficiency is a measure of network integration. A high efficiency indicates that pair's of nodes on average have short communication distances and can be reached in a few steps. The local efficiency is just the efficiency calculated over the local neighborhood of that node.

## 5.3 Analysis of the Edge Between-ness centrality

In the brain network of 70 nodes, we represent all the edges by a number obtained its position in the column major order of edges. Thus there are 4900 edges. Our analysis of edge-betweenness centrality across different sexes indicate that there is one edge (namely edge no: 841) which is discriminative across sexes. The figure below shows the mean edge between-ness centrality of each edge (for Sex 0 and Sex 1) (we show only a small range instead of all 4900)

Figure 7: Pareto Chart for the Clustering Coefficient of Node 55 (Sex 1)

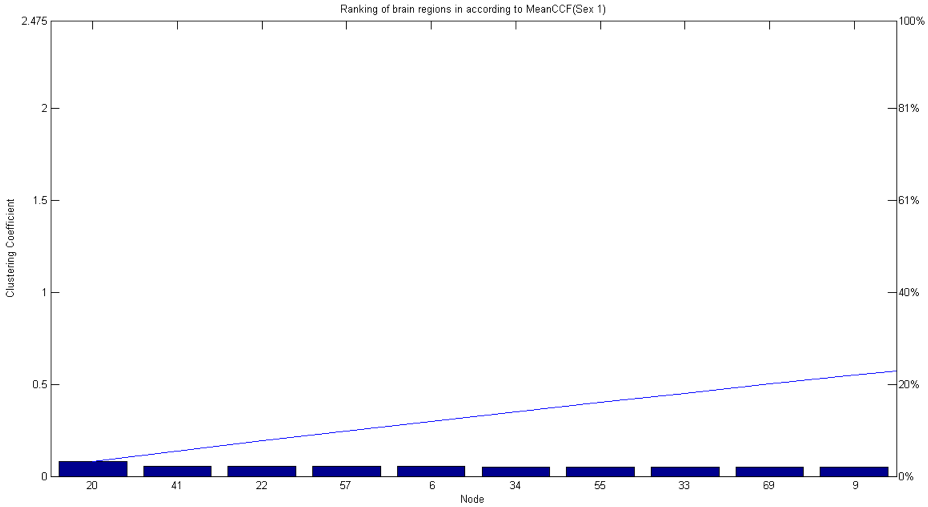


Figure 8: CDF of clustering coefficient of Node 55(Blue: Sex 0 , Red: Sex 1)

