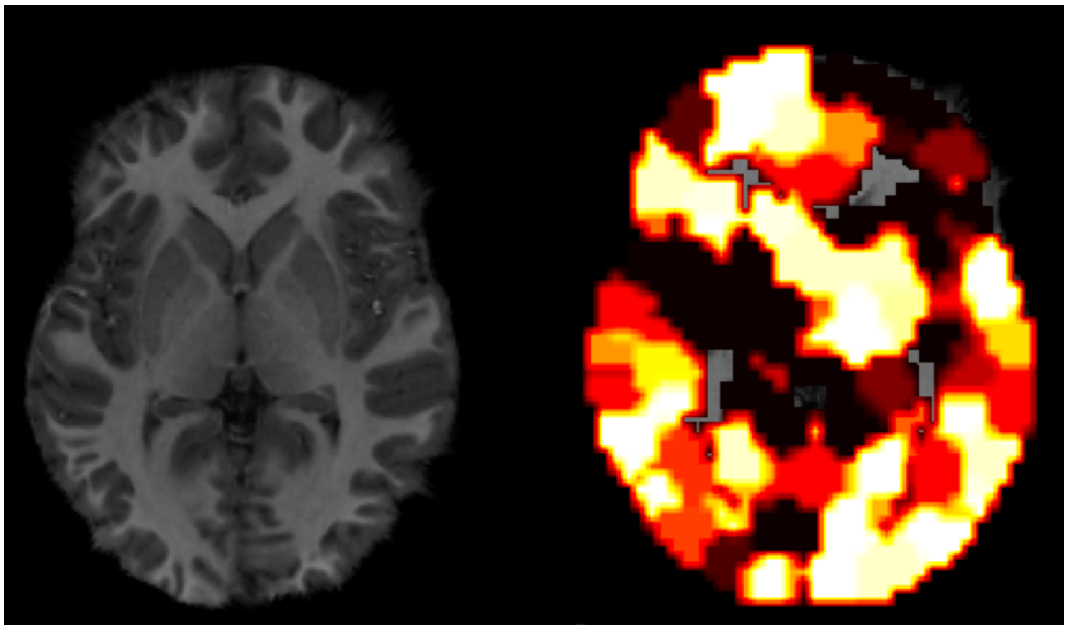
Cortical Parcellation and Classification using PageRank Clustering and the Small-Worldness of ADHD

Master's Thesis



Niels Arjan van Galen Last

Cortical Parcellation and Classification using PageRank Clustering and the Small-Worldness of ADHD

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE

by

Niels Arjan van Galen Last
born in Naarden, the Netherlands

Man-Machine Interaction Group
Department of Mediamatics
Faculty EEMCS, Delft University of Technology
Delft, the Netherlands
http://www.mmi.tudelft.nl

Stanford University
Institute for Computational and
Mathematical Engineering
Stanford, CA, USA
http://icme.stanford.edu

 Cover picture: Axial slice at 0mm of a subject diagnosed with ADHD, left: anatomical image, right: result of functional clustering in 71 partitions with k-prmeans algorithm.

Cortical Parcellation and Classification using PageRank Clustering and the Small-Worldness of ADHD

Author: Niels Arjan van Galen Last
Student id: 1533096
Email: n.a.vangalenlast@student.tudelft.nl

Abstract

Networks and graphs are all around us and can represent a variety of models: from airline connections, the World Wide Web to the way people are connected on Facebook. There are are a few classes of networks with interesting properties such as structured and random networks. One particular interesting class is the small-world network, with a short path length between nodes and a high degree of clustering. This network emerges in many (real-world) phenomena, like the models mentioned above, as well as biological models such as protein-protein interaction and neuronal connections in the human brain. We discuss the differences in small-world properties in the anatomical brain for typically developing children and children diagnosed with Attention-Deficit/Hyperactivity Disorder. Using the differences we investigate the possibilities for classification between the populations. Secondly, we propose a clustering algorithm that is based on spectral clustering and uses the importance of nodes in a network to find clusters. This algorithm is similarly applied to fMRI data and used to reduce the dimensionality before classification.

Thesis Committee:

Chair:	Prof. Drs. Dr. L.J.M. Rothkrantz
University Supervisor:	Prof. Dr. Ir. M. Gerritsen
Committee Member:	Dr. ir. P. Wiggers
Committee Member:	Ir. H.J.A.M Geers

PREFACE

The human brain. Understanding the human brain, from a computational point of view, has been one of my long term fascinations.

This passion can be traced back to a period in time, about a decade ago, shortly after I started programming. It did not take long before I got bored with the applications I wrote and wanted to develop something that exhibited some form of intelligence. This led me to the exciting field of Artificial Intelligence, and turned my passion into my goal of understanding and studying all aspects of it, thoroughly.

The first three weeks of my Masters at the Delft University of Technology I participated in a study visit to the USA. This experience turned out to be life changing. We visited many inspiring companies and universities and upon coming back to the Netherlands, I was very motivated to do my thesis project at a university in the United States. In November of 2009 I explained my ambitions to Professor Rothkrantz who put me into contact with Professor Gerritsen of Stanford University. A year later, in November 2010, I was at Schiphol Airport with two suitcases, the Fulbright Scholarship, Huygens Talent Scholarship and the EEMCS scholarship and departed for my research project to the USA. Nine months later I am proudly writing the preface to my thesis, looking back at an incredible time I could not have done without the help of many people around me.

First and foremost I would like to express my gratitude and appreciation to my supervisor, Professor Margot Gerritsen of Stanford University for taking me on as a student. She provided me the freedom to find my thesis topic, helped me to stay on track and taught me to think critically about the *how* and *why* questions of research. I could not have wished for a warmer and more pleasant welcome at Stanford University.

Furthermore, I would like to thank Léon Rothkrantz for his guidance during my Master's. I would not have come this far if it wasn't for his help in realizing my dreams. His suggestions and feedback in the final stages of writing my thesis helped me improve it significantly.

Additionally, I thank dr. ir. Pascal Wiggers and Ir. H.J.A.M. Geers for being part of my thesis committee.

On a personal note I want to express my gratitude to my parents (Pim and Annemieke van Galen Last) and sister (Jolien) for their everlasting, unconditional support and love, no matter which path I take to pursue my dreams and passions. Finally, I want to show my appreciation to my girlfriend, Sabine, for proofreading and editing my thesis. Her endless patience during the days and nights I am working, or when I am simply not physically with her, is truly admirable.

Thank you.

Niels Arjan van Galen Last
Delft, the Netherlands
August 23, 2011

Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

This chapter provides an informal introduction to this thesis. In the following section we describe the underlying motivation for performing this research and which problems we are targeting. In Section 1.2 we formulate the research questions that drive this thesis. Finally, Section 1.3 lays the outline for the sequence of chapters.

1.1 Motivation / Problem Description

Recent studies show that attention-deficit/hyperactivity disorder (ADHD) is now the most frequently diagnosed neurological disorder for children in the USA [2000]. The disorder shows the first symptoms in childhood and many children, who are diagnosed with ADHD, continue to experience the effects of the disorder throughout the rest of their life. ADHD is mostly diagnosed in children and young adolescents. Sometimes adults are diagnosed with ADHD but they oftentimes show slightly different symptoms, such as restlessness instead of hyperactivity.

The symptoms of ADHD, as the name implies, include inattention and/or hyperactivity and impulsivity. Most of the children who are diagnosed with ADHD display one or more of these traits. There are three different types of ADHD:

***combined* ADHD**, in which all symptoms (inattention, hyperactivity and impulsivity) are present

***inattentive* ADHD**, which requires the attention and concentration issues.

***hyperactive-impulsive* ADHD**, marked by hyperactivity without inattentiveness or concentration problems.

For the diagnosis of ADHD, the symptoms need to be present before the age of 7 and in different settings. For example, children have to be inattentive both at home and at

school, and it seriously has to affect and influence the ability to function or the performance. The diagnosis must be done at a young age because the disease can seriously affect the academic performance, influence the social relationships or may lead to other psychiatric disorders when untreated [2000].

The diagnosis of attention-deficit/hyperactivity disorder in children has been monotonically increasing over the past 10 years [Davis and Williams, 2011]. More boys are diagnosed with the disease than girls of the same age. As a matter of fact, boys are about 3 to 4 times more often found positive for ADHD than girls. Recent research however, shows that this ratio has decreased during the past few years [2010]. This report describes an increase from 7.8% to 9.5% of children (age 4 to 17) diagnosed with ADHD between 2003 and 2007, which is a substantial 21.8% increase in just 4 years. Research by Mandell et al. [2005] showed an increase of 39% between 1989 and 2000 in diagnosis of several psychiatric disorders. They found an increase of 381% of ADHD diagnosis during this timespan.

Although some might think the increase is solely due to overzealous doctors who make the diagnosis and parents who have their children labeled with ADHD because they are restless, it is almost certain that ADHD is indeed over-diagnosed. The rapid increase in ADHD diagnosis does therefore not necessarily reflect a true increase of such a magnitude. The second argument to support the increase in diagnosed ADHD cases is more awareness and an increased understanding of the symptoms. The diagnosis is now set by the physician, based on his interpretation of the information he receives from the parents and the behavior of the child.

Because the diagnosis of attention-deficit/hyperactivity disorder is based on opinions and interpretation we are going to explore a novel approach for the diagnosis of ADHD based on the structure of the a child's brain derived from a functional Magnetic Resonance Image.

With the ever-increasing creation and generation of data and the emergence of massive information networks, many algorithms are often no longer feasible. This is largely due to the prohibitive time and space complexity of the algorithms. For example, computing the shortest path between each set of pages on the internet would take, even when all computers in the world would be used at the same time, too long to be meaningful. Instead of applying algorithms to a large network there is a more recent research area that focusses on *local* algorithms, applied to a subset of the data. Graph partitioning and finding local communities in a network is a well-studied problem that is relevant in many areas. The partition algorithms are typically developed using divide-and-conquer approaches and because exact solutions for graph partitioning is known to be NP-hard, various approximations have been developed. There are many examples of networks that turn out to belong to a special class known as the small-world, one of the most remarkable and interesting that belongs to this class is the neuronal network and the internet. For finding local communities or clusters on the internet, algorithms used to rely on the Euclidean distance between nodes. However, this fails to capture the underlying structure and properties of a small-world network such as the short characteristic path length between nodes and the high cluster coefficient. Page et al. [1998] then introduced the PageRank algorithm, to find order in the web using a form of spectral clustering. In

this thesis we investigate whether it is possible to combine the idea of PageRank with a clustering algorithm. We will determine the performance of this algorithm with graphs that exhibit small-world characteristics similar to the ones found in the brain.

1.2 Research Question

This project is centered around the question:

Are there significant differences in small-world properties in a graph representation of the brain between normal children and children diagnosed with attention-deficit/hyperactivity disorder?

This question is supported by several smaller questions which we will explore in the following chapters of this thesis, such as:

- What is the 'small-world' phenomenon and how does it relate to the brain?
- Can we use a cluster algorithm like k-means to partition a graph irrespective of the structure?
- Is it possible to built a classifier for fMRI data?
- Can we detect differences in fMRI data for children with ADHD?
- Are there differences between properties between normal children and children who are diagnosed with ADHD? If so, what are the differences and are they significant?
- If there are differences, can we utilize them to classify whether somebody has ADHD based solely on these characteristics?

1.3 Outline

This thesis is divided in four parts, hereafter we give a brief overview of the contents of each part.

Theory

In the first part we provide the foundations for the topics discussed in the sequence of parts. We start with an overview of *graph theory* in Chapter 2. In this chapter we discuss a number of metrics to 'measure' different properties of graphs and various topologies such as the small-world network and some derived models. In Chapter 3 we introduce the PageRank, a simple algorithm that drastically changed the way we search and retrieve relevant information on the internet. An introduction to computational neuroscience is given in Chapter 4. The discussion starts with the human brain, using functional magnetic resonance imaging to create an anatomical or functional image of the brain and explore some research that combines graph theory and the brain.

Clustering the brain

The second part of this thesis is about clustering the brain. In Chapter 5 we introduce the k-means algorithm, one of the most, if not the most, used cluster algorithm as well as a recent improvement known as k-means++. Then we explain two methods for approximating a personal PageRank for a vertex in a graph and use this to define a new distance metric: the *pagerank-distance*. We propose a new cluster algorithm, *k-prmeans*, that uses this distance metric and provides good results in small-world graphs where the Euclidean distance fails. Finally we provide an algorithm called *stable-k-prmeans* to find a 'good' partitioned graph using repeated clustering and the conductivity of clusters. Experimental results of this algorithm are shown in Chapter 6, in which we apply this algorithm to the small-world network and several derivations.

The Small-Worldness of ADHD

In the third part we investigate the small-world properties of children with and without attention-deficit/hyperactivity-disorder. Chapter 7 explains the data and how it is preprocessed. Then we explain how we derive a graph from the fMRI data in which the nodes correspond to anatomical regions and the edges to activity between them. We use these graphs to investigate the small-world properties between these groups of children and provide an extensive physiological analysis of our findings.

We use the small-world properties derived from an anatomical template to classify the groups of children in Chapter 8. Different types of dimension reduction are explored and classification results are presented. In Chapter 9 we explore classification using cluster coefficients derived from fMRI data clustered with the stable-k-prmeans algorithm.

Conclusions

In the final part of this thesis we present the conclusions. In Chapter 10 we provide a summary of this thesis, present our findings and give suggestions for future research.

Part I

Theory

CHAPTER 2

GRAPH THEORY

In the year 1735, the eminent mathematician and physicist Leonard Euler (1707-1783) solved the 'Seven Bridges of Königsberg' problem. The old city of Königsberg was set around two islands in the Pregel River. The islands were connected by a total of seven bridges to each other and to the shores of the mainland. The original map of Königsberg with depicted bridges and river is shown in Fig. 2.1. The problem can be formulated as follows: *find a walk through the city, across each of the seven bridges where every bridge is to be crossed once and only once, and a bridge can only be crossed in its entirety.* Euler came up with a negative resolution by representing the problem as an abstract network: a 'graph'. This is now considered as the first proof in the field that is now known as graph theory.

This chapter lays the graph theoretical foundations of this thesis. Section 2.1 describes the representation of graphs and the most important topologies. Then in Section 2.2 several graph metrics are discussed. Lastly, Section 2.1.3 presents several other derivations and characteristics of other network topologies.

2.1 Measuring Graphs

In this section we explore the representation of graphs, the structure of graphs and small-world networks.

2.1.1 The Definition of a Graph

As explained in the introduction, many models that involve connections or relations can be represented in an abstract form, a network, known as a graph. For example, the graph representation of the problem is shown in Fig. 2.2. The green circles can be considered

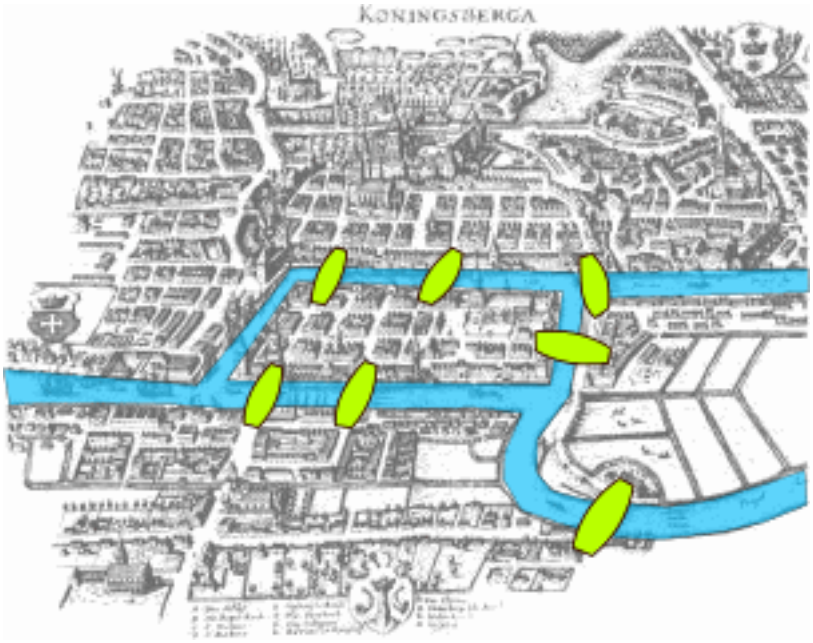


Figure 2.1: Original map of Königsberg with the river and seven bridges

as the mainland, the blue circles represent the islands and the lines between the circles represent the bridges.

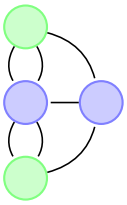


Figure 2.2: Abstract representation of the *Seven Bridges of Königsberg* problem

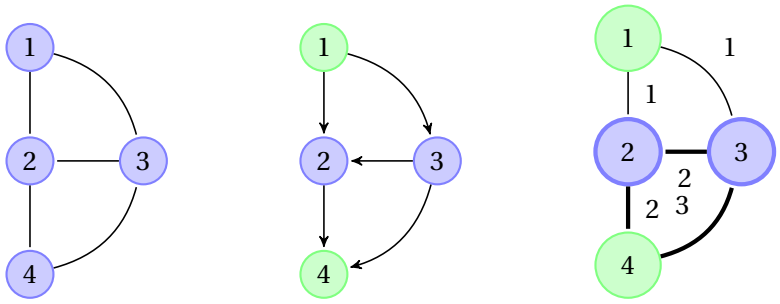
Formally, we will consider a *graph* G, consisting of a *vertex* (node) set $V = \{v_1, \ldots, v_n\}$ and an *edge* (connection) set E, denoted as $G(V, E)$. Depending on the problem, the presence of an edge between two nodes indicates a certain relationship between two vertices, e.g. in Euler's case, an edge represents a bridge. The connections between every node can be modeled as an adjacency matrix A. If and only if there is a connection between two vertices v_i and v_j, the entry in the adjacency matrix is: $A_{i,j} = 1$. The number of edges of a given vertex is called the *degree*, denoted by $d(v_i)$. From the degree of each vertex we can construct a diagonal matrix D, where $D_{i,i} = d(v_i)$. Every vertex has a certain degree and we can model this for the graph by the degree distribution $P(d(v_i))$, this describes the likelihood that a randomly chosen vertex v_i has degree $d(v_i)$.

When it comes to the edges there are more characteristics than the degree of a vertex. A graph can be *undirected*, in this case information can flow from v_i via the edge

to v_j and in the other way around. In a *directed* graph information can only flow in the direction that the vertices are connected. For a directed graph it is possible to differentiate between the *in degree* and *out degree* distribution. This is defined by the number of edges pointing to a vertex and from this vertex, respectively. Furthermore, a graph in which all edges are considered as equally important is called a *unweighted* graph. A graph in which some edges are of higher importance, or indicating a stronger connection between vertices, is called a *weighted* graph.

Basic examples of the different types of graphs are shown in Fig. 2.3. Fig. 2.3a displays an undirected and unweighted graph. In this graph one can traverse from each set of edges in both directions and each edge has the same importance. The edges in Fig. 2.3b are also of equal importance but one can only traverse the edges in the direction the arrow is pointing. That is, one can go from either node two or node three to node four, but once one is in node four it is impossible to traverse to another edge since there are no edges out of the node. Finally, Fig. 2.3c shows a graph in which the nodes can be traversed in both directions but are assigned a 'weight', which gives some numerical information about an edge. For example, the value can express the cost of traversing the edge or the distance between two edges.