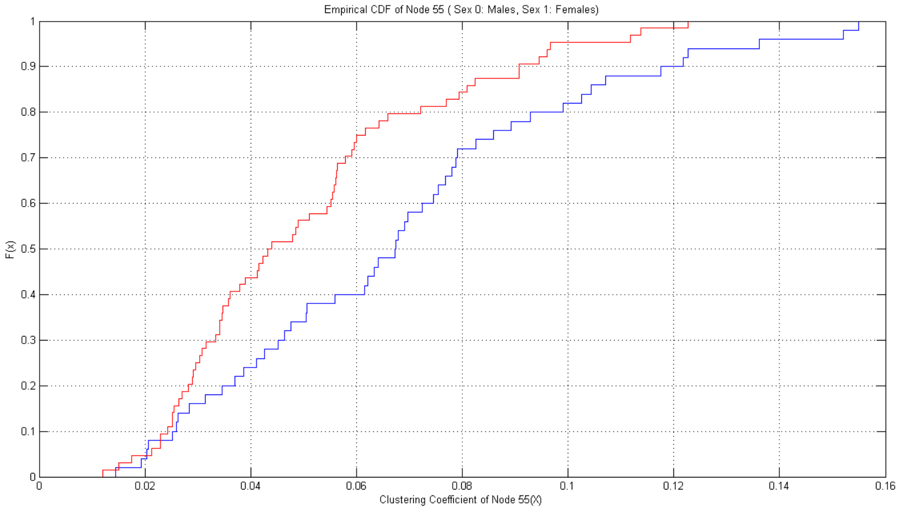


Figure 9: Histogram of Clustering Coefficient of Node 55 in Sex 0

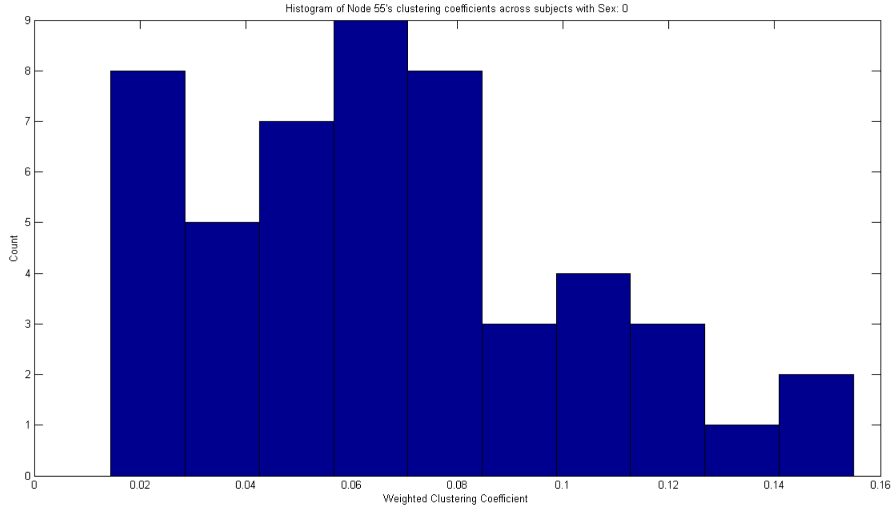


Figure 10: Histogram of the clustering coefficient for Node 55 for subjects of Sex:1

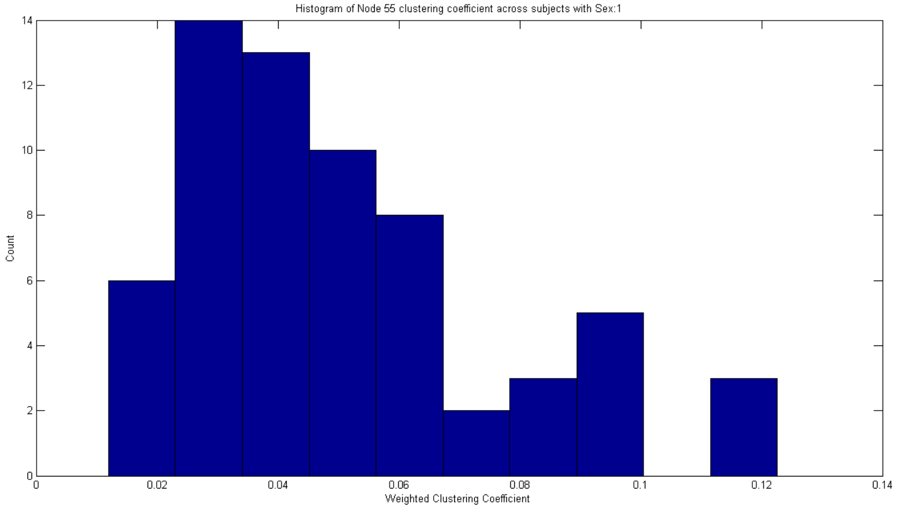
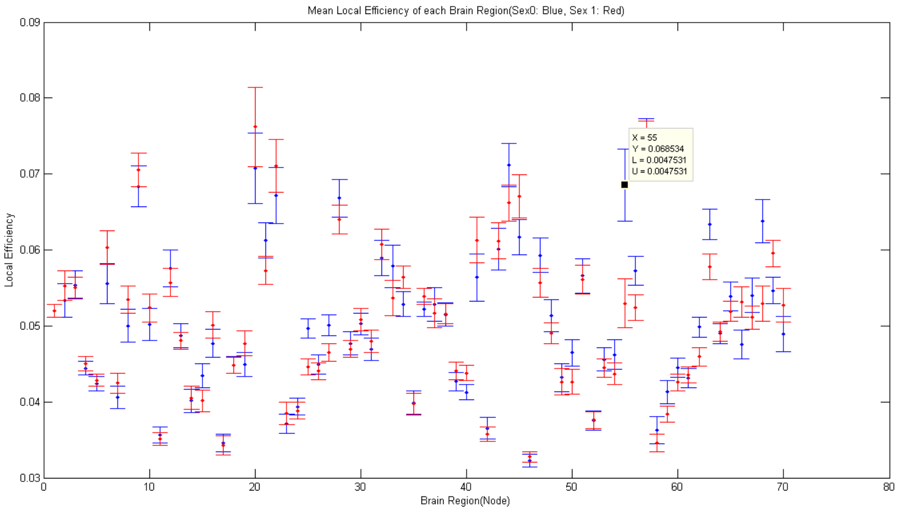


Figure 11: Mean local efficiency of each node



whil

Figure 12: Visualization of ranking of nodes according to Clustering Coefficient : Sex 0

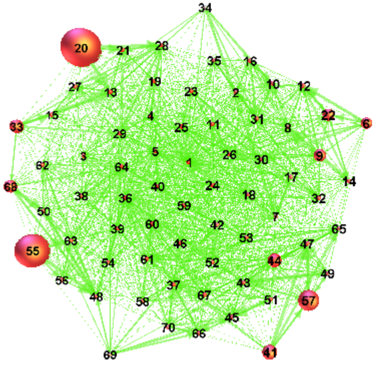


Figure 13: Visualization of ranking of nodes according to Clustering Coefficient: Sex 1

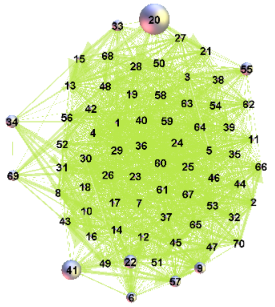
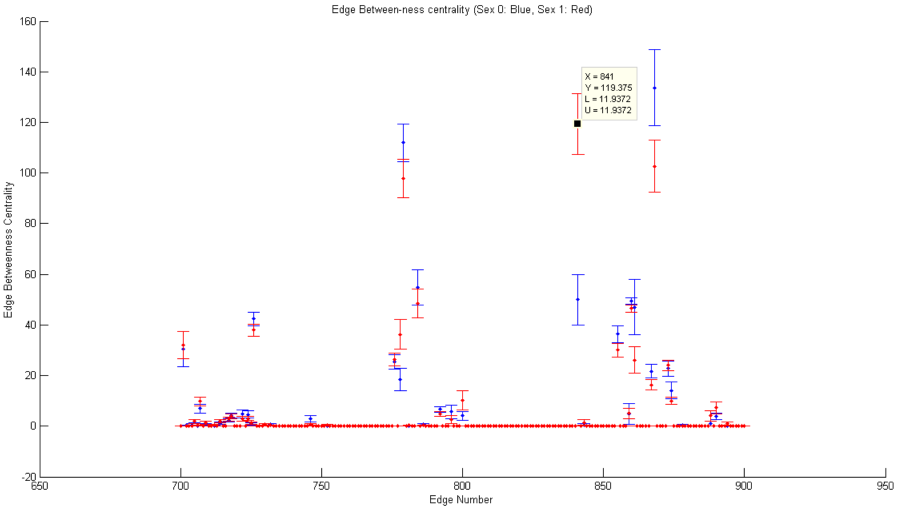