As explained before, we can summarize the a graph using an adjacency matrix. The adjacency matrices corresponding to the example graphs of Fig. 2.3 are shown below in equation 2.1. The left matrix is a symmetric matrix for which each edge is represented by a 1. The middle matrix is asymmetric because the edges cannot be traversed in both directions by default. If there is an edge adjoining two edges in the weighted graph, the entry in the matrix on the right shows the associated weight.



(a) Undirected (b) Directed (c) Weighted

Figure 2.3: Different Graph Structures

$$Fig\ 2.3a \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} \quad Fig\ 2.3b \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix} \quad Fig\ 2.3c \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 2 & 2 \\ 1 & 2 & 0 & 3 \\ 0 & 2 & 3 & 0 \end{pmatrix}. \tag{2.1}$$

2.1.2 The Structure of a Graph

Small graphs such as shown in Fig. 2.3 are easily to visualize and interpret. This holds for graphs consisting of tens or hundreds of vertices for which it is easy to draw the graph. Then, with our visual perception ability we can get a good understanding of the structure of the network. However, as the graph becomes larger with thousands up to billions of vertices, it becomes impossible to visualize it. When analyzing a graph of this size, it is convenient to fall back to statistical methods to make useful statements about the graph.

The first, and simplest, type of graph is the *random graph*. It was first discussed by Rapoport [Rapoport, 1957], who especially focussed on the degree distribution, and Erdös and Rényi [1959], who proved many important theorems. Rapoport [Rapoport, 1957] were the first to define a model for large random networks, called the "random net". Nearly a decade later Erdös and Rényi independently published a similar model what they called the "random graph". Their model of the random graph used a Poisson process for the random wiring of the edges. It is utterly simple to construct a model of a random graph. Start by considering n vertices and connect each vertex to every other vertex with probability p and the random graph is complete. Random graphs are extensively studied and a many properties have been proved by Erdös and Rényi [Erdös and Rényi, 1959, 1960, 1961].

When varying the degree distribution of a graph from one to zero, the rewiring probability decreases and the structure goes from random to an ordered, regular lattice, graph. Every vertex in an *ordered graph* is connected to its k nearest neighbors. How the notion of nearest neighbors is defined, depends on the underlying network that is to be modeled by this graph. It does not come as a surprise that there are not many natural phenomena that can be modeled using either a lattice or random graph. When researchers attempted to model real networks they stumbled upon one phenomenon that they could not explain using these networks: the calculated 'distances' in sparsely connected networks turned out to be much smaller than estimated.

2.1.3 Small-World Networks

The implications of the short distances between entities was first recognized in 1929 by the Hungarian writer Frigyes Karinthy in his experiment 'Chains', [Karinthy, 1929]. In this article, he argues that it is unlikely for a graph consisting of all connections between people in the world, the average 'distance' between any two people to be larger than five persons. His estimate turned out to be very accurate and let to the discovery and introduction of a new class of networks. The empirical evidence of this phenomena came from Stanley Milgram, who studied behavior of social networks. In one of his, now most-famous, experiments, he sent out letters to random selected people in the USA in which he asked them to forward the letter to a specific person located in Boston. If they didn't know that person, he asked them to forward the letter to only one person, who are either living closer to Boston or had a large likelihood of sending the letter in the right direction. Many letters reached the designated endpoint and the letters were sent only 5.5 times. This seemed to correspond to Karinthy's expectations that the social connections

in the world are very tightly coupled. These phenomena was coined the 'small-world' phenomenon and is also known as the 'six degrees of separation' [Milgram, 1967; Travers and Milgram, 1969]. Several other experiments confirmed the existence of the 'small-world' phenomena in natural networks, but for a long time nobody could give any valid explanation for this. The neuronal network of the *Caenorhabditis elegans* exhibits small-world properties [Watts and Strogatz, 1998], world-wide airport networks [Guimera and Amaral, 2004], metabolic networks [Jeong et al., 2000; Kaiser and Hilgetag, 2004] and collaborations between actors [Collins and Chow, 1998] are all examples of networks that exhibit these characteristics. These networks also have shown to be more persistent to random 'attacks' and a higher synchronizability.

In fact, why such diverse networks all show these properties remained a mystery for a long time. This all changed when Watts and Strogatz [1998] published an article in Nature, proposing a very simple model to explain this behavior. Their model is based on the phase transitions using a rewiring procedure when interpolating between a regular and random network. They start with an ordered graph shaped in a ring with n vertices and k edges per vertex, shown in Fig. 2.4. For a given probability p, each edge is rewired at random. As p varies from 0, the ring lattice, to 1, a random graph, it allows for probing the structures in the intermediate region $0 < p < 1$. The structural properties of the graphs are then quantified using two metrics known as the *characteristic path length* and the *clustering coefficient*, which will be explained in the sequent section.

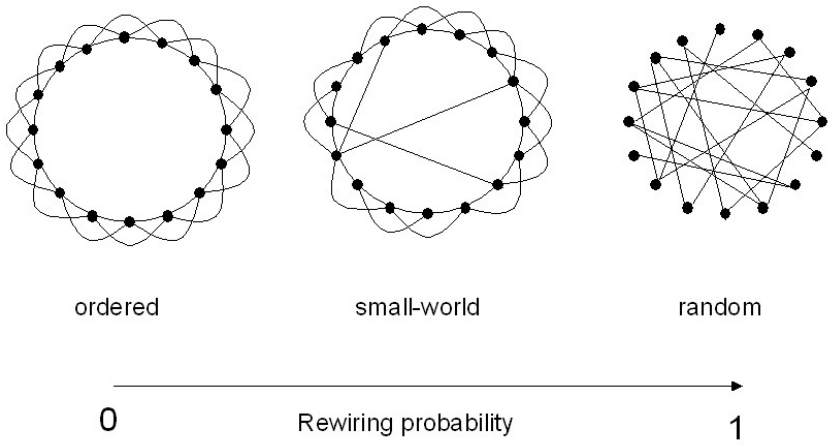


Figure 2.4: Rewiring a network from a lattice to a random network with the intermediate small-world network [Stam and Reijneveld, 2007]

2.2 Graph Metrics

Now that we introduced some different topologies, we can define some metrics for each type. There are two characteristics that are oftentimes used to explain the *local* or *global* structure of the graph [Watts and Strogatz, 1998; Newman, 2003b]. These characteristics are the *clustering coefficient C* and the *characteristic path length L*.

The clustering coefficient is a statistic that gives some information about the local clustering, when it is calculated for a vertex, or about the density of the entire graph. To amount of local clustering for some vertex is based on the number of edges between the direct neighbors of that vertex.

The idea of the clustering coefficient is based on triplets, i.e. connection between three nodes. A triplet is a set of three nodes that are connected by a number of edges ranging from zero to three, and was first introduced by Luce and Perry [1949]. A greater number of triangles, sets of three edges between three neighboring vertices, between the neighbors of a vertex indicate a higher clustering coefficient. Formally, the clustering coefficient C_i of vertex i with degree k_i is defined by the ratio of edges between its neighbors to the possible number of edges between them. The vertices directly adjacent or connected to a vertex v_i are said to be in the *neighborhood* of that vertex, defined by $N_i = \{v_j : e_{i,j} \in E \wedge e_{j,i} \in E\}$. The clustering coefficient is defined slightly different for undirected and directed graphs. In the undirected model, for each edge $e_{i,j}$ there is also an edge $e_{j,i}$, and therefore maximum number of edges is $\frac{k_i(k_i-1)}{2}$.

The clustering coefficient for an undirected graph can thus be defined as [Watts and Strogatz, 1998][1]:

$$C_i = \frac{2\,|\,\{e_j\}\,|}{k_i(k_i-1)} : e_j \in N_i. \tag{2.2}$$

For a directed graph, the maximum number of edges that could exist between directly adjacent to v_i is $k_i(k_i-1)$, and the cluster coefficient is given by:

$$C_i = \frac{|\,\{e_j\}\,|}{k_i(k_i-1)} : e_j \in N_i. \tag{2.3}$$

It is immediately clear that for each vertex the clustering coefficient is in the range between zero and one. This allows us to use the local clustering coefficient to obtain the mean clustering coefficient of the entire graph:

$$C = \langle c \rangle = \frac{1}{N}\sum_{i=1}^{N} c_i. \tag{2.4}$$

Another metric of interest is the path length, $d_{i,j}$, defined as the minimum number of edges that have to be traversed when going from v_i to v_j. This path is also known as the geodesic path between i and j. Using the geodesic distance we can formulate the characteristic path length, a global metric of distances in the graph. The characteristic path length is defined as the average path length between all vertices:

$$L = \frac{1}{N(N-1)}\sum_{i,j\in N, i\neq j} d_{i,j}. \tag{2.5}$$

[1] Newman [2003b] introduces a slightly different formulation for the clustering coefficient.

As discussed in the previous section, the characteristic path length and clustering coefficient are of great importance when trying to investigate the structure of a graph. Oftentimes, the (natural) networks of interest consist of many vertices with sparse connections. In this case *'sparse'* implies that the network consist of a low number of edges, but sufficiently more than whenever an edge at random would be removed the graph would be disconnected. Hence this means that $n \gg k \gg \ln(n) \gg 1$, where $k \gg \ln(n)$ guarantees the graph to be connected. Staring with a lattice ($p \to 0$), we find the theoretical values of $L_{ordered} \sim n/2k \gg 1$ and $C_{ordered} \sim 3/4$, indicating a highly clustered, large network where $L_{ordered}$ grows linearly with n. On the other hand, for a random graph as ($p \to 1$), we find $L_{random} \sim \ln(n)/\ln(k)$ and $C_{random} \sim k/n \ll 1$, a poorly clustered, small world (in terms of vertex popularity), where L_{random} only grows logarithmically with n. These extreme cases hint that a large characteristic path length is associated with a high cluster coefficient and a small characteristic path length with a low cluster coefficient. Then what about the range of networks that lie between the ordered and random graph as $0 < p < 1$? Watts and Strogatz [1998] showed that there is a whole range of networks, the small-world networks, for which L is almost as small as L_{random} but with $C \gg C_{random}$. The graphs thus have a very short path length between any two nodes, but with a high clustering coefficient. Small-world networks exhibit these characteristics especially for $0.001 < p < 0.1$. However, to date there is no known analytical solution of L and C as functions of p [Newman, 2003a].

Finally, there is another type of network that we should include in this analysis: the *scale-free* network. This network follows, at least asymptotically, a power law degree distribution, for which the ratio P of nodes in the network with k connections follows:

$$P(k) \sim ck^{-\gamma}, \tag{2.6}$$

where c is a normalization constant and γ a model parameter. It can be created when starting with a very sparse lattice and growing it using preferential attachment [Barabasi and Albert, 1999] or other models [Boccaletti et al., 2006]. A growth model such as preferential attachment is analogous to *the rich get richer* model and will be explained in the subsequent section.

It is important to point out that a small-world network is not a scale-free network, for the latter has a very small path length in the order of $\ln\ln(n)$, but not necessarily a high clustering coefficient like a small-world network [Cohen and Havlin, 2003].

2.3 Miscellaneous Topologies

The subsequent sections discuss several other topologies, similar or derived from the small-world network.

2.3.1 Preferential Attachment

In the previous section we briefly touched upon the preferential attachment network. It was proposed by Barabasi and Albert [1999], and the basic idea is to model a random

network with a scale-free degree distribution. A preferential attachment model can be created as follows. First, the edge degree d for each node is specified and we start out with a single node and let the network grow by adding nodes up to n. Each new node added to the network is connected to d other nodes. The nodes to which a new vertex v_i is connected is determined by a probability p that is proportional to the current degree of the nodes that are already in the network. In other words, the nodes which are added in the beginning are better connected than the node which has be added last. This is analogous to the *'rich get richer model'*, for the nodes which are added at the beginning tend to become the nodes with the highest edge degree.

2.3.2 Range Dependent Model

Another biological-inspired network, which is actually a generalization of the Watts-Strogatz small-world model, is the range dependent network. It was first described by Xenarios et al. [2000], who used it to model a protein-protein interaction (PPI) network. These interactions can be translated into a network where an edge (interaction) between two proteins is presented if the proteins have been observed, by experimentation, to interact. The ordering of nodes in proteins are always fixed, and as such, each pair of nodes (v_i, v_j) is connected with an edge with probability $\alpha\lambda^{|j-i|-1}$, where α regulates the number of connections and the geometric λ controls the ratio of long-range edges over short-range edges. That is, if v_i and v_j are far apart, the probability of having an edge joining the vertices is small. An example of such a graph is shown in Fig. 2.5.

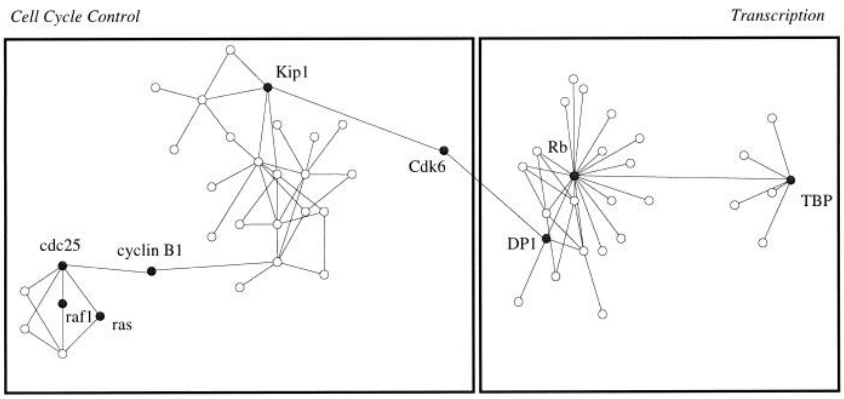


Figure 2.5: An example of a diagram of interacting proteins functioning in cell cycle [Xenarios et al., 2000]

Grindrod [2002] proposes a class of range dependent networks for modeling proteome networks. He explores the reordering property of range dependent graphs, that is, such as how to reorder a graph such that most vertices are closely connected and only few long paths exists. A numerical analysis based on this work is presented in Higham [2003]. Another application for the range dependent network is percolation theory [Xenarios et al., 2000].

2.3.3 Geometric Graph

The two-dimensional, non-periodic, *geometric random graph* with n nodes is generally defined as follows. Initially, each of the nodes are placed in a unit squared according to a uniformly random distribution. Formally, this implies that vertex v_i is placed at (x_i, y_i) where the every set of coordinates are independent and identically distributed with a $U(0,1)$ distribution. Next a radius r is specified, and each node is connected by an edge to every other node that is within a squared distance of the node. That is, given r, v_i and v_j are connected if and only if

$$(v_i(x) - v_j(x))^2 + (v_i(y) - v_j(y))^2 \leq r^2. \tag{2.7}$$

We now have a graph with connections which can be represented by an adjacency matrix, hereby ignoring the positions of the nodes since they are only required while the network is constructed. Note that the Euclidean distance can be replaced by another norm which is necessary to model the network. The network can be used to model many biological networks as shown by Przulj et al. [2004]. A very in-depth analysis and overview of geometric networks is presented in Penrose [2003].

2.3.4 Kleinberg Graph

In [Kleinberg, 2000] another variation of the Watts-Strogatz model is proposed. This model is called the Kleinberg model and particularly exploits the short-cuts in a network. In contrast to the geometric model which is non-periodically distributed in a two-dimensional space, the vertices in a Kleinberg model are equally spaced in the grid. Each vertex is placed at position $(v_i(x), v_i(y))$ equidistant from each of the nodes in each of the directions from it. Every vertex is then connected to every other vertex which are at most a Manhattan distance p away. To create short-cuts in the network, each node is then connected to m nodes of at least a distance p away with a probability uniformly proportional to $r^{-\alpha}$, where r is the distance and α are controls the probability.

2.4 Summary

This chapter provided a thorough review of graphs. We explored the conversion of a model into the abstract notion of a graph and how to 'measure' a graph. Several important metrics were then discussed such as the cluster coefficient, a metric that describes the 'connectedness' of direct neighbors for a given vertex, and the characteristic path length, the mean distance one needs to traverse between each set of nodes. These two metrics characterize an important class of networks known as the small-world networks; networks with a high clustering coefficient and a low characteristic path length. Finally, we discussed several networks which are generalizations, derivations or closely related to the small-world networks.

CHAPTER 3

PAGERANK

In this chapter we explore the PageRank algorithm. Why such an algorithm is necessary is given in Section 3.1. We then explore the computation of the PageRank vector of a matrix in Section 3.2. Finally Section 3.3 gives a brief introduction to the personalized PageRank, which will be further discussed in Chapter 5.

3.1 Introduction

For nearly three decades, the continuous growth of information on the world-wide web has been unprecedented. The current indexed size of the web is estimated to be at least 14.17 billion pages [de Kunder, 2011]. This number sounds even more dramatic when compared to the volume of the web in 1999, when the size was estimated at 800 million pages [Lawrence and Giles, 1999]. This shows a 17-fold increase of the volume that search engines are indexing in just 12 years. When the volume of the web exceeded several 100 million pages, two problems became apparent. First, there is the issue of the heterogeneous nature of content on the web. Whether it is somebody on the far end of the world who wrote a tweet or posted a video on YouTube about his breakfast or the latest news about natural disasters, you can find information about it on the web. Due to the changes in content on the web, from plain text to hypertext and other semantics, it became clear that search engines had to start looking for other information retrieval methods than simply the content of a website when responding to a user's search query for information. The second problem is the volume of the web. When the web reached several 100 million pages it became significantly more complex for search engines to index the entire web [Chakrabarti et al., 1999]. This problem is as relevant now as it was back then, considering the exponential growth of the web [LivingInternet]. Take the growth of Twitter for example [Twitter, 2011]. It took 3 years, 2 months and 1 day until

the billionth Tweet was sent on May 22, 2009. Now it takes less the 1 week for the users to send a billion Tweets. Another example is YouTube [ReelSEO, 2011], which was founded in February 2005. Currently, every minute there is 35 hours of video footage uploaded to the website. All this information has to be indexed by search engines before it can show up in search results.

The growth of the web forced the search engines to start relying on client-side filtering and analysis of results, and thus *personalized* search results became essential. In the late 1990's, Page et al. [1998] presented PageRank, a novel technique to identify what the important pages are on the web. PageRank became the foundation and the key factor resulting in the success of Google's search engine and a shift in discovering qualitative and personalized information on the web.

3.2 PageRank

In their well-known paper Page et al. [1998] presented the novel idea of the PageRank, a ranking method for every web page on the World Wide Web using the graph-structure of the web itself, with applications to browsing, traffic-estimation and of course search. PageRank is inspired by work in analysis of citations [Garfield, 1995] and research focussed on spectral clustering. Now looking back at these separate research efforts it is only natural that eventually one would combine these and develop a search engine. The question to ask would be: *"What would we need to build a personalized search engine?"*.

A search engine has to perform several tasks. To start with, it needs to collect and maintain information available in the multitudinous pages on the web. Obtaining information from websites is known as *crawling*.

3.2.1 The PageRank Matrix

We can view the web as a very large network, or *directed graph*, where the pages can be represented by nodes and a link from one page to another is analogous to an edge in the graph. In this representation we can denote each page v, corresponding to a node ($v \in V$) and every link on that page to another page as an edge e, ($e \in E$). The goal is to give every page in the web a PageRank value, which is a measure (or score) of importance. This score of a page is based on the links that are pointing to this web page. A link from page v_i to page v_j is called the *backlink* for page v_j. This can be viewed in the context of the academic citations: the importance of a publication is depended on the number of citations. Each citation indicates a vote of confidence in a publication and hence increases the importance score of the paper.

Let us consider a small web of n pages where each page is indexed by an integer k, where $1 \leq k \leq n$, like the one depicted in Fig. 3.1, for which there are five pages $\{v_1, \ldots, v_5\}$ and eight links. Obviously, not every page has a link to every backlink and therefore we can consider the web as a *directed graph*. The corresponding adjacency matrix to this

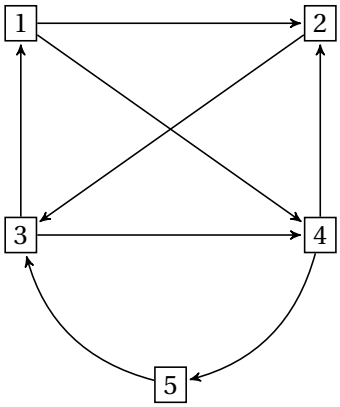


Figure 3.1: A small web consisting of five nodes and eight edges

graph is shown equation 3.1.

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}. \tag{3.1}$$

We will denote the importance score of each page k in the web by x_k. The importance of a page is non-negative and $x_j \geq x_i$ implies page x_j is at least if note more important than page x_i. To define the score of a page we can start by considering the number of backlinks:

$$x_k = \sum_{j \in B} x_j, \tag{3.2}$$

where B is the set of nodes linking to page x_k, i.e. the backlinks for page k. Then the network in Fig. 3.1 would produce a PageRank vector with the following scores:

$$\mathbf{x} = \begin{bmatrix} 1 & 2 & 2 & 2 & 1 \end{bmatrix}^\top. \tag{3.3}$$

This would imply that for this network the pages two, three and four are equally most important, whereas pages one and five are least important.

Obviously this is not a very good idea because it fails to take into account two properties. First, this ignores an intuitive property of an importance score: it is better to have a backlink from an important page than from an unimportant page Second, a person with bad intentions could easily create thousands of pages all linking to one page to improve its importance score.

The first problem can easily be solved when the importance of a page v_i is based on the importance scores of the backlinks. This yields the following set of equations:

$$\begin{aligned} x_1 &= x_3 \\ x_2 &= x_1 + x_4 \\ x_3 &= x_2 + x_5 \\ x_4 &= x_1 + x_3 \\ x_5 &= x_4 \end{aligned} \tag{3.4}$$

As such, every page will propagate a part of its importance through the network. Each of the equations corresponds to a row in the adjacency matrix (3.1).

The second problem can be solved when the PageRank value of page v_i is defined by the sum of the PageRank values for all pages linking to v_i normalized with respect to the total number of outgoings links for each page. This slightly changes the notation of 3.2, which now becomes:

$$x_k = \sum_{j \in B} \frac{x_j}{n_j}, \tag{3.5}$$

where n_j is the number of outgoing links from page v_j. This is reflected in the adjacency matrix:

$$\mathbf{P} = \begin{bmatrix} 0 & 0 & \frac{1}{2} & 0 & 1 \\ \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{3.6}$$

Each of the columns sums up to exactly one. Such a matrix is called a *column stochastic* matrix, which has convenient properties for the computation of the PageRank. Several propositions and theorems with proofs are given in [Bryan and Leise, 2006]. In fact, every column in a stochastic matrix sums up to 1, this formally implies that $P_{i,j} \geq 0$ and $\mathbf{e}^T\mathbf{P} = \mathbf{e}^T$, where $\mathbf{e}$ is a vector consisting of ones. Now, we want to find the vector $\mathbf{x}$ with importance scores that solves the equation

$$\mathbf{Ax} = \mathbf{x}. \tag{3.7}$$

This is apparently an easy eigenvalue problem, for which there are many mathematical tools available. We compute the eigenvalues of $\mathbf{A}$ and the eigenvector corresponding to eigenvalue equal to one (i.e., for which λ equals 1). After solving these equation and normalizing the eigenvalues, such that they sum up to one, we find the following importance scores, or PageRank values:

$$\mathbf{x} = \begin{bmatrix} 0.154 & 0.192 & 0.307 & 0.230 & 0.115 \end{bmatrix}^\top. \tag{3.8}$$

We can see that page three is the most important, or influential, page in this network with a PageRank value of 0.307, whereas page five is the least important page with a PageRank value of 0.115.

This seems like a nice accomplishment, for we are able to find an, apparent, valid set of PageRank values for our network. The real problem, however, is the enormous computation required to find the eigenvector of a matrix corresponding to the actual web with billions of entries. This procedure will also fail when there are *dangling* nodes in the graph, i.e., nodes which have an edge to themselves $\{\exists e \in E : (v_i, v_i)\}$. Even when we would assume that there are no dangling nodes (which we will for the remainder of this section), there is the problem of solving the eigenvector problem when the graph is disconnected, which is obviously the case when computing the eigenvector for the real web (cf. [Langville and Meyer, 2004] for a broad discussion on the structure of the web and implications on the PageRank algorithm).

3.2.2 The PageRank Equation

It is, however, possible to solve the problem (proof can be found in Bryan and Leise [2006]) of having a disconnected web (under the assumption that there are no dangling nodes). Let $\mathbf{S}$ denote a square matrix of length n for which all entries are $1/n$. Like the matrix $\mathbf{A}$, this matrix is column stochastic, and we replace matrix $\mathbf{A}$ in equation 3.7 by

$$\mathbf{M} = \alpha\mathbf{A} + (1-\alpha)\mathbf{S}. \tag{3.9}$$

The α parameter is known as the 'teleportation' parameter, and is oftentimes set to 0.85, which is the same value as Google reportedly used [Langville and Meyer, 2004]. A comprehensive discussion on the influence of this parameter can be found in Gleich [2009]. The matrix $\mathbf{M}$ is now a weighted average of $\mathbf{A}$ and $\mathbf{S}$ and is always column stochastic [Bianchini et al., 2005] and hence can be used to compute a PageRank vector with unique scores. If we let α be 0.85 we can solve our original problem using this method. This will yield the final PageRank values:

$$\mathbf{x} = \begin{bmatrix} 0.157 & 0.192 & 0.301 & 0.225 & 0.124 \end{bmatrix}^\top. \tag{3.10}$$

Even though the values are not the same as in equation 3.8, the ranking is maintained, but the algorithm no longer has the shortcoming of failure when the network is disjoint.

3.3 Personalized PageRank

The output of the original PageRank algorithm for computation of a ranking of Web pages for search-query results, is a single PageRank vector that represents the 'importance' for all pages. This PageRank vector is independent of the query and the output will be the same for whoever the query enters and irrespective of the query itself. Haveliwala [2002, 2003] proposed an algorithm to compute a set of PageRank vectors, each biased in a way such that it captures a different notion of importance with respect to a particular search query. Using these PageRank vectors, they generate topic-sensitive importance scores for pages that belong to the same category as the query. Using query-specific PageRank vectors is known as *personalization*. This was first suggested, however not explored, by Brin et al. [1998]. The main idea is to use a non-uniformly distributed vector $\mathbf{s}$, to bias the

computation of the PageRank vector. This yields the following, frequently used, formula for computing a personalized PageRank vector:

$$pr_\alpha(\mathbf{s}) = (1-\alpha)\mathbf{s} + \alpha pr_\alpha(\mathbf{s})\mathbf{W}. \tag{3.11}$$

$\mathbf{W}$ is the lazy random walk transition matrix $W = \frac{1}{2}(I - D^{-1}A)$, where D is the diagonal edge degree matrix. Chung and Zhao [2008] examine the relationship between PageRank and random walks on graphs. They propose a generalized version of personalized PageRank with an additional parameter that controls the rate of diffusion in the graph.

A more in-depth analysis of the computation of personalized PageRank vectors is presented in chapter 5. There we will give two algorithms to compute the vectors and how to use them to cluster a graph.

3.4 Summary

In this chapter we introduced the PageRank algorithm. We showed how it is derived and can be computed and what it means as a metric for a graph. Finally, we briefly touched upon the personalized PageRank.

CHAPTER 4

GRAPH THEORETICAL NEUROSCIENCE

We present a brief overview of the biological structure of the human brain 4.1, followed by a short introduction to fMRI 4.2 and how fMRI data can be transformed into the notion of structural and functional connectivity and how we can apply graph theory to get a better insight in the structure of the brain 4.3.

4.1 The Human Brain

The brain is the center and the most important part of the nervous system in animals. It is generally considered as the most complex organ in the human body. Although most of the time people talk about 'the brain' as a unified whole, neuroscientist identified several areas with specific functions within the brain. The brain can be considered at three different levels: micro, mesa and macro. Inasmuch as at the micro-scale individual neuronal connections are considered, at the macro level we are concerned with functional regions of the brain.

The brain is symmetrically divided in a left and right hemisphere. Both hemispheres are organized in three major, interconnected, components: the central core, the limbic system and the cerebral cortex. The first two are inside the cerebral cortex. These three components are shown in Fig. 4.1 and control almost all aspects of our daily life, whether it is movement, thinking or emotions. Every vertebrate's brain contains a central core, which, as the name implies, controls the basic functions that are required to stay alive. Examples include the control of basic and subconscious functions like breathing, arousal, managing the sleep process, moving and it is also the first stop for processing stimuli.

The **central core** (shown in Fig. 4.2) can be further divided in five different regions. The *medulla* rhythmically stimulates the diaphragm allowing us to breathe without think-

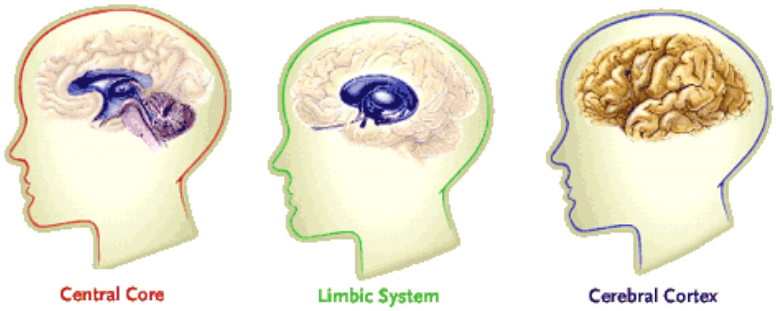


Figure 4.1: The three major parts of the cerebrum [Psychology]

ing about it. Next to that, it also regulates the blood flow and the heartbeat. The *pons* is some kind of intermediate station which carries sensory signals from the cerebral cortex to the cerebellum. The *reticular formation* receives sensory input (e.g., sound) and relays the signals to the thalamus and cerebral cortex. It is also allows us to dream when we sleep and being in a state of arousal. The *thalamus* is the first step in processing and relaying sensations and motor signals. It classifies stimuli in categories such as good/bad, then sends the signal to the cerebral cortex. Finally, the *cerebellum*, Latin for 'little brain', plays an important role in motor control. The cerebellum does not regulate movement but provides a support role to fine movement, equilibrium and motor learning. While it is very similar to the cerebral cortex, in that it has two hemispheres and a highly folded surface, its size is only 10% of the cerebral cortex. Nevertheless, the number of neuronal connections are about the same.

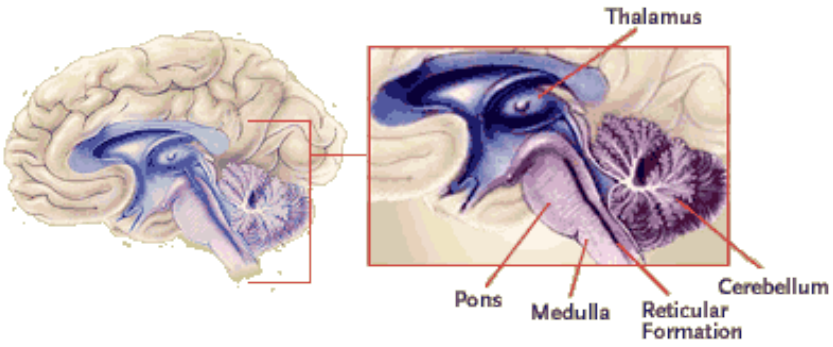


Figure 4.2: The five main regions in the central core [Psychology]

The **limbic system** (Fig. 4.3) is only found in mammals and, like the central core, buried in the cerebral cortex. The limbic system is often referred to as the 'emotional brain', because it allows mammals to express emotions. On the other hand it is responsible for regulation of important life processes such as the body temperature and blood pressure. The *hypothalamus* is involved in emotions that oftentimes lead to actions such

as emotions, thirst and hunger. The *amygdala* is responsible for emotional states such as anger and fear. The *hippocampus* manages the task of transforming the short term memory into a more permanent memory. The hippocampus also stores information regarding spatial relationships in the world around us.

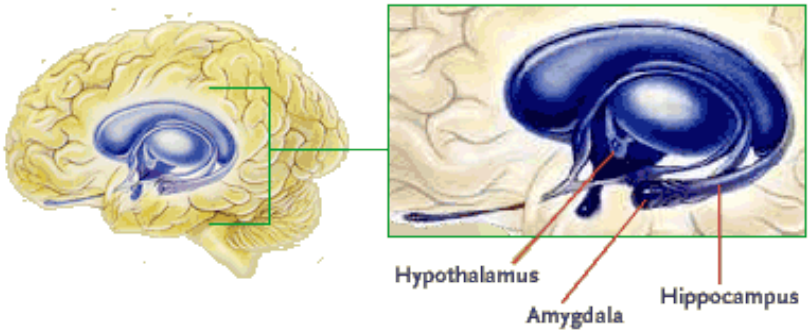


Figure 4.3: The three important regions in the limbic system [Psychology]

The **cerebral cortex** as shown in Fig. 4.4, the outer layer of the brain, is highly folded. This makes the brain more efficient because it significantly increases the surface area of the brain and therefore the number of neurons that it can contain. The cerebral cortex can be divided into two, anatomically almost symmetrically, hemispheres. It has, however, been shown that the functions of the left and right hemisphere are not equal. The left hemisphere is associated with logic abilities, whereas the right hemisphere has been shown to be associated with creativity.

The cerebral cortex can be subdivided into four lobules: the *Frontal Lobe*, the *Parietal Lobe*, the *Occipital Lobe* and the *Temporal Lobe*. Each of these lobules are known to control different functions. The frontal lobe contains the most dopamine-sensitive neurons of the cerebral cortex, which is associated with short-term memory, reward, attention and planning. In the parietal lobe are various sensory information systems integrated which allows us to make sense of spatial information and help us to navigate and manipulate objects. One of the most important functions of the occipital lobe is the processing of visual information and it therefore is the part where the dreams come from. Finally, the temporal lobe is the region in the cerebral cortex that is involved in auditory perception and processing. It contains the hippocampus and plays a very important role in the transformation of short-term information to long-term memories and knowledge.

Now that we have some idea of the different regions in the brain, we can discuss a technique to measure activity in the brain. More specifically, we will discuss functional Magnetic Resonance Imaging, a specialized technique that can measure neural activity in the brain.

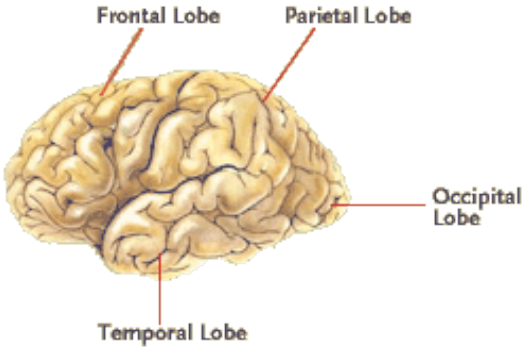


Figure 4.4: The cerebral cortex with the four important lobules [Psychology]

4.2 Functional Magnetic Resonance Imaging

Functional Magnetic Resonance Imaging (fMRI) is a relatively recent technique which can create functional brain maps in humans. In this section we will first discuss what fMRI is, followed by the procedure and techniques to create fMRI scans.

4.2.1 What is fMRI

In 1890, Roy and Sherrington [1890] performed experiments leading to the idea that cerebral blood flow (CBF) could reflect neuronal activity. This seminal publication is the basis for all of todays hemodynamic-based brain imaging research and techniques. Simply put, *an increase in CBF can be considered as a relation to increased neuronal activity* because the CBF and glucose metabolism are closely related [Raichle, 1987]. Hence when we measure the changes in CBF, we can create a mapping of regions in the brain which are activated during particular tasks.

Researchers at AT&T Bell Laboratories reported that functional brain mapping is possible using the venous *blood oxygenation level-dependent* (BOLD) Magnetic Resonance Imaging (MRI) contrast Ogawa et al. [1990b]. Changes in deoxyhemoglobin (dHb) result in an endogenous paramagnetic contrast which is the key for the BOLD contrast [Ogawa et al., 1990a]. The BOLD contrast is basically the difference in magnetic properties between oxygenated (diamagnetic) and deoxygenated (paramagnetic) blood. In other words, brain mapping using MRI relies on the localization of changes in the, by neural activity induced, magnetic environment. The area of functional brain mapping with MRI is continuously growing ever since it emerged about two decades ago. FMRI uses MRI equipment to detect changes in the cerebral metabolism or oxygenation while the subject is performing a specific task. At a neuronal level, it is known that neurons do not store glucose or oxygen, and therefore, whenever there is neural activity it requires glucose and oxygen to be delivered to the blood stream. In a process called the hemodynamic response, blood releases the glucose to the neurons, which results in a overabundance of oxyhemoglobin in the blood vessels. The contrast, or difference, in

oxyhemoglobin at the specific region of the brain which required the glucose for a specific activity, is measurable and is the key to MRI.