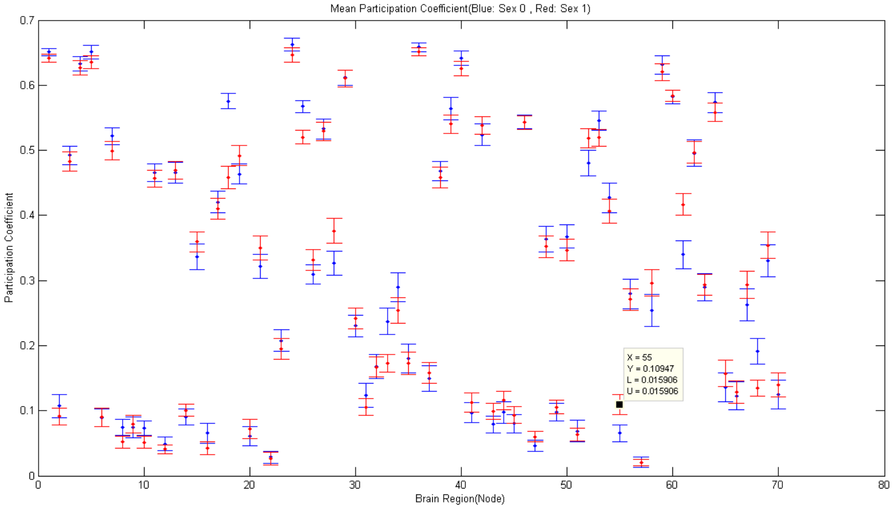
Figure 14: Mean Edge Between-ness centrality for each edge



### 5.4 Analysis of Participation Coefficient:

The participation coefficient is a measure based on modularity. It represents the diversity of inter-modular connections of a given node. Intuitively the participation coefficient of a node is close to one if it's links are uniformly distributed across all modules and 0 if all its links are within its own module. A node with a high participation coefficient thus represents a connector hub in the brain. We then decided to investigate whether Node 55 which was most discriminative was a hub. We note that although there is a difference in the participation coefficient in Node 55 among the sexes, we see that there are other nodes having higher participation coefficients thus indicating that Node 55 is unlikely to be a connector hub. We however note that although the clustering coefficient of Node 55 is higher in Sex 0 than in Sex 1, the participation coefficient is lower in Sex 0 than in Sex 1. This seems to indicate that the brain region corresponding to Node 55 connects closely with its neighbors (is densely clustered) with its own module in one of the sexes(namely Sex 0).

Figure 15: Participation Coefficient of each region

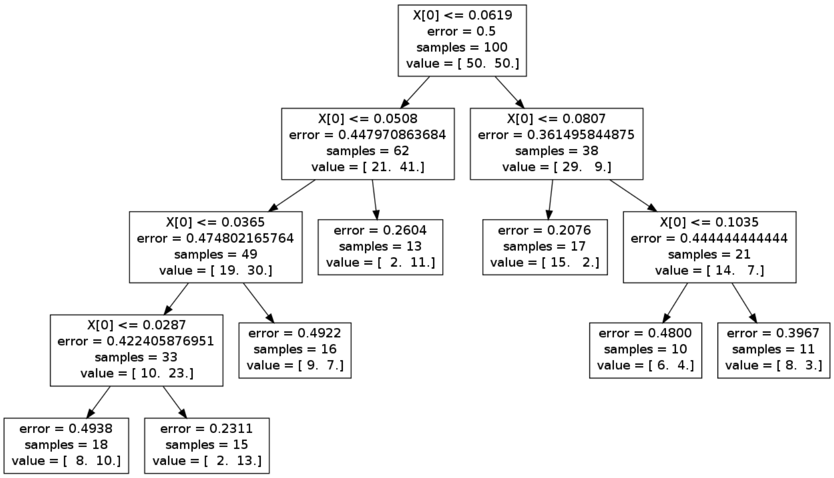


Having established differences in network measures , across sexes, we now cherry pick a set of features and use these features to train and evaluate a classifier which we will discuss in the next section.