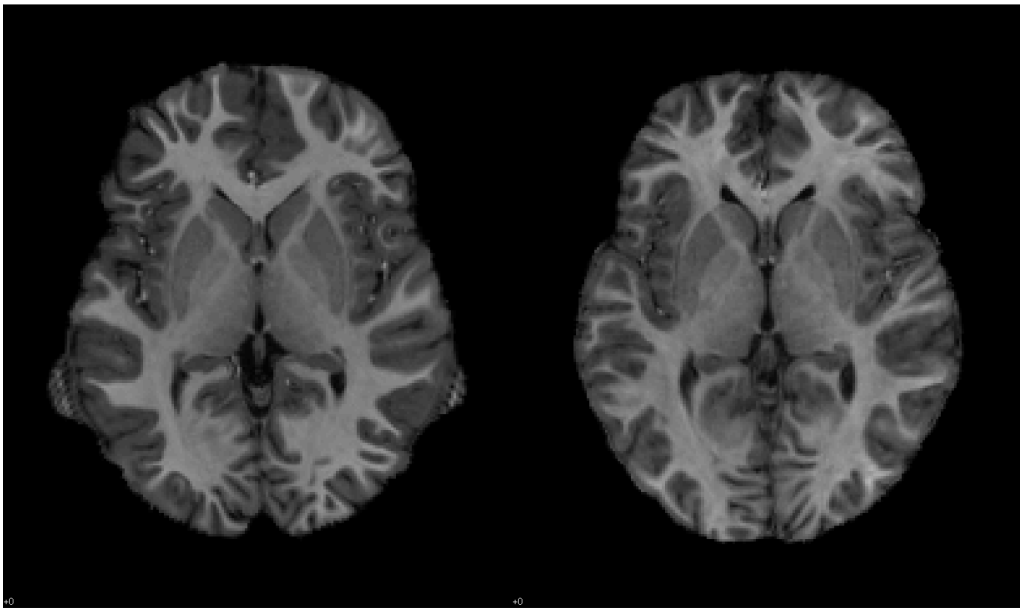
4.2.2 Processing and analysis

A single fMRI scan generates a vast amount of four-dimensional data which needs to be processed in order to make some sense of it. Most of the research that uses fMRI data is hypothesis-driven and voxel-based. That is, the researchers start out with a certain hypothesis and perform voxel-based group-comparison to determine whether different activation response exist. An analysis is always preceded by a multi-step pre-processing scheme which routinely consists of a number of typical steps [Nielsen et al., 2006]: spatial realignment of individual scans to correct for head motion, coregistration between the high resolution anatomical and low resolution functional scans, normalization of scans for all subjects, such that the scans are of the same size. Then the data is generally smoothened using a Gaussian filter, whereafter for each group a 'summary image' is generated. Finally, a statistical test can be performed on the voxels to determine whether there are any significant differences between the groups. There are numerous packages and tools publicly available for preprocessing and testing hypotheses, such as SPM, FSL, AFNI, LIPSIA and BrainVoyager.

The goal of voxel-based analysis is to identify activated brain regions for a given stimulus. The data is almost always gathered using the BOLD contrast. The block design or block-scheme is the most common experimental setup. In this setup the participant performs, for a fixed time, a task or is subjected to a stimuli. Then the participant is scanned while in a resting state. This scan is used as a baseline to which the scans acquired during the task or stimuli can be compared to. Statistical parametric mapping or multi-voxel pattern analysis is then performed to detect small, significant differences in activations that correspond to the regions being used in the brain.

4.3 Functional and Structural Connectivity

The brain can be analyzed from two broad categories: from a structural and a functional point of view. Structural brain imaging aims to map the coarse structure of the brain and is used for the diagnosis of large-scale intracranial pathophysiological diseases, such as tumors, hemorrhages and blunt traumas. Structural, or anatomical, brain images are static images of the anatomy and generally of a much higher resolution than the functional images. The structural images can also be used to measure the diffusion with a recently developed technique called Diffusion Tensor Imaging (DTI). Functional imaging is used to measure brain functions with the goal of understanding the relationship or correlation between activity in a specific brain area and a specific task or stimuli. An example of two structural images is shown in Fig. 4.5. The left image shows the central axial slice of a typically developing male. The right image displays the slice of a male who has been diagnosed with attention deficit hyperactivity disorder. However, in this thesis we are specifically interested in the functional imaging.

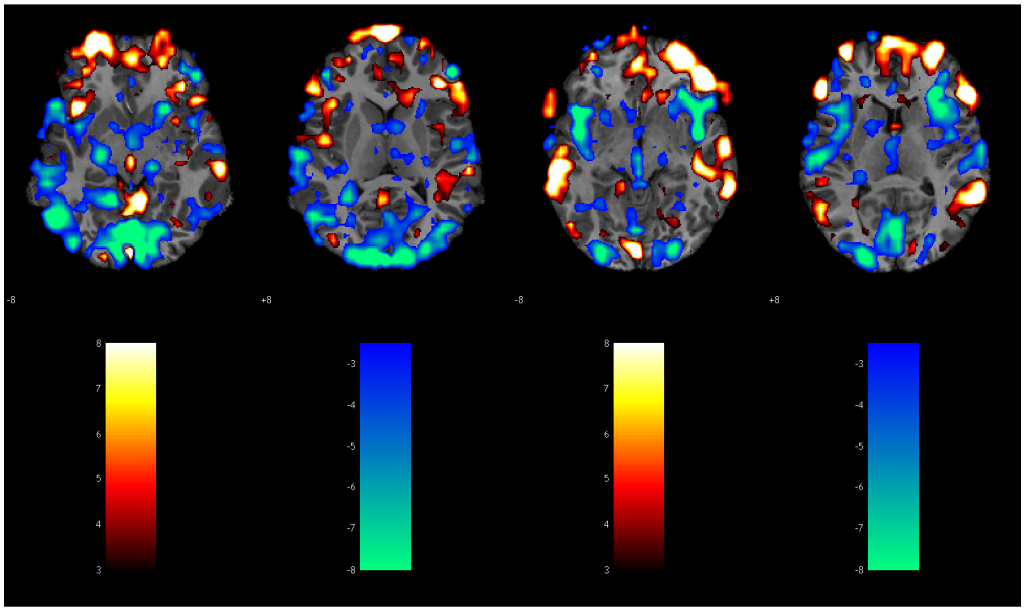


(a) Typically Developing (b) ADHD

Figure 4.5: Structural Brain Images

In contrast to structural imaging, functional imaging creates images of the 'activity' (the hemodynamic response) of the brain over time. This response is actually rather slow, and it may be lagging for a period as long as four or five seconds. The lag is, however, relatively stable and can be compensated using a linear time-invariant filter during preprocessing. The temporal resolution of a scan is limited by the number of measurements a scanner can make (this is called the repetition time, TR) and bounded by the volume that needs to be scanned. The TR is typically between one and three seconds. With standard scanners the voxels in which the brain is segmented can be as small as 1 millimeter and the state of the art scanners even have a higher spatial resolution. The BOLD changes generally cover a larger area than the specific neural activity which is one of the limitations for the spatial resolution. Similarly to the anatomical images shown above, we can create an activity map for the typically developing and ADHD subject, shown in Fig. 4.6. The regions which are colored according to the 'hot' color table (left), show an positive hemodynamic response, whereas the regions colored with the 'winter' scheme show a negative response.

One of the questions explored in this thesis is whether it is possible to classify the diagnosis of children based on their functional brain connectivity maps.



(a) Typically Developing (b) ADHD

Figure 4.6: Functional Brain Images (left colorbar indicates an increase in hemodynamic response, whereas the right indicates a decrease for that region)

4.4 The Brain and Graph Theory

The area of computational neuroscience is becoming increasingly popular over recent years. In order to get a better understanding and more insight into the organization of the brain and the functions, understanding the development of the brain is critical. Understanding the functional and structural brain is especially important when investigating disorders such as attention-deficit/hyperactivity disorder (ADHD), autism spectrum disorders (ASD), Alzheimer's and Parkinson. These disorders have in common that the normal developmental process is disrupted.

The goal of this section is to give a brief overview of recent applications of graph theory to neuroscience. In chapter 2 we discussed the class of networks known as small-world networks. These networks can relay information very fast because of the short path length and the high cluster coefficient. This suggest that this type of network is very beneficial for information processing in neural networks. There are three approaches to get a better understanding of the human neuronal network: simulations, real neuroanatomical networks using *in-vitro* and *in-vivo* analysis and functional networks. The analysis using simulations is outside the scope of this thesis and will not be discussed.

4.4.1 Neuroanatomical Networks

Coincidentally, the paper by Watts and Strogatz [1998] was also the first paper to explore the brain from a graph theoretical point of view. They created a complete graph of the anatomical connectivity in the nervous system of the *Caenorhabditis elegans*. This graph consists of 282 vertices connected by an average of 14 edges per vertex. The characteristic path length is $L = 2.65$ and has a cluster coefficient $C = 0.28$.

After this initial evidence for the existence of small-world networks in the nervous system, researchers investigated the nervous systems in vertebrates and primates. Hilgetag et al. [2000] studied the compilations of corticocortical connections in the visual, somatosensory and whole cortex from the macaque and the cat. They found the complex networks to be organized into densely intra-connected clusters of areas and concluded that the networks processes all properties of small-world networks. Sporns and Zwi [2004] studied the datasets also explored datasets of the macaque and cat cortex. They found little evidence that the graph has a scale-free degree distribution and noted that the small-world architecture must play an important role in information processing in the cortex.

4.4.2 Functional Networks

Supekar et al. [2009] explored the development of small-world properties, including the path length and cluster coefficient, in children and young adults. In both groups they found a similar small-world organization at the global level and significant differences in hierarchical organization and interregional connectivity. The efficiency of the brain using small-world properties was investigated by Achard and Bullmore [2007] and found that the efficiency decreases as the age increases.

4.5 Summary

This chapter provided a general top-down overview of different anatomical regions in the brain. How we can make functional images of the brain using fMRI was then introduced, followed by an explanation of the differences between functional and structural connectivity. Finally we provided some references to research that use graph theory to analyze the brain from a functional or structural perspective.

Part II

Clustering the Brain

CHAPTER 5

PAGERANK CLUSTERING ALGORITHM

The problem of finding a cut in a graph or network where the ratio between the number of edges crossing the cut and the volume of the smaller side of the cut is small is one of the main problems in algorithm design in network theory. There is much literature published on this topic with applications in many areas. In this chapter we first explain the traditional k-means algorithm and the initial seeding improvements of k-means++ in Section 5.1. Section 5.2 presents two algorithms to approximate the personalized pagerank for a node in a graph. Then in Section 5.3 a novel algorithm coined K-prmeans is presented, that uses the personalized pageranks to cluster data represented by a (small-world) network.

5.1 Traditional K-means

K-means is the most frequently used method for unsupervised cluster analysis. The basic objective of the algorithm is to partition a set of observations into different clusters where each observation is assigned to the nearest 'mean', represented by a centroid (cluster center). Formally, given a set of n observations $\{x_1, x_2, \ldots, x_n\}$ in a m-dimensional space represented by the features in each of the observations, the goal of the k-means algorithm is to partition the observations into k different sets. The number of sets, or clusters, $C = \{C_1, C_2 \ldots, C_k\}$ is at most equal to the number of observations ($k \leq n$) such that it minimizes the sum of squares within each of the clusters. That is, it minimizes

$$\underset{C}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x_j \in C_i} \left|\left|x_j - \mu_i\right|\right|^2, \tag{5.1}$$

where μ_i is mean of all points assigned to cluster C_i. The traditional k-means algorithm was first suggested by Seth Loyd in 1957, never formally published until 1982, [Lloyd,

1982]. The algorithmic procedure for traditional k-means is shown in Algorithm 1.

1: Place k centroids $C = c_1, \ldots, c_k$ at random in the n-dimensional space represented by the objects being clustered.
2: **repeat**
3: {Assign each object to the nearest centroid according to some distance metric}
4: $C_i^{(t)} = \left\{ x_j : \left|\left| x_j - m_i^{(t)} \right|\right| \leq \left|\left| x_j - m_{i^*}^{(t)} \right|\right| \text{ for all } i^* = 1, \ldots, k \right\}$
5: {Move the clusters to the center of the objects assigned to it}
6: $m_i^{(t+1)} = \frac{1}{|C_i^{(t)}|} \sum_{x_j \in C_i^{(t)}} x_j$
7: **until** positions of the clusters do not change

Algorithm 1: Traditional k-means algorithm

While the algorithm is very easy to implement, efficient and applicable to many different datasets, there are several drawbacks that make it less than perfect. Even in this simple algorithm we can identify three major drawbacks: (1) algorithmic complexity, (2) Distance metric and (3) location of the clusters. The algorithm is an *NP*-hard problem in Euclidean space even for two clusters [Dasgupta, 2008] and for any number of clusters even in the two-dimensional plane [Mahajan et al., 2009]. The algorithm is implicitly based on spherical clusters that are separable such that the centroids converge towards the cluster centers. For some datasets the algorithm works well, whereas it fails completely on others. Due to the random initialization of the cluster centers, the clustering can be arbitrarily bad compared to optimal clustering.

The third problem was addressed by Arthur and Vassilvitskii [2007]. They proposed a modified version of k-means algorithm know as *k-means++*, an approximation for the NP-hard problem of finding the right clusters. The selection of the initial clusters now works under the assumption that the spread between cluster centers is likely to be large. The first data point is uniformly chosen from all data points to be the cluster center. Every subsequent cluster center is then chosen with a probability proportional to the squared distance to the previously selected centroids. The full algorithm is shown in Algorithm 2. K-means++ has been shown to be a huge improvement and to outperform traditional k-means.

The k-means++ algorithm can also be used to cluster graphs or meshes. The algorithm works reasonable when using normal graphs or networks but fails to produce good clusters when the network is a small-world network because the path length tends to be very small and the cluster coefficient high. In other words, the (Euclidean) distance metric fails because the distance between each pair of nodes is generally very short. To overcome this problem we need to come up with a new distance metric that is not vulnerable to this problem.

1: Assign the first centroid c_1 from the centroids $C = \{c_1, \ldots, c_k\}$ uniformly at random to one of the objects.
2: **for** $i = 2 \rightarrow k$ **do**
3: choose $c_i = x' \in X$ as the next center with probability $\frac{D(x')^2}{\sum_{x \in X} D(x)^2}$
4: **end for**
5: **repeat**
6: {Assign each object to the nearest centroid according to a distance metric}
7: $C_i^{(t)} = \left\{ x_j : \left|\left| x_j - m_i^{(t)} \right|\right| \leq \left|\left| x_j - m_{i^*}^{(t)} \right|\right| \text{ for all } i^* = 1, \ldots, k \right\}$
8: {Move the clusters to the center of the objects assigned to it}
9: $m_i^{(t+1)} = \frac{1}{|C_i^{(t)}|} \sum_{x_j \in C_i^{(t)}} x_j$
10: **until** positions of the clusters do not change

Algorithm 2: K-means++ algorithm

5.2 Approximating Personalized PageRank

In Section 3.2 we described how to compute the PageRank vector for a graph; a metric to describe the 'importance' of each page (vertex) in each graph. Intuitively we can expect the nodes with many incoming links to have a higher PageRank value than a vertex with few incoming links. As a consequence when the graph is visualized, the pages with a high PageRank are probably in the center of many nodes, surrounding it with many incoming edges.

We briefly described the possibility of computing a personalized PageRank vector for a graph in Section 3.3. The computation of a personalized PageRank requires a non-uniformly distributed initial PageRank value for each node in the graph. The vector **s** in equation 5.2 shows an example of an initial personalized PageRank vector. The first node in this graph is initially assigned a higher PageRank than the other vertices.

$$\mathbf{s} = \begin{bmatrix} 0.5 & 0.1 & 0.2 & 0.1 & 0.1 \end{bmatrix}^\top. \tag{5.2}$$

Having a non-uniform distribution of **s** allows the bias according to a user-specified set of important pages. Another example 5.3 shows an initial PageRank vector for node v_2.

$$\mathbf{s} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \end{bmatrix}^\top. \tag{5.3}$$

This initial vector has a special property: upon convergence of the PageRank computation, the result is a PageRank vector representing the relative importance for each node in the graph with respect to v_2.

Jeh and Widom [2003] first introduced a technique for approximating a PageRank vector. [Andersen et al., 2006] showed an algorithm based on Jeh and Widom [2003] to compute a PageRank vector with a teleport probability α for finding a local cut in a graph.

The easiest way to do approximate a PageRank vector is by means of a geometrical sum. Consider a graph $\mathcal{G}(V, E)$ with n vertices, $V = \{v_1, \ldots, v_n\}$ and k edges between the

vertices. All the connections can be described by a (undirected[1]) connection matrix A, where $a_{i,j} = 1$ if there is an edge joining v_i and v_j. Let D be the diagonal matrix where $d_{i,i} = d_k$, the degree of vertex v_i. From Section 3.2 we know that we can compute the PageRank as

$$pr_\alpha(s_i) = (\alpha\mathbf{A} + (1-\alpha)s_i)pr_\alpha.$$

This is equal to a personalized PageRank vector for a vertex v_i which can be computed by the geometrical sum:

$$pr_\alpha(s_i) = \alpha \sum_{k=0}^{\infty} (1-\alpha)^k s_i W^k, \tag{5.4}$$

where α is the 'teleportation parameter', which is in general set to 0.85, s is a zero vector of length n with $s_i = 1$ and W is the random walk transition matrix $W = D^{-1}A$. An implementation example and proofs of several propositions can be found in Andersen et al. [2006]. Another way to approximate the personalized PageRank vector uses the heat kernel, with a temperature parameter $t \geq 1$ instead of α.

$$pr_t(s_i) = e^{-t} \sum_{k=0}^{\infty} \frac{t^k}{k!} s_i W^k. \tag{5.5}$$

The difference comes down to the computation of the geometric sum in 5.4 versus the exponential sum in 5.5. The heat-kernel PageRank formula consists of two parameters t and the initial PageRank distribution s_i. Basically, the heat-kernel PageRank algorithm computes the exponential sum of random walks from the initial distribution (s_i), regulated by the temperature t. The advantage of the heat-kernel PageRank is the satisfaction of the heat-kernel equation (5.6) and that a subset of vertices S satisfies the Dirichlet boundary conditions. Proofs of derivations, the Dirichlet boundary conditions and other properties can be found in [Chung, 2009].

$$\frac{\partial}{\partial t}\rho_{t,s_i} = -\rho_{t,s_i}(I - W). \tag{5.6}$$

5.3 K-PRMEANS

Now that we have two approaches for computation of a personalized PageRank vector, we will discuss how we can use them to obtain a partitioned graph. Consider the computation of a personalized PageRank matrix P, where the p_i is the personalized PageRank vector computed using either the geometric or heat-kernel approximation algorithm. We can now introduce a new distance metric which we call the *pagerank-distance*. Pagerank-distance is defined as the 'distance of two personalized PageRank vectors', i.e., given two personalized PageRank vectors, p_i and p_j we compute the distance l as:

[1] the graph can also be directed, but for sake of simplicity we assume an undirected graph.

$$l = \left|\left| p_i \cdot d(v_i)^{-1} - p_j \cdot d(v_j)^{-1} \right|\right|, \tag{5.7}$$

where $d(v_i)^{-1}$ is the inverse edge degree of vertex v_i. Iterating over all sets of vertex pairs (v_i, v_j) yields the symmetric distance matrix L, where $l_{i,j}$ is the pagerank-distance between vertices i and j. The major advantage of the pagerank-distance matrix is that the importance of vertices is preserved.

Combining this procedure with the k-means++ algorithm, using the advantageous initial cluster seeding, gives the following algorithm 3 which we call *k-prmeans*.

Require:
A: $n \times n$ connection matrix
k: number of clusters.
1: **for all** $v_i \in V$ **do**
2: {compute a personalized pagerank vector using the geometrical sum or heat kernel}
3: $p_i \leftarrow pagerank_\alpha(v_i)$ or $p_i \leftarrow pagerank_t(v_i)$
4: $P = [p_1|p_2|\ldots|p_n]$
5: **end for**
6: **for all** $p_i, p_j \in P$ **do**
7: {compute the pagerank distance matrix L}
8: $l_{i,j} = \left|\left| p_i \cdot d(v_i)^{-1} - p_j \cdot d(v_j)^{-1} \right|\right|$
9: **end for**
10: Assign the first centroid c_1 from the centroids $C = \{c_1, \ldots, c_k\}$ uniformly at random to one of the objects.
11: **for** $i = 2 \rightarrow k$ **do**
12: choose $c_i = x' \in X$ as the next center with probability proportional to $\frac{D(x')^2}{\sum_{x \in X} D(x)^2}$
13: **end for**
14: **repeat**
15: {Assign each object to the nearest centroid according to the pagerank distance}
16: $C_i^{(t)} = \left\{ x_j : \left|\left| x_j - m_i^{(t)} \right|\right| \leq \left|\left| x_j - m_{i^*}^{(t)} \right|\right| \text{ for all } i^* = 1, \ldots, k \right\}$
17: {Move the clusters to the center of the objects assigned to it}
18: $m_i^{(t+1)} = \frac{1}{|C_i^{(t)}|} \sum_{x_j \in C_i^{(t)}} x_j$
19: **until** positions of the clusters do not change
20: **return C**, the vector with cluster assignments for all vertices.

Algorithm 3: K-prmeans algorithm

Although the novel seeding algorithm of k-means++ is incorporated in k-prmeans, we are still dealing with a probabilistic selection of initial cluster centers. We can make

two observations. First, if the algorithm starts with a very inconvenient set of cluster means, the likelihood of producing a good clustering C is low. Second, if there is some intrinsic structure in the graph, there is a higher probability that, when the algorithm is run several times, the number of different outcomes are lower than when a random graph is clustered. Third, if there exists natural clusters in the data, the number of edges that are in the cut are low. With these remarks in mind, we propose the *stable-k-prmeans* algorithm that bridges these issues. This algorithm can be build as a wrapper with a function call to the k-prmeans algorithm.

We will constructively introduce the stable-k-prmeans algorithm. The first issue can be partially solved by repeated clustering, i.e., performing the k-prmeans algorithm several times. The second problem can be solved by computing the probability distribution for each unique outcome. Algorithm 4 addresses both these issues.

Require:
A: $n \times n$ connection matrix
k: number of clusters
h: number of k-prmeans iterations
1: **for** $i = 1 \rightarrow h$ **do**
2: $c_i \leftarrow k - prmeans(A, k)$
3: $C = [c_1|c_2|\ldots|c_h]$
4: **end for**
5: {Determine the unique set of cluster outcomes (D) and the corresponding probability distribution p.}
6: $D \leftarrow unique(C)$
7: **for** $i = 1 \rightarrow length(D)$ **do**
8: $p_i \leftarrow \frac{\text{frequency} D_i \in C}{h}$
9: **end for**

Algorithm 4: First version of the stable-k-prmeans algorithm

We approach the third problem separately using the notion of *conductivity*. First we define the volume of a subset $S \subseteq V$ as

$$\text{vol}(S) = \sum_{v_i \in S} d(v_i). \tag{5.8}$$

Obviously, the volume of an undirected graph is $vol(V) = 2m$, where m is the number of edges in the graph. The *edge boundary* of a subset $S \subseteq V$ is defined as the set:

$$\partial(S) = \{\{x, y\} \in E | x \in S, y \neq S\}. \tag{5.9}$$

Then the conductance of a set is:

$$\Phi(S) = \frac{|\partial(S)|}{\min(\text{vol}(S), 2m - \text{vol}(S))}. \tag{5.10}$$

Basically the conductance of a set describes the ratio of edges in the cut to the lesser of the total number of edges in the set and its complement. A lower conductance implies that there are less edges in the cut and hence means that the vertices in the set are better connected. Therefore to find well-structured clusters, the cluster C_i needs to have a low average conductivity and has to occur frequently. The complete algorithm is shown in Algorithm 5. The next chapter discusses experimental results using the stable-k-prmeans algorithm.

Require:
A: $n \times n$ connection matrix
k: number of clusters
h: number of k-prmeans iterations
1: **for** $i = 1 \rightarrow h$ **do**
2: $c_i \leftarrow k - prmeans(A, k)$
3: $C = [c_1|c_2|\ldots|c_h]$
4: **end for**
5: {Determine the unique set of cluster outcomes (D), corresponding probability distribution p and the conductivity E for each cluster.}
6: $D \leftarrow \text{unique}(C)$
7: $t \leftarrow \text{length}(D)$
8: **for** $i = 1 \rightarrow t$ **do**
9: $p_i \leftarrow \frac{\text{frequency} D_i \in C}{h}$
10: **for** $j = 1 \rightarrow k$ **do**
11: $E_{j,i} \leftarrow \Phi(D_i(j))$
12: **end for**
13: **end for**
14: $C \leftarrow \min\{\bar{E_1} \times p_1, \ldots, \bar{E_t} \times p_t\}$
15: **return C**, the vector with cluster assignments for all vertices.

Algorithm 5: Improved version of the stable-k-prmeans algorithm

5.4 Summary

In this chapter we provided an introduction to the k-means algorithm and the improvement known as k-means++. These algorithms can produce good results, although it depends a lot on the domain. The algorithms fail when the data is a network of the small-world class, for which the average path length is very short. We noted that vertices which are the cluster centers oftentimes have many edges and are likely to have a high PageRank (cf. chapter 3). With this in mind we introduced a new distance metric: the PageRank-distance between two vertices based on the two corresponding personalized PageRank vectors. Using this distance we proposed the k-prmeans algorithm, to cluster a graph using PageRank distances with (better) initial seeding. Finally, to over-

come the disadvantage of probabilistic initial seeding of the cluster means and to define a 'good' cluster numerically we suggested the stable-k-prmeans algorithm. This algorithm clusters a network repeatedly and uses the frequency and conductivity of clusters to determine the best output.

CHAPTER 6

K-PRMEANS EXPERIMENTS

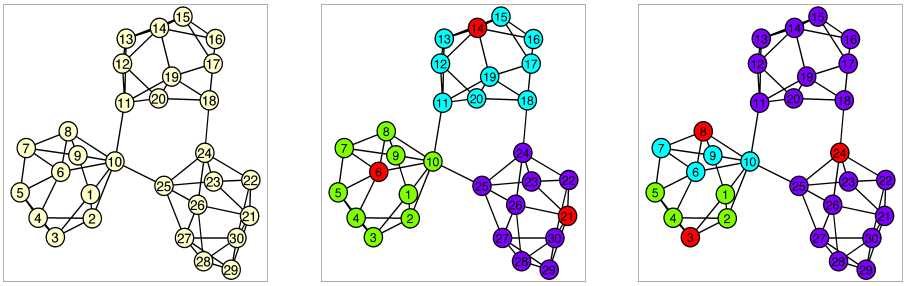
In this chapter we use the k-prmeans algorithm, as described in the previous chapter, to cluster several different graph models that belong to the small-world class. For each model, we let the cluster coefficient and the characteristic path length be in the same range as for small-world networks. Visual inspection sometimes reveals what clusters would be good clusters and we test this using the cluster algorithm. To investigate the advantage of initial seeding with k-prmeans, we compare the frequency of the 'optimal' partitioning outcome for each graph to the frequency of the 'optimal' outcome with random seeding.

6.1 Small-World

Consider the artificial small-world network depicted in Fig. 6.1. It consists of 3 small-world graphs with a total of 30 nodes, each node is connected to its immediate two neighbors and each connection is randomly rewired with probability 0.2. The cluster coefficient of this graph is 0.45 and the characteristic path length is 2.98.

We are interested in how often the algorithm would cluster this graph in the three correct partitions. We found that after 200 iterations the *k-prmeans* algorithm found 11 different solutions of which 177 times the 'correct' solution, shown in Fig. 6.1b. The average conductivity when the graph would be partitioned in this way is 0.0448. The second most frequently occurring solution is shown in Fig. 6.1c. The algorithm only resulted five times in this partitioning, which has a mean conductivity of 0.2263.

These results show a big improvement when we compare the frequency of the optimal solution to the frequency when the initial seeding was chosen at random. That way the optimal clustering is 'only' found 137 times out of 200 test runs and the second best solution 13 times. Without the improved seeding, the algorithm also found a larger



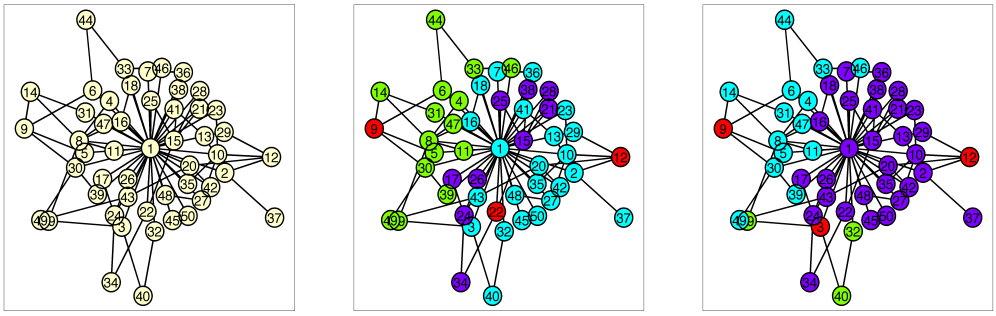
(a) Unclustered Graph (b) Clustered Graph (c) Suboptimal Solution

Figure 6.1: Small-World Network ($n = 30$)

number of different solutions, namely 16.

6.2 Preferential Attachment

The second model we tested is the preferential attachment model as described in Section 2.3.1. A graph was generated with 50 nodes and an average path length is 2.1880, much lower than the small-world network. It has a cluster coefficient of 0.5075. The cluster coefficient is slightly higher than that of the small-world network because it is a scale-free graph, for which the nodes are added successively. The unclustered graph is shown in Fig. 6.2a.



(a) Unclustered Graph (b) Clustered Graph (c) Suboptimal Solution

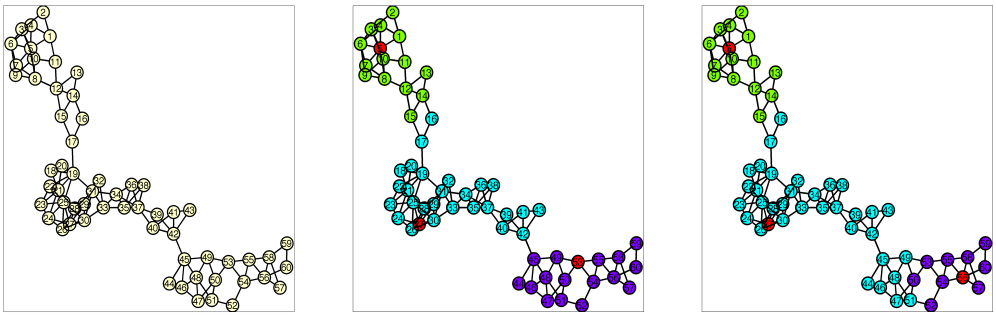
Figure 6.2: Preferential Attachment ($n = 50$)

The *k-prmeans* algorithm found 39 different solutions of which the most frequent solutions were found after 53 and 30 iterations. They are shown in Fig. 6.2b and Fig. 6.2c, respectively. The mean conductivity of these clusters are very close together, 0.3252 and 0.3203, but it is clear from the figures that the most frequent solution is also the best solution because the number of nodes is more evenly distributed throughout the clusters.

When clustering the graph without initial seeding, we obtain similar results. The algorithm then finds 35 different solutions, of which the best solution is found 69 times and the second-best 21 times. This implies that in a scale-free graph the initial seeding is not necessarily better than random seeding.

6.3 Renga

We also tested the algorithm on a graph generated using the RENGA algorithm (cf. 2.3.2). The λ parameter (base for edge probability) was chosen to be 0.6 and the α parameter as 0.9. These settings yield a graph with an average cluster coefficient of 0.44 (similar to the previously discussed Small-World Network), but with a higher characteristic path length of 6.54. The higher path length is clearly visible in Fig. 6.3a.



(a) Unclustered Graph (b) Clustered Graph (c) Suboptimal Solution

Figure 6.3: Renga Network ($n = 60$)

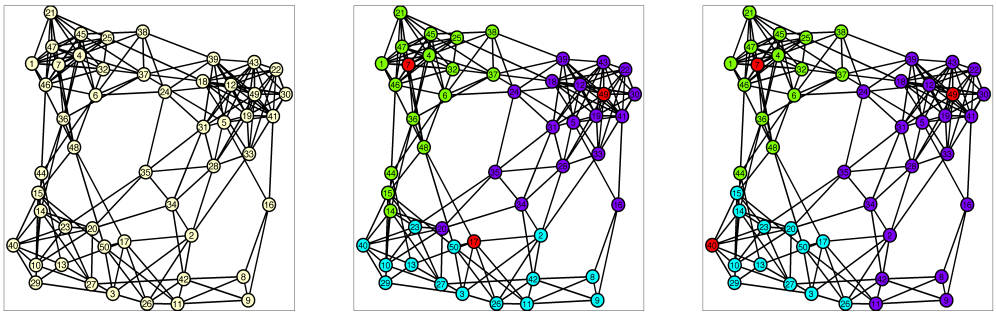
After 200 iterations the best clustering out of the 16 different solutions found with *k-prmeans* is shown in Fig. 6.3b. It has an average conductivity of 0.0254 and occurred 73 times. The second best solution, Fig. 6.3c appeared 51 times and has an average conductivity of 0.0876. Visual inspection supports that the most occurring solution is intuitively a better one, because (1) it cuts through less edges and (2) the size of the clusters are more balanced.

When the initial cluster locations are chosen at random, the algorithm finds 15 different solutions. Only 52 iterations result in the best outcome. The second best solution is found 26 times. All the other solutions of lesser quality appear more frequent, which shows the clear advantage of the initial seeding.

6.4 Geometric Network

The next model we will discuss is the geometric graph. The simulated model consists of 50 nodes. The characteristic path length is 2.6568 and the cluster coefficient is 0.5579.

These values are slightly below and above the values of the small-world network respectively.



(a) Unclustered Graph (b) Clustered Graph (c) Suboptimal Solution

Figure 6.4: Geometric Network ($n = 50$)

The best solution, shown in Fig. 6.4b, was found 34 times by *k-prmeans* and has a mean conductivity of 0.1810. The partitioning shown in Fig. 6.4c is the second-best solution. This solution, however, has a mean conductivity of 0.1496. Visual inspection tells us why the best solution is slightly the better of the two: the graph is more balanced.

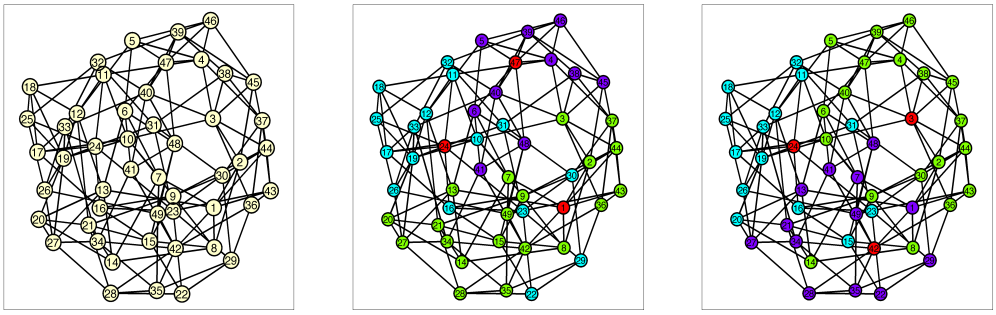
Running the algorithm with random cluster seeding resulted 38 different solutions after 200 iterations. Both the best and the second best solutions were found 28 times. Such an output isn't completely surprising given the large number of different solutions and because they are both pretty good. Non-random initial seeding shows again the better results.

6.5 Kleinberg Graph

The next graph we examined is the Kleinberg model consisting of 50 nodes. This model contains much more edges than a traditional small-world network, but also explores the short path length between nodes far apart. In this example there are 284 edges with an average path length of 2.4856 and a mean cluster coefficient of 0.2252. The dense graph is shown in Fig. 6.5a.

In this graph the clusters are not very clear and after 200 iterations of the *k-prmeans* algorithm there are 79 different solutions found. However, the particular partitioning shown in Fig. 6.5b was the outcome of 26 iterations, which is significantly more frequent than the other solutions. Not unexpected, this solution also has the lowest conductivity 0.2795. The second frequent occurring solution, 14 times, is shown in Fig. 6.5c, which might seem better, but has a much higher conductivity of 0.3544.

The algorithm with random seeding yields 83 different results of which 23 iterations yielded the best results and 14 times the second-best result.



(a) Unclustered Graph (b) Clustered Graph (c) Suboptimal Solution

Figure 6.5: Kleinberg Network ($n = 50$)

6.6 Summary

In this chapter we investigated the experimental results of the k-prmeans cluster algorithm with and without initial seeding. In many cases the output of the stable-k-prmeans algorithm corresponded to visual inspection of the graph. In each experiment we found better results when we used the initial seeding advantage, because the initial positions of the cluster means are less likely to be close together. The fact that the frequency of the best clusters are not coming from a uniform distribution and the combination of frequency and conductivity show the advantage of the stable-k-prmeans algorithm.

Part III

The Small-Worldness of ADHD

CHAPTER 7

GRAPH THEORETICAL PROPERTIES OF ADHD

In this chapter we will investigate the graph theoretical properties of ADHD. The chapter starts with a brief description of the data in Section 7.1. In Section 7.2 we explain and exemplify how we transform fMRI data into a graph. One of the parameters the transformation depends on is a threshold parameter, which influence is explored in Section 7.3. Finally, Section 7.4 provides an in-depth physiological analysis of the cluster coefficient.

7.1 Data

The NeuroBureau[1] is an initiative that is trying to get more people involved in computational Neuroscience by releasing pre-processed datasets. Over the past few months they have been releasing parts of a large dataset of fMRI scans of children and teenagers. These scans are part of the *ADHD 200 Sample*[2], an initiative dedicated to accelerating the understanding of the neural basis of ADHD through open data-sharing. This is supported by the unrestricted public release of 776 resting-state fMRI and anatomical scans. These scans were recorded at eight different sites such as universities (Brown University, New York University and Peking University), and research institutes (Kennedy Krieger Institute). Out of the 776 scans, there are 491 obtained from typical developing young adolescents (*TD*) and 285 from children and teenagers with ADHD. Every scan from each site was pre-processed in the same way, using scripts and tools that are publicly available. Processing every scan in the same way allows for easier comparison between scans. Besides the scans, they also released phenotypic information about the participants such as the diagnostic status, ADHD symptoms, age and gender.

[1] http://neurobureau.projects.nitrc.org/NeuroBureau/Welcome.html
[2] http://fcon_1000.projects.nitrc.org/indi/adhd200/

A few months ago, the *ADHD 200 Global Competition* was announced. This is an open competition for researchers across all disciplines. The results presented in the next chapter of this thesis are submitted to the contest of *The Highest Performance Imaging-Based Diagnostic Classification Algorithm for ADHD*.