### 5.5 Classification

We used the features obtained from previous analysis and trained 2 simple classifiers on them. We decided to use Decision Trees to get an understanding of the best split and is very simple to use. We also trained a support vector

Figure 16: Decision Tree Best Split on Node 55's clustering coefficient alone



machine classifier with a linear kernel on this classification task. We evaluated our classifier using a 10-fold cross-validation. We describe both of these below separately:

### 5.5.1 Decision Trees

We decided to investigate the best split obtained when using clustering coefficient of Node 55 alone as a feature. Based on the analysis if clustering coefficient, we would expect the decision tree to use a point at around 0.06. To avoid over fitting, we ensured the decision tree will always have a minimum of 10 elements in the leaf nodes. The decision tree indeed reported the best split at around that value. The best split was found to be around 0.0619. The 'error' field represents the Gini Coefficient[4] before the split. This confirms our intuition that the best split would be around 0.06.

We tried various combinations of features to assess the best set of features which results in optimal accuracy. We outline them below:

| Network Measure | Best Feature Set | Accuracy |
|---|---|---|
| Clustering Coefficient | Nodes 25, 55, 68 | 0.68(±0.04) |
| Edge Between-ness Centrality | Edge: 841 | 0.68(±0.06) |
| Participation Coefficient | Nodes 18,61 | 0.68(±0.1) |

### 5.5.2 Support Vector Machines

We evaluated how a simple support vector machine with a linear kernel performs on this classification based on cherry-picked features. We outline our results below(the best feature set):

| Network Measure | Best Feature Set | Accuracy |
|---|---|---|
| Clustering Coefficient | Nodes 20, 55 | 0.68(±0.05) |
| Edge Between-ness Centrality | Edge: 841 | 0.58(±0.05) |
| Participation Coefficient | Nodes 18,61, 68 | 0.73(±0.05) |

We note that both classifiers report comparable accuracies with our cherry-picked features. It would be useful to evaluate the classifier on a larger data set to get better accuracy bounds.

## 6 Future Work

It would be useful to understand the biological significance of these differences (which we have not investigated in this project). It would also be important to test on larger data sets as the data set we used was small and we may not have enough power of the test to identify discriminative features which require a higher power of resolution. It would also be useful to establish statistical significance of the features we identified as discriminative(perhaps by a boot strapping procedure).

It would also be useful to investigate the use of motif's and discriminative sub graphs(sub graphs which occur frequently and are discriminative) for classification. One could also extend the above analysis to explore differences in math talent, agree-ability , neurotic-ism etc.

## 7 Conclusion

We just summarize our results succinctly here :

- We have shown that there exists sex differences in network measures like clustering coefficients, edge between-ness centralities and participation coefficient.
- We have also shown these differences perform moderately well on classification and these few set of features boost accuracy.

## 8 Acknowledgements

We are extremely grateful to Prof.Leman Akoglu for providing us the opportunity to work on this exciting research project and guiding us all along. We thoroughly enjoyed the insightful discussions we had regarding this project.

# References

[1] Vogelstein, J. T., Gray, W. R., Vogelstein, R. J., & Priebe, C. E. (2011). Graph Classication using Signal-Subgraphs : Applications in Statistical Connectomics. Applied Physics, 1-15.

[2] Hierarchical topological network analysis of anatomical human brain connectivity and differences related to sex and kinship:Julio M. Duarte-Carvajalinoa, et al.

[3] Network analysis of intrinsic functional brain connectivity in Alzheimer's disease: Supekar K, Menon V, Rubin D, Musen M, Greicius MD.

[4] http://en.wikipedia.org/wiki/Gini_coefficient

[5] Brain Connectivity Toolbox: https://sites.google.com/site/bctnet/

[6] Networks of the Brain: Olaf Sporns

[7] http://www.humanconnectomeproject.org/2012/03/mapping-out-a-new-era-in-brain-research-cnn-labs/

[8] http://faculty.psy.ohio-state.edu/myung/personal/course/826/bootstrap_hypo.pdf