The results of this chapter are produced using the data made publicly by New York University. Preprocessing was performed with scripts[3] developed by Cameron Craddock using AFNI and FSL. The following steps were performed to preprocess the data[3]. The first 4 EPI volumes were removed and the timing for each slice was corrected. The dataset was deobliqued and reoriented to RPI orientation. All EPI volumes were corrected such that the fifth volume was set as the first image of the time series. A mask was created to exclude non-brain data and an 'average brain' image was computed using the average of the volumes. The mean EPI image was co-registered to the corresponding high-resolution anatomic image. The fMRI data and mean image were written into template space at 4 mm x 4 mm x 4 mm resolution. The White Matter and Cerebral Spinal Fluid masks were matched to EPI resolution and the time-courses were extracted. Finally, the slices were band-pass filtered ($0.009 < f < 0.08$ Hz) to exclude frequencies that are not part of the resting state functional connectivity and blurred using a 6mm Full Width-Half Maximum Gaussian filter.

7.2 Functional and Anatomical Parcellations

After preprocessing each of the fMRI scans, the time courses can be extracted from filtered resting state data using different functional and anatomical atlases. Examples of anatomical atlases are the automated anatomical labeling (AAL), Eickhoff-Zilles (EZ), Talairach and Tournoux (TT), and Harvard-Oxford (HO). To create a functional parcellation the time series are typically parcellated into k clusters using the *ncut* cluster algorithm. In this chapter we use the AAL template, a parcellation of an anatomical brain [Tzourio-Mazoyer et al., 2002]. This template allows us to create a mapping from the voxel-space to 116 anatomical defined regions using the Statistical Parametric Mapping Package.

Using this atlas we obtain the time courses for each of the individuals in the dataset. The time courses consist of a $time \times ROI$ matrix, or an $m \times n$ matrix. The number of ROI's (n) depends on the parcellation scheme and the number of time steps (m) depends on the site where the scan was recorded.

The values of an $m \times n$ time series matrix are the MRI signals measured at each voxel location at each sampling epoch. These are the BOLD values and are all relative to each other. This makes it difficult to compare between scans, because the values have no physiological meaning and can vary substantially between subject or scanner. Even the room temperature or the number of times the MRI scanner has been used that day can be of influence to the measured values. We solved this problem by normalizing each of the signals with respect to the variance, hence we compute the z-score for each of the

[3]Further information can be found at `http://www.nitrc.org/plugins/mwiki/index.php/neurobureau:AthenaPipeline`

ROI's:

$$z = \frac{x - \mu}{\sigma}. \tag{7.1}$$

Because our goal is to extract a connectivity matrix between regions in the brain, we have to know which of the areas are connected to each other. In other words, we want to know which areas show related behavior over time and hence are likely influenced by each other. We can do this by computing the correlation matrix $\mathbf{P}$ where each entry $\rho_{i,j}$ shows the correlation between region i and region j.

The values in the correlation matrix are in the range $[-1, 1]$ and indicate the degree of relationship between the functional or anatomical regions i and j. We can now make two observations. First, the correlations on the diagonal are always 1; $\rho_{i,i} = 1$. Second, we can treat negative correlations in the same way as if the correlation was zero, because both imply that there is no positive influence between the regions. This gives some suggestions on how to extract a connectivity graph from this matrix: it can be done by choosing a threshold that determines the minimal required correlation we need to conclude that there is a strong enough connection between those regions. Hence, we set each entry of the matrix as follows:

$$\rho_{i,j} = \begin{cases} 1 & \text{when } \rho_{i,j} \geq \theta, \\ 0 & \text{when } \rho_{i,j} < \theta. \end{cases} \tag{7.2}$$

This process is shown in the example below.

Example: From correlation matrix to connection matrix

The following, partial, time series data is obtained from a participant at the NYU site. The participant is an 11.8 year old, typically developing, male. In this example the fMRI data was parcellated using the AAL template into 116 anatomical regions and for each region we have 172 measurements.

$$T = \begin{pmatrix} -0.1799 & -0.0281 & -0.0742 & 0.0802 & \cdots & t_{1,116} \\ -0.1709 & 0.0768 & 0.0277 & 0.0279 & \cdots & t_{2,116} \\ -0.0096 & 0.0051 & 0.0712 & 0.1386 & \cdots & t_{3,116} \\ 0.0418 & 0.0984 & -0.0015 & 0.0936 & \cdots & t_{4,116} \\ \vdots & & & & \ddots & \vdots \\ t_{172,1} & \cdots & & & \cdots & t_{172,116} \end{pmatrix}. \tag{7.3}$$

The next step is to normalize each of the columns using the z-score. Then we compute the column wise correlations between each of the anatomical regions. This results in a correlation matrix $\mathbf{P}$ of 116×116 correlations.

$$P = \begin{pmatrix} 1.0000 & 0.4690 & -0.2959 & -0.0096 & \cdots & \rho_{1,116} \\ 0.4690 & 1.0000 & -0.1745 & -0.0342 & \cdots & \rho_{2,116} \\ -0.2959 & -0.1745 & 1.0000 & 0.0634 & \cdots & \rho_{3,116} \\ -0.0096 & -0.0342 & 0.0634 & 1.0000 & \cdots & \rho_{4,116} \\ \vdots & & & & \ddots & \vdots \\ \rho_{116,1} & \cdots & & & \cdots & \rho_{116,116} \end{pmatrix}. \tag{7.4}$$

Then the matrix is thresholded according to a threshold θ to create a binarized verion of correlations. Let the threshold θ be 0.2 to obtain the following matrix $\mathbf{P_b}$.

$$P_b = \begin{pmatrix} 1 & 1 & 0 & 0 & \cdots & \rho_{1,116} \\ 1 & 1 & 0 & 0 & \cdots & \rho_{2,116} \\ 0 & 0 & 1 & 0 & \cdots & \rho_{3,116} \\ 0 & 0 & 0 & 1 & \cdots & \rho_{4,116} \\ \vdots & & & & \ddots & \vdots \\ \rho_{116,1} & \cdots & & & \cdots & \rho_{116,116} \end{pmatrix}. \tag{7.5}$$

When this threshold is selected, only regions one and two are connected. Also, every region is connected to itself, because the correlation of a region with itself is always 1. Because it is inconvenient to have self-loops in a network and they do not add information to the graph, it is safe to remove them and we obtain the final graph.

$$P_b = \begin{pmatrix} 0 & 1 & 0 & 0 & \cdots & \rho_{1,116} \\ 1 & 0 & 0 & 0 & \cdots & \rho_{2,116} \\ 0 & 0 & 0 & 0 & \cdots & \rho_{3,116} \\ 0 & 0 & 0 & 0 & \cdots & \rho_{4,116} \\ \vdots & & & & \ddots & \vdots \\ \rho_{116,1} & \cdots & & & \cdots & \rho_{116,116} \end{pmatrix}. \tag{7.6}$$

Important and interesting questions that arise are *how do we choose the 'best' threshold and what is the influence of the threshold*? These questions are the topic of discussion of the next section.

7.3 Influence of the threshold

In this section we have a look at the process of thresholding the correlation matrix. In section 7.3.1 the implications on the graph properties will be discussed, followed by the changes in small-world properties in section 7.3.2. Besides the discussion of general properties we they are also compared when the subjects are divided in groups such as TD/ADHD, Male/Female, Male TD/ADHD and Female TD/ADHD.

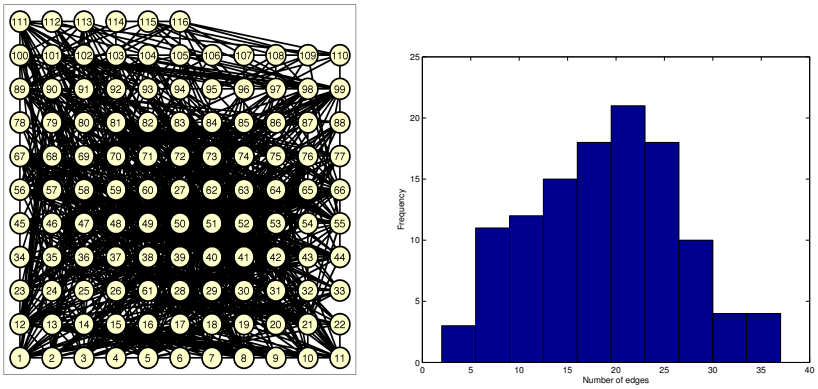
7.3.1 On network properties

Now that the process of extracting a network from fMRI time series data is explained we can inspect the properties of such a network. The chosen threshold is quite important because it determines how strong the network is connected. Intuitively, we can assume that the connectivity of the network is inversely proportional to the threshold. Inasmuch as setting the threshold higher will only allow the regions with a higher correlation to be connected and hence less connection will be in the final model.

Let's have a look at the final graph described in the example of the previous section. The graph is shown in Fig. 7.1a. This also is a good example why it is more convenient

to look at numerical properties of graphs than just to rely on visual inspection because even with such a, relatively, low number of nodes and edges, it looks very cluttered.

There are several properties of a graph we can analyse by simple inspection or computation. Since the graph is parcellated using the AAL anatomical template, we have 116 regions, or nodes, and using $\theta = 0.2$ yields a network of 2230 connections (the network is only for 17% saturated). The least connected region is only connected to two other areas, whereas the maximum is 37. The average number of edges per node is 19.22. A histogram of the edge distribution is shown in Fig. 7.1b and follows a normal distribution.



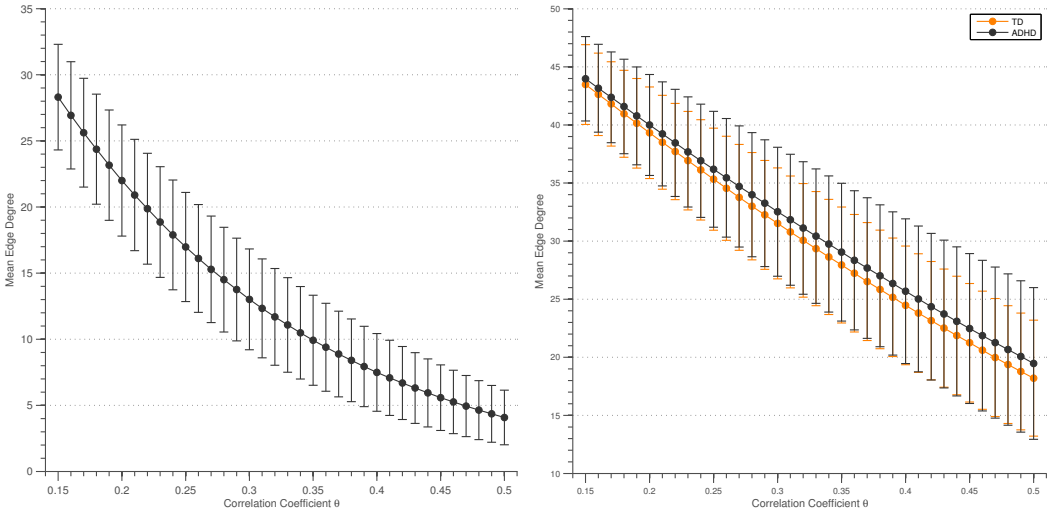
(a) Extracted Network Graph (b) Edge Distribution Graph

Figure 7.1: Extracted Network

Even though the properties of one participant are interesting, the changes in parameters measured over all participants could tell us much more. We start by preprocessing all 202 participants, irrespective of gender, from the NYU site to get the correlation matrices. The plot is shown in Fig. 7.2a, and, not surprisingly, the mean edge degree decreases monotonically when the correlation threshold θ increases.

One may wonder whether there are differences in the mean edge degree between male or female participants, male participants with ADHD or males who are typically developing. The NYU dataset consists of 73 female and 129 male participants. For better comparison only the first 73 male participants are selected to create two balanced groups. The two graphs are shown in Fig. 7.2b. It shows that for a low correlation threshold the mean edge degrees are nearly identical, but as the threshold becomes higher, the means start to deviate. Even though they are different, statistical tests allow to conclude whether the data between two groups are significantly different. It turned out that the apparent differences between mean edge degree between males and females is not significantly different when examined with a t-test and ANOVA, for any of the correlation values. For each case, the p-value is higher than 0.05 (the minimum was 0.18), and hence it is not possible to conclude whether they are significantly different.

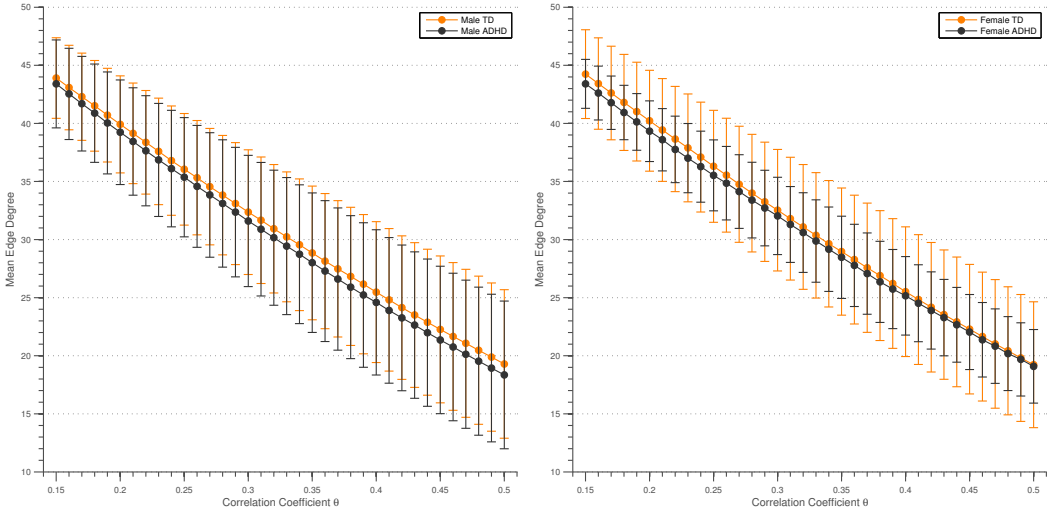
In a similar way, balanced groups of typically developing and ADHD males and females can be created and compared. Comparison of the mean edge degree of 45 males in both groups is shown in Fig. 7.3a. The comparison of 25 females in each group is



(a) All NYU Participants (b) Male - Female Comparison

Figure 7.2: Mean Edge Degree

shown in Fig. 7.3b. Not surprisingly, the differences between male TD and ADHD are not significant. Even though the means of the female groups are apart, statistical testing does not give support for rejection of the null hypothesis.



(a) Male TD / ADHD (b) Female TD / ADHD

Figure 7.3: Mean Edge Degree between male and female populations

These experiments are performed for each of the scan sites and all anatomical and

functional parcellation atlases. Unfortunately, none of the statistical tests have shown a significant difference between the mean edge degree. However, we are still looking for a property that is significantly different between typically developing children and children with ADHD that can give more insight in their brain structure. This would allow us to explain such a disease based on the properties of the brain. In the subsequent section similar comparisons are made but then based on the small-world properties of the brain.

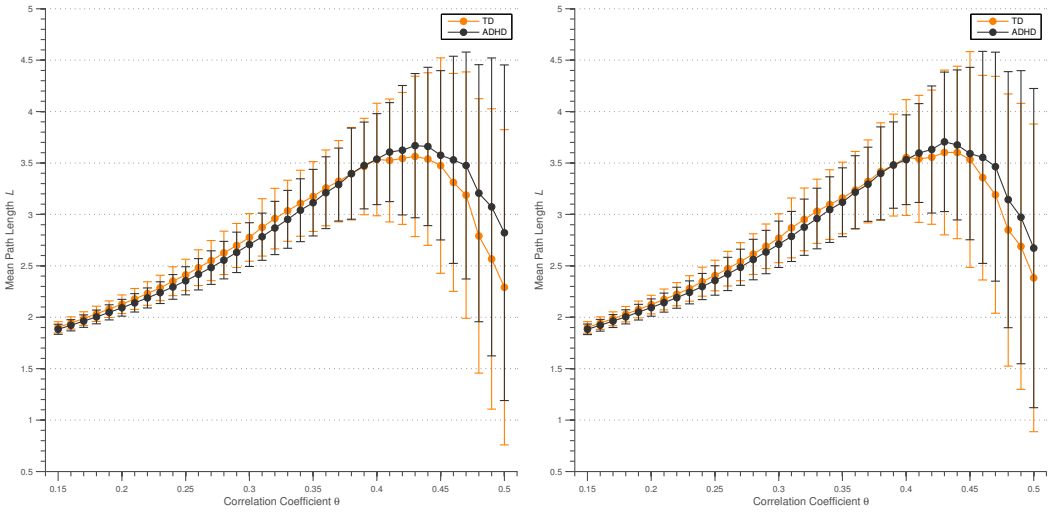
7.3.2 On small world properties

Whereas we looked in the previous section at the edge degree of functional and anatomical parcellations of the brain, we will now turn to small-world properties calculated in the extracted graphs. In section 2.2, we discussed several small-world metrics of which two are the most important and characterizing for small-world networks. These are the characteristic path length and the clustering coefficient. We will first discuss the characteristic path length followed by the clustering coefficient. In this section we, again, use the NYU dataset but the results obtained hold for other sites.

Characteristic Path Length

Let's ponder for a moment what we know about the characteristic path length and the brain to try to predict the outcome based on the observations. First, the CPL gives a measurement of distance in the brain computed as the number of steps it takes from each node, i.e. functional or anatomical region, to every other node. Second, in theory, when computing the path length between disconnected edges, the distance is infinity. Then, as the characteristic path length is defined as the average of each nodes to every other node, the mean path length becomes infinity. This is inconvenient so we define each infinite path as zero and have to take into account that for disconnected graphs the CPL is prone to be lower. Third, because we found differences in edge distribution, albeit not significant, we can reasonably expect this to influence the CPL and therefore might have a physiological influence on the individual. Now, when taking these remarks into account, we can expect the CPL to be short when the correlation threshold is low and it will be longer as the threshold becomes higher because the graph becomes more disconnected. However, setting the path length to zero when there is no real path between two regions might introduce a peculiarity when the graph becomes very disconnected. Inasmuch as there are many zeros in the mean path length, the average path length might not increase anymore or could even decrease.

In Fig. 7.4 the between group comparisons are shown. On the left, Fig. 7.4a, shows the comparison of path length between individuals with TD and ADHD, irrespective of their gender. The trend is somewhat like we predicted: the means are very close together with a low standard deviation when θ is low, both group means tend to be higher as θ increases until the network starts to fall apart. This is evident when the group means are decreasing and the standard deviations are becoming very large. The first column of Table A.1, listed in Appendix A, shows the p-values for each of the paired t-tests of the



(a) TD / ADHD (b) Male / Female

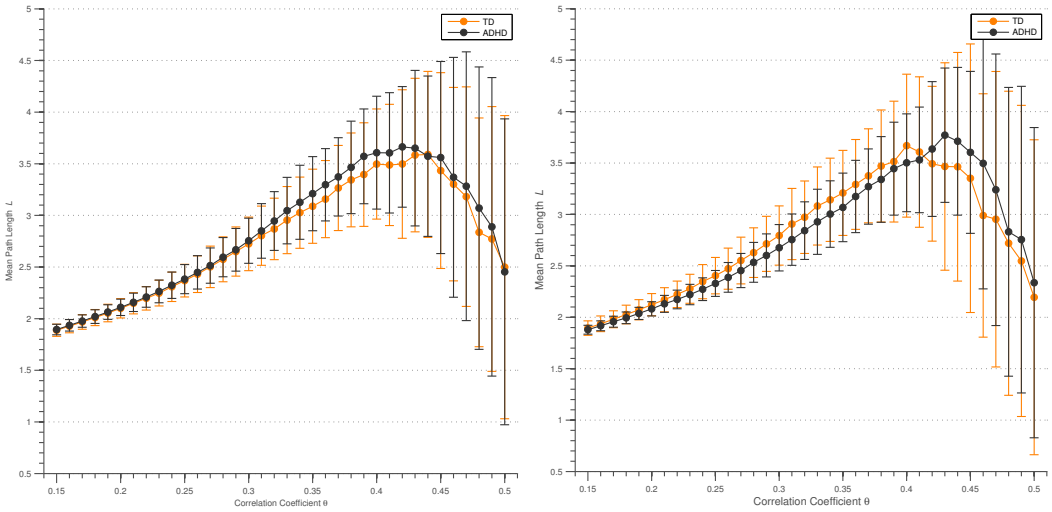
Figure 7.4: Mean CPL between male and female populations

hypothesis that the two groups come from distributions with equal means. For various values of θ the mean characteristic path length is significantly different, i.e., for each entry where the value is lower than 0.05. There are also entries where the p-value is even lower than 0.02, the 98% confidence interval, giving even more support to the hypothesis that they are significantly different. There are many cases, for both a low as well as a high threshold, indicating significance. For the high threshold it is not surprising that they are different for the means are very much apart and both have high standard deviations. However, apparently there is also a significant difference between mean path length between children with ADHD and TD children when the threshold is set to a low value and the graphs are fully connected.

In Fig. 7.4b the comparison between males and females is shown, without taking the diagnostics into account. If the means follow a trend similar to that shown on the left, i.e., means close together with low standard deviations and the graphs breaking apart as θ increases. The second column in A.1 shows the p-values for the paired t-test that the mean path length of males and females are equal. There are only two values of theta giving support that the means are not equal, hence supporting rejection of the null hypothesis. In all other cases the null hypothesis cannot be rejected, it can be the case that the number of samples is not large enough or that there really is not a difference. Also, when considering that there are only two cases that show support in favor of rejection, it might also just be the case that they are different because the sample size is only 73.

The within group comparisons are shown in Fig. 7.5. The comparison between male participants is shown in Fig. 7.5a. The graph also follows the same trend as the graphs previously discussed. However, as can be seen in the third column of Table A.1, none of the differences are significant. This is likely the case due the low number of scans

available for males. Finally, the comparison between females with and without an ADHD diagnosis are shown in Fig. 7.5b. The graph is again similar to the others and like the male comparison, none of the t-tests, the fourth column of Table A.1 show a significance in favor of rejection that the two groups come from distributions with equal means.



(a) Male TD / ADHD

(b) Female TD / ADHD

Figure 7.5: Mean CPL within male and female populations

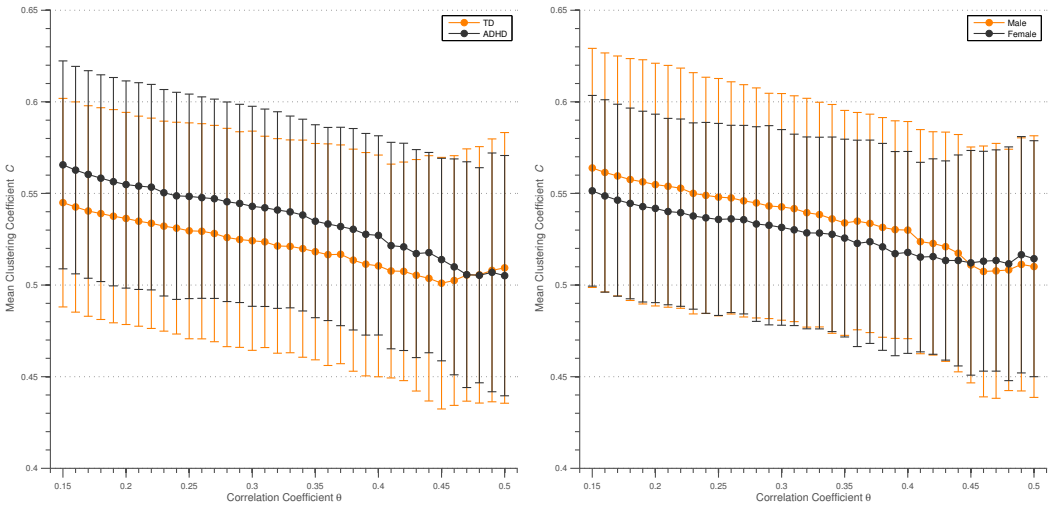
Clustering Coefficient

From Section 2.2 we know that the clustering coefficient is a metric that describes the degree of clustering of the neighbors for a given node. In this case it describes the extend to which adjacent functional regions are connected for a given region. We know that the graph is reasonably well-connected and that it becomes less connected as the correlation coefficient is set higher. Hence we can expect the mean clustering coefficient to start out at a certain point and decrease as the the correlation coefficient increases.

Fig. 7.6a shows the mean cluster coefficient between TD children and children with ADHD. The trend is similar to what we expected: the mean cluster coefficient decreasing as the threshold increases. When the threshold is higher than 0.4, the means are diverging from this trend and tend towards each other. This is probably due to the graph becoming more disconnected. The first column of Table A.2 shows the significant values for the t-test of the hypothesis that the means come from the same distribution. We find the first 23 values all to be significant ($p < 0.05$) and the rest is all not significant ($p \geq 0.05$). We can see that as the means in the graph are starting to diverge from the expected trends the p-values are also becoming very high.

The comparison between male and female participants is shown in Fig. 7.6b. Again, a similar trend is clearly visible, with odd behavior as the threshold becomes high. In

the second column of Table A.2 we find the p-values for the male / female comparison. Even though the means are seemingly as far apart as in the TD / ADHD comparison, none of the t-tests show a significant support for rejection of the hypothesis that they are statistically different.



(a) TD / ADHD (b) Male / Female

Figure 7.6: Mean cluster coefficient between groups

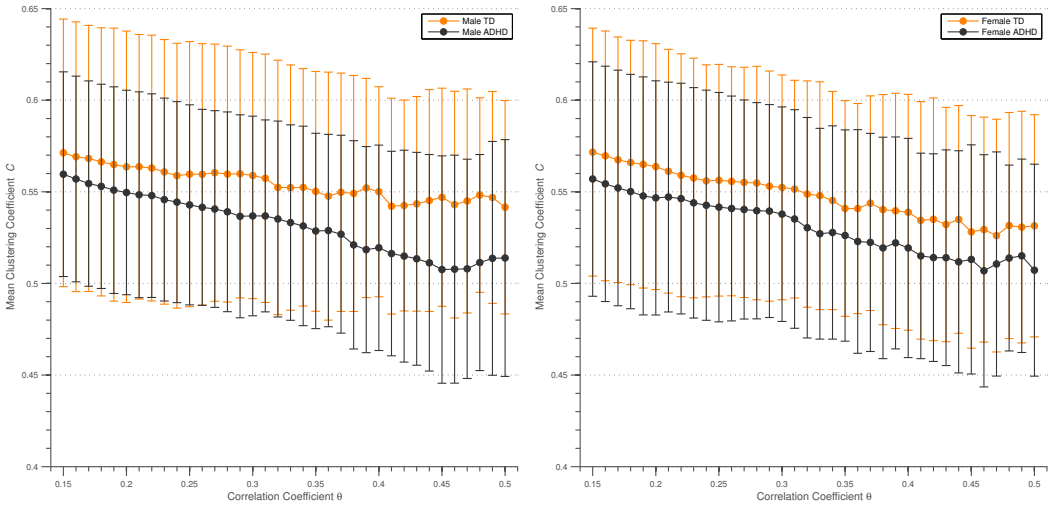
The comparison between males with and without ADHD is shown in Fig. 7.7a. We see that the means are starting out closer together and start to diverge whenever the threshold is higher. In the third column of Table A.2 we see that none of the p-values are significant. This time it could be due to the small number of samples available.

Finally, in Fig. 7.7b we can see the within group comparison for females. This time the means follow a more similar trend, however, although none of the p-values are significantly different. This is not surprising because the sample size is even smaller than that of the males.

An overview is shown in Fig. 7.8. Here the differences between means within the various populations are clearly visible. Unfortunately, none of the means are significantly different, possibly due to the small population size.

7.4 Physiological Analysis of the Cluster Coefficient

In the previous section we found an indication that the mean cluster coefficient is a might be a good metric to differentiate between children with ADHD and normal developing children. In this section we are going to explore the physiological implications of this metric. Even though no statistical significant differences were found within the male participants or the female participants, we will still perform an analysis for each region within those groups to see if we find anything interesting.



(a) Male TD / ADHD (b) Female TD / ADHD

Figure 7.7: Mean cluster coefficient within male and female populations

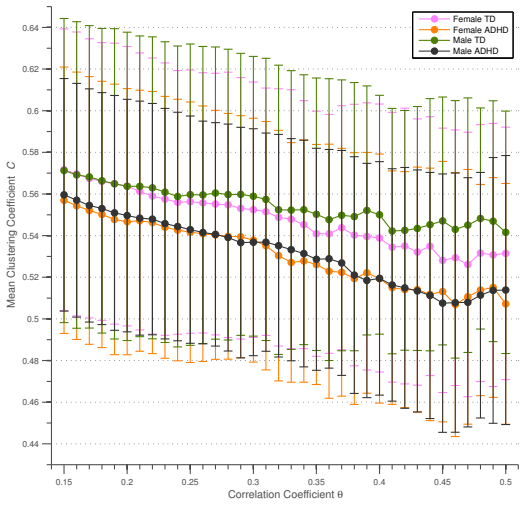


Figure 7.8: Mean cluster coefficient within populations

The mean cluster coefficients turned out to be statistically significantly different for a large number of correlation thresholds. Fig. 7.9 shows a plot of the p-values for the mean cluster coefficient comparison for TD/ADHD children. We can see that for the first 22 values of the correlation threshold, the means are statistically significantly different between groups. For each of the subsequent values, the results are not significant. This could either be the result of the graphs are becoming too disconnected or having too few samples.

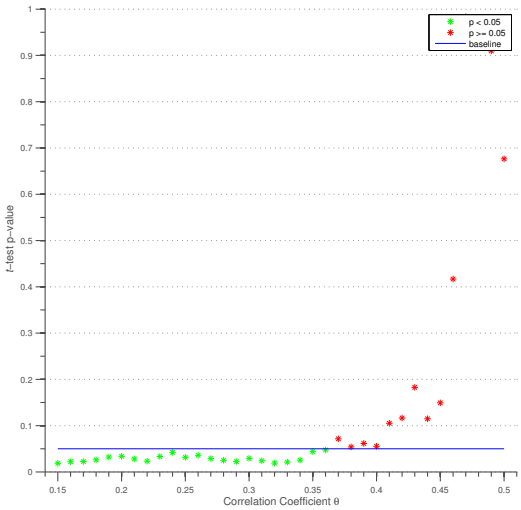


Figure 7.9: T-test p-values for mean cluster coefficient comparison of TD/ADHD

Now that we know that the means of the cluster coefficients are different, it is very interesting to investigate which of the brain regions are causing this. We can perform a t-test for each of the 116 regions to inspect if any of the regions are significantly different or whether only the overall means are different. Whereas we used a 95% confidence level when comparing the means, we will use a 99% confidence level when rejecting the null hypothesis that the two groups are coming from distributions with the same means. Also, instead of using all the correlation threshold values, we will restrict ourselves to only the first 22 for which we rejected the null hypothesis. After performing t-tests for each of the regions and for each of the 22 correlation values we find the 23 different regions which occur at least two times. We chose the minimum frequency of two to leave the regions out which occur only once, because this might be caused by the sample size. Table 7.1 shows the frequency of each region which has a significantly different cluster coefficient. The next thing we have to do is to find the corresponding regions in the AAL atlas and determine for each of the regions whether they are regions that are known to 'cause' ADHD. We can look up each of the labels in the second column of Table 7.1 in the corresponding paper which describes the AAL template by Tzourio-Mazoyer et al. [2002]. Each of the labels has a matching corresponding anatomical description, and the -D and -G at the end of labels indicate the right and left hemisphere, respectively. Next we will discuss each of the regions in order of appearance in Table 7.1. The first 13 regions are located in the cerebrum, whereas the last 10 regions are in the cerebellum.

7.4.1 Anatomical Regions

First, the superior frontal gyrus, dorsolateral (**F1G** and **F1D**) are part of the lateral surface of the frontal lobe. Cognitive control, such as the ability to suppress thoughts and actions is known to be different in children with ADHD compared to typically developing chil-

TABLE 7.1: Frequency table of significant regions

AAL Template Region	Label	Frequency
3	F1G	5
4	F1D	5
7	F2G	6
19	SMAG	2
21	COBG	3
22	COBD	2
31	CIAG	4
32	CIAD	6
39	PARA_HIPPOG	4
40	PARA_HIPPOD	2
61	P2G	4
76	PALLD	3
89	T3G	3
93	CER	7
97	CER4_5G	4
101	CER7BG	6
102	CER7BD	2
103	CER8G	19
104	CER8D	2
105	CER9G	15
106	CER9D	10
107	CER10G	3
114	VER8	18

dren. Durston et al. [Durston et al., 2003] examined the neural basis of these differences. They found differences in the network of regions between these groups, more specifically the activations in frontostriatal regions are different. ADHD children rely on a more diffuse network of regions in specific trials, including more the dorsolateral frontal regions.

The middle frontal gyrus (**F2G**) is also located in the lateral surface of the frontal lobe. In the same study by Durston et al. the MFG has also been shown to respond different for people with ADHD [Durston et al., 2003].

The supplemental motor area (**SMAG**) is part of the medial surface of the frontal lobe. This area is also different in children diagnosed with ADHD according to various studies. Differences in physical behavior in terms of response-time to actions is localized to the supplementary motor area in the studies by Suskauer et al. [Suskauer et al., 2008b]. Suskauer et al. [2008a] show that children with ADHD show greater prefrontal activation than TD children, adding evidence that dysfunction of premotor systems may contribute to impaired response inhibition in children with ADHD.

Differences in the olfactory cortex (**OCG** and **OCD**), located in the orbital surface of the temporal lobe, between TD and ADHD adults is shown in [Schecklmann et al., 2011].

The studies suggest an association of cortical olfactory processing with hyperactivity and impulsitivity in ADHD, i.e., patients with ADHD revealed a diminished activation in olfaction associated brain regions. Significantly ($p < 0.01$) poorer olifaction is also shown in children with ADHD by [Karsz et al., 2008].

The previously discussed anatomical regions are all located in the frontal lobe, for which it is also known to be significantly different for children with ADHD compared to TD children. [Rosack, 2004] shows a reduction of the size of the frontal lobe for children with ADHD.

The anterior cingulate and paracingulate gyri (**CIAG** and **CIAD**) are located in the limbic lobe and play a central role in attentional processing. This is done by modulating stimulus selection (i.e., focussing and managing attention) and mediating the response selection. It is not surprising to find differences in these areas between populations, for it intuitively makes sense that anterior cingulate disfunction might contribute to produce one of the core features of ADHD: inattentive and impulsive behavior. This hypothesis is significantly supported as shown in [Bush et al., 1999] who used fMRI and the Counting Stroop experiment. There are many more papers showing support for this hypothesis, such as [Qiu et al., 2010] focussing both on the structural and functional differences between the brains of children, Seidman et al. [2011] use voxel-based morphology to look at alterations in the gray matter in adults with and without ADHD.

The parahippocampal gyrus (**PARA_HIPPOG** and **PARA_HIPPOD**) is also located in the limbic lobe. The limbic structures are implicated by the origin of ADHD by the presence of mood and cognitive changes or elevations and have been suggested to be involved in the pathophysiology of ADHD. Carmona et al. [2005] show a decrease of about 5.4% in volume of the whole brain in children and adolescents. They localized the differences to various regions, such as the cingulate cortex and the parahippocampal gyrus and the cerebellum. A significant difference is also found using DTI in children [Peterson et al., 2011]. Plessen et al. [2006] also found a significant ($p < 0.002$) difference in the morphology in the (para)hippocampal gyrus for children with ADHD.

The inferior parietal lobule (**P2G**) is part of the parietal lobe and is involved in the perception of facial stimuli and interpretation of sensory information. Abnormalities in this area is common in disorders involving inhibitory control, such as ADHD. Research has shown that this part of the brain is different for children with ADHD. The study of [Rubia et al., 2010] focusses on ADHD and obsessive-compulsive-disorder (OCD) and found increased activation caudate, cingulate and parietal regions, specifically in the inferior parietal lobe. Cao et al. [2008] studied the neurophysiological deficits of TD and ADHD children and found less activation in the frontal (middle and superior gyrus) and inferior parietal lobe, regions in which we also found significant differences.

The lenticular nucleus pallidum (**PALLD**) is located in the sub-cortical gray nuclei. More specifically, it is part of the basal ganglia, which is associated with functions such as voluntary motor control and procedural learning related to behavior and emotions. A large number of brain-imaging evidence supports the role for the basal ganglia in ADHD. Four major studies of the pallidum have shown children with ADHD to have smaller volumes in either the left (**PALLG**) or right part (**PALLD**): Castellanos et al. [2001, 1996, 2002]; Aylward et al. [1996]; Overmeyer et al. [2001]. Seidman et al. [2005] provide a comprehen-

sive literature overview on various regions, including all previous mentioned areas.

Finally, the inferior temporal gyrus (**T3G**) is part of the temporal lobe. Cao et al. [2007] found abnormal neural activity, using the recent regional homogeneity (ReHo) approach, in resting state fMRI in children with ADHD compared to TD children. A similar difference in temporal lobe was found by [Rosack, 2004].

Traditionally the cerebellum (entries 14 to 23 in Table 7.1) has been thought to be involved only in support for motor control. However, clinical and research experiments have shown that the cerebellum is also involved in cognitive and affective processes. Middleton and Strick [2001] have demonstrated that the cerebellum influences the prefrontal cortex via connections in the thalamus which provides a substrate in the pathophysiology of ADHD. Recently, Durston et al. [2004] performed volumetric measurements on the cerebellum and showed a significant volume reduction in the right part of the cerebellum in children with ADHD of 4.9% ($p = 0.026$). Berquin et al. [1998] also performed a volumetric study, solely with male participants, and found a similar significant decrease in volume. Hill et al. [2003] found, amongst others, a significantly smaller total brain, superior prefrontal (**F1G**, **F1D**) and cerebellar lobules.

Conclusions

We found significant differences in clustering coefficient for 14 different anatomical regions, counting the cerebellum as one region. Many of the regions are directly associated with the characteristics of ADHD, because they control the important roles such as the focus of attention, behavior or emotional control. Each of the regions we found is supported by (statistical) findings in other studies based on children with and without ADHD that it influences and characterizes the disease. Based on all this support we can conclude that the cluster coefficient is a remarkable metric for localizing the important areas that are important in diagnosis of ADHD.

Male Group Analysis

We can do a similar analysis for the regions between typical developing male participants and participants with ADHD. Each region was tested between groups using the t-test. The regions which were found statistically significantly different ($p < 0.01$) are shown in the Table 7.2 below. The right column shows the number of times the test found a certain region different between groups.

We can immediately identify several regions that match with the anatomical regions found in the TD/ADHD comparison as shown in Table 7.1. Namely, the superior frontal gyrus (**F1D**) and the olfactory cortex (**COBG**) in the temporal frontal lobe, the inferior temporal gyrus (**T3G**) in the temporal lobe and the lenticular nucleus, pallidum in the sub cortical grey nuclei (**PALD**). Also we see that several areas in the cerebellum are matching (**CERCRU2G**, **CER7BG**, **CER8G**, **CER9G**, **CER9D**, **VER8**).

In the lateral surface we also found differences in the middle frontal gyrus and the opercular part of the inferior frontal gyrus (**F2G** and **F3OPD**, respectively). A study by Sasayama et al. [2010] showed some results that the middle frontal gyrus was different

TABLE 7.2: Frequency table of significant regions in males

AAL Template Region	Label	Frequency
4	F1D	6
7	F2G	14
10	F2OD	3
12	F3OPD	3
21	COBG	4
33	CINMG	7
37	HIPPOG	2
42	AMYGDD	2
76	PALLD	2
82	T1D	3
89	T3G	3
93	CERCRU2G	18
101	CER7BG	5
103	CER8G	15
105	CER9G	4
106	CER9D	8
114	VER8	2
116	VER10	4

in amplitude of low-frequency fluctuation. However, they only found significant results for the right part of the **F2**, whereas we found differences in the left part. In another studies, Yang et al. [2011] found significantly smaller volumes of gray matter in children with ADHD in regions such as the bilateral amygdala and the left middle frontal gyrus. Grey matter reductions and differences in cortical development in the inferior frontal gyrus is more well-known and found in several studies, such as [Shaw et al., 2011; Batty et al., 2010]. Depue et al. [2010] are specifically focussed on reductions in the inferior frontal gyrus in the right hemisphere.

In the orbital surface we found significant differences in both the orbital part of the middle and inferior frontal gyrus (**F2OD** and **F3OPD**, respectively). Unfortunately, there are no papers (yet) that support this finding. Hence it might very well be the case that this finding is incorrect, due to the small sample size.

The temporal lobe is known to be different, volumetrically, in children with ADHD [Rosack, 2004]. At the moment there is only one recent paper [Cao et al., 2009] that supports the differences which were found in the superior temporal gyrus (**T1D**).

The hippocampus (**HIP**) in the limbic lobe is also known to be different for children or adults with ADHD. This finding is supported by several studies such as [Plessen et al., 2006; Verkhlyutov et al., 2010; Amico et al., 2011]. The study of [Verkhlyutov et al., 2010] found differences in the volume of cerebral ventricles. Inasmuch as most studies are focussed on children, Amico et al. [2011] confirmed abnormalities in various areas like the prefrontal cortex, the cingulate cortex, the hippocampus and the amygdala in adults.

Finally, the amygdala (**AMYG**) is located in the sub cortical gray nuclei and is known to play a part in ADHD. This is shown by several studies such as [Amico et al., 2011; Sasayama et al., 2010; Perlov et al., 2008]. Alas, all of the mentioned studies are performed on adults and this makes it more difficult to confirm that our findings are correct.

Conclusions

We found several regions when doing the within group analysis for males that correspond to regions found in the between group analysis. However, there are also regions that we did not find before, but for which we could find support in the literature. We also found several regions which were only weakly supported, or not supported at all, making the findings not as strong as for the TD/ADHD comparison. This is likely because the total sample size was less than half TD/ADHD comparison, making it more vulnerable to more variance in the data which can lead to wrong conclusions and findings (such as regions with no support). Results might be better supported when using a larger sample size.

Female Group Analysis

Table 7.3 lists an overview of all regions which we found significantly different ($p < 0.01$) for females with and without ADHD. Because we only have a small number of samples ($N = 50$), it is much harder to find regions that are different, which can be concluded from the low number of regions found. There are a few regions in agreement with our previous findings. The pallidum (**PAL**) found in both the TD/ADHD comparison as well as the male TD/ADHD comparison. The opercular part of the inferior frontal gyrus (**F3OPD**) was also found in the male comparison, there are, however, no papers yet published with findings of the influence of this part on ADHD.

TABLE 7.3: Frequency table of significant regions in females

AAL Template Region	Label	Frequency
12	F3OPD	3
16	F3OD	3
26	FMOD	2
58	PAD	5
59	P1G	10
75	PALLG	4
76	PALLD	3
87	T2AG	8
113	VER7	3

There is no literature available on our findings in differences in the orbital part of the inferior frontal gyrus (**F3OPD**), the medial orbital part of the superior frontal gyrus (**FMOD**) and the middle temporal pole (**T2AG**). We did find some weak support for two other regions.

The postcentral gyrus in the central region was also found to be significantly different. There is a study by Dibbets et al. [2010] who observed differences in activity in several regions including the postcentral gyrus. However, the participants in their study were only adults and hence we have to be extra careful with concluding that this is a valid result.

Tamm et al. [2006] collected behavioral and brain activation data on a small group of adult participants while they performed a visual oddball task. The subjects with ADHD showed significantly less activation in the bilateral parietal lobes, including the superior parietal gyrus (**P1**). This may support our finding of differences in the **P1G**. Again, we have to be extra careful because we only found differences in the left hemisphere and our dataset consists of only of young females in contrast to their mixed adult subjects.

Conclusions

While we did find a few regions that are in agreement with both our findings in the TD/ADHD as well as the male TD/ADHD comparison, there was almost no support for our other findings. This makes it hard to conclude that we found strong evidence that the cluster coefficient is a good metric to find the regions that are specifically important for differentiating between TD/ADHD children. It does not come as a surprise because we have only 25 participants per group.

7.4.2 Typical Developing Group Analysis

Now that we performed the within group analysis for males and females, it is also interesting to explore the existence of possible gender differences between the groups. Therefore we created two balanced male and female groups ($n = 25$, per group). Below we will explore the existence of differences in cluster coefficient between typically developing males and females and in the subsequent section (7.4.3) we will look into the gender differences for children with ADHD.

TABLE 7.4: Frequency table of significant regions in TD males and females

AAL Template Region	Label	Frequency
1	FAG	5
2	FAD	2
57	PAG	4
58	PAD	2
86	T2D	3
87	T2AD	3

An overview of the significantly different regions is shown in Table 7.4. Surprisingly, a total of six regions are different between the typically developing males and females. The question is if these observed differences are real differences, i.e. the clustering coefficient of the brains are really different, or are these results depending on the threshold that we choose? If the former is true, when looking at the significant regions for each

threshold separately we would find these regions spread through the table. In case the latter hypothesis is true, we would find the regions not to be randomly distributed in such a table.

Table A.3 shows which regions are significantly different for the thresholds 0.15 up to 0.34. Except for the 4th and 9th entry the table shows that for low thresholds the cluster coefficients are not different for typically developing males and females. As the threshold is set higher, the extracted networks start to become disconnected and more regions are becoming significant. Because the frequency of each of the regions which are significantly different is low, we can assume that the difference arise because the graphs are falling apart differently and not because of real gender differences. Using a larger population would allow us to make stronger conclusions regarding the gender differences in typically developing children.

7.4.3 ADHD Group Analysis

We performed a similar analysis as described in the previous section, but this time we compared the males and females with ADHD. Because last time we were unable to find strong evidence for gender differences, except for a few regions for high thresholds, we can expect the same thing to happen in the analysis for children with ADHD. The Table 7.5 below shows the regions which are different for children with ADHD. Only the orbital part of the superior frontal gyrus was found different. This could again be the case that it is not a real difference, but only occurs because the threshold is set too high or the population is too small.

TABLE 7.5: Frequency table of significant regions in ADHD males and females

AAL Template Region	**Label**	**Frequency**
17	F1OG	5

Table A.4 shows for thresholds 0.15 up to 0.34 the regions that are significantly different. Again we find that when the threshold is low, there are no differences between the cluster coefficient per region. The region that differs only shows up in the last five entries of the table, when the threshold is high. In conclusion, it could be the case that there are differences between males and females with ADHD but that these are not significant due our small sample size, or that there are no real differences and our observations are correct.

7.5 Conclusions

In this chapter we performed a graph analysis and comparison using different metrics. Only for the cluster coefficient we found statistically significant differences for both typical developing children and children with ADHD, for males with and without ADHD as well as for diagnosed females. We traced the differences back to the anatomical regions

and found for every significant region other studies that confirmed differences in these regions for children with ADHD.

CHAPTER 8

ANATOMICAL CLASSIFICATION OF ADHD

In this chapter we will investigate and report the results on the classification of children with ADHD and typically developing children. Whereas many studies are performing voxel-based classification, we are using the cluster coefficient as features. In this section we will first look at the classification of TD / ADHD children in Section 8.1, irrespective of their gender This is followed by within-group classification of males (8.2) and then the results of the TD and ADHD female population (8.3). Concluding remarks are given in Section 8.4.

8.1 TD / ADHD Classification

In this section we will look at the classification of typical developing children and children with ADHD. The dataset that we will construct is based on 186 children, equally balanced with 93 TD children and 93 children who are diagnosed with ADHD. For each individual we compute the cluster coefficient for each of the 116 regions in the AAL template as described in the previous chapter 7. Hence, for each participant we get a vector of 116 values in the range of zero to one. We use these values as features for the classifiers that we are using to classify the children. The dataset is thus represented as a 186×116 matrix, where each row is the vector of 116 cluster coefficients.

Let us start by using the whole dataset and training and testing the classifiers using *10-fold* cross validation. Cross validation is used to estimate the error of the untrained classifier. The dataset is randomly ordered and split in 10 sets of approximately the same size. The classifier is then trained on nine sets and tested on the other one. This is then leaving each of the parts out only once and the classification error is computed as the average of all errors obtained.

For classification we used PRTools. PRTools [Duin et al., 2007] is a toolbox written for

MATLAB which includes many classification algorithms. The dataset with features was converted to the PRTools standard format. The classifiers are listed in Table 8.1

TABLE 8.1: Classification results using the cluster coefficient of all anatomical regions in the AAL template as features

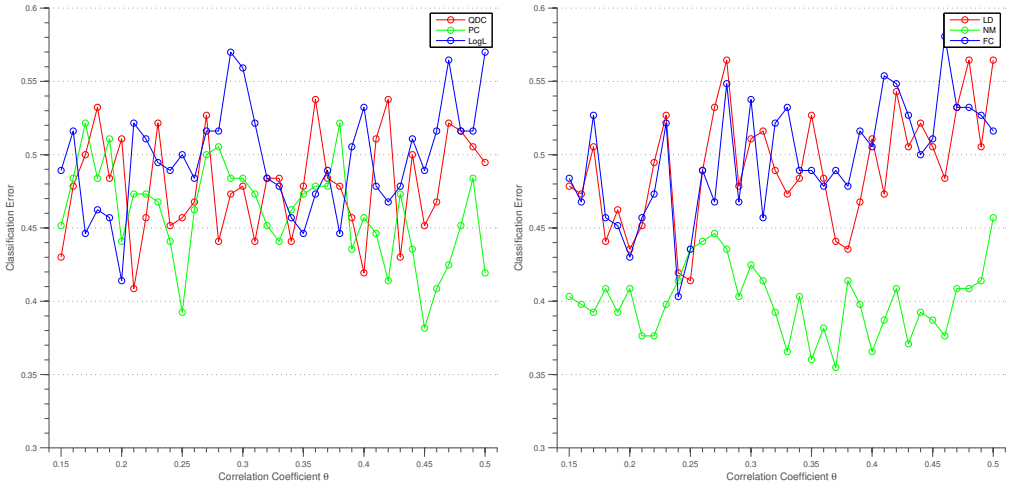
Classifier	Classifier
Quadratic Discriminant Classifier (QDC)	Support Vector Machine (SVM (p=2))
Parzen Classifier (PC)	Support Vector Machine (SVM (p=2))
Logistic Linear Classifier (LogL)	Support Vector Machine (SVM (p=2))
Linear Bayes Normal Classifier (LD)	Support Vector Machine (SVM (r=5))
Nearest Mean Classifier (NM)	Support Vector Machine (SVM (r=10))
Fisher's Linear Classifier (FL)	Support Vector Machine (SVM (r=15))

TABLE 8.2: Classification results using the cluster coefficient of all anatomical regions in the AAL template as features

Classifier	Min Error	Classifier	Min Error
QDC	0.4086	SVM (p=2)	0.3978
PC	0.3817	SVM (p=3)	0.3710
LogL	0.4140	SVM (p=4)	0.3656
LD	0.4140	SVM (r=5)	0.3387
NM	0.3548	SVM (r=10)	0.3978
FL	0.4032	SVM (r=15)	0.3871

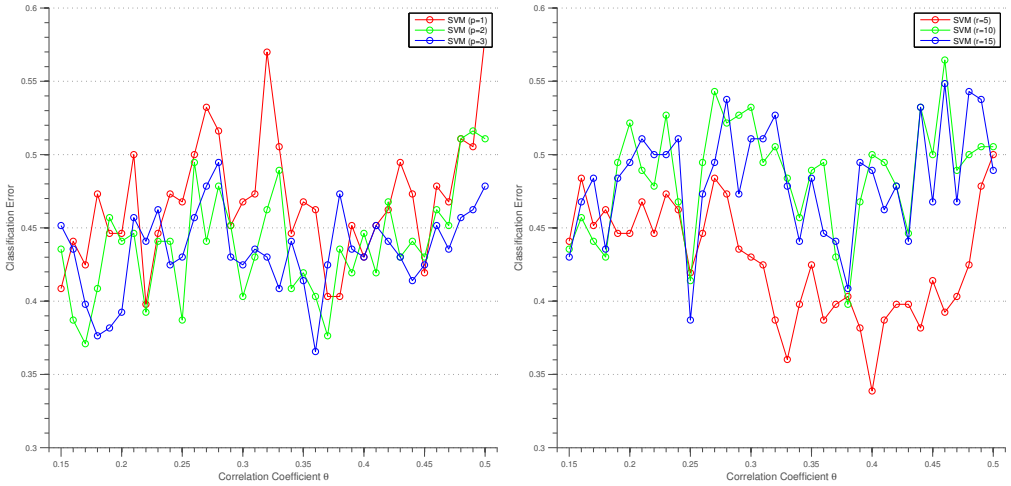
Table 8.2 lists the error for each of the classifiers tested on the entire dataset. With the lowest, or 'best', classification error only being 0.3387, it is obvious that some form of feature reduction is necessary to improve the results, even though classifying about two-thirds of the children correct based on the cluster coefficient is not a bad start. The results are only slightly better than the expected error, for an error of 0.5 would be random guessing. This can be explained because the means of the data are very close together (see Fig. 7.6a). As a matter of fact, they are even closer together than one standard deviation of the data. Hence, it is not surprising that the classifiers are having problems seeing the forest for the trees.

All classifiers for each correlation threshold are shown in Fig. 8.1. The top-left Fig. 8.1a shows the Quadratic, Parzen en Logistic classifiers. We can see that the classification error for each of the classifiers is not dependent on the correlation threshold because none of the classifiers seem to improve or worsen for a varying threshold. Also, none of the classifiers seem to be much better than the others. In the top-right Fig. 8.1b are the plots of the Linear, Nearest Mean and Fisher's Linear classifier. It is clear that the NMC almost consistently outperforms the LD and FC classifier and that also for these classifiers the correlation threshold is not of much influence to the classification error. Support Vector classifiers with a polynomial kernel are shown in the bottom left Fig. 8.1c. Oftentimes the error is around 0.45 and is not determined by the correlation coefficient. Finally, the support vector classifiers with a radial basis kernel are shown in the bottom right in Fig. 8.1d.



(a) Quadratic, Parzen and Logistic Classifiers

(b) Linear, Nearest Mean and Fisher's Linear Classifiers



(c) SVM with polynomial kernels

(d) SVM with radial basis kernels

Figure 8.1: First classification results using all anatomical regions as features

The best performance is shown when using a radial basis with $r = 5$, otherwise the radius is set too large and too many objects are misclassified.

Because the correlation threshold is not of any significant influence, it is set it to 0.2 for the remainder of this chapter unless stated otherwise. There are two reasons for this, first, we found in the previous section that the mean cluster coefficients are still significantly different and, second, the graphs are still connected. The errors obtained for each of the classifiers are shown in Table 8.3 for easier comparison to the errors we get when using some form of feature reduction.

We will next explore three different types of feature reduction: Principal Component

TABLE 8.3: Errors obtained for each of the classifiers ($\theta = 0.2$)

Classifier	Error	Classifier	Error
QDC	0.5000	SVM (p=2)	0.4731
PC	0.4624	SVM (p=3)	0.4301
LogL	0.4462	SVM (p=4)	0.4409
LD	0.4624	SVM (r=5)	0.4785
NM	0.4140	SVM (r=10)	0.4462
FL	0.4839	SVM (r=15)	0.4785

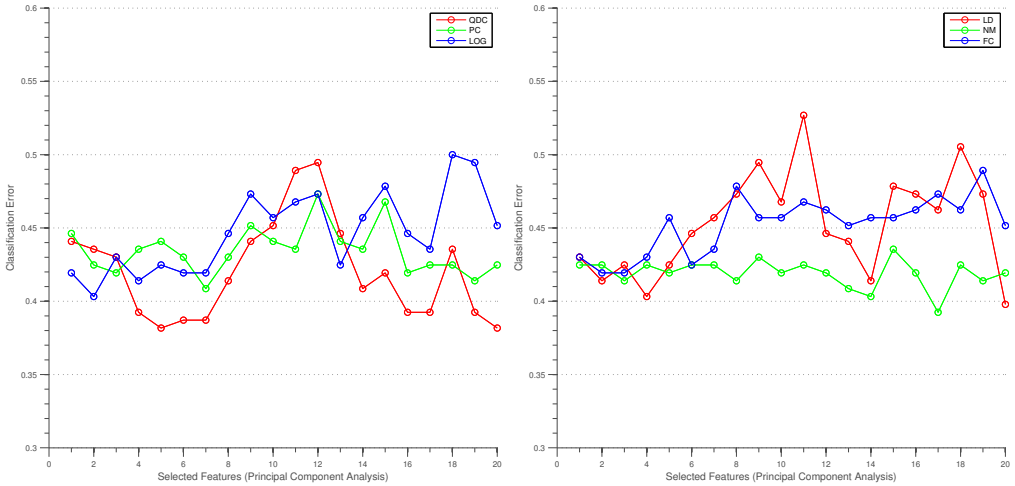
Analysis, Feature Selection and the Fisher Linear Mapping.

8.1.1 Principal Component Analysis

We explored classification and feature reduction using Principal Component Analysis (PCA). PCA is a technique to identify patterns in data, and to transform the data such that the differences between features are maximized while minimizing the similarities. More specifically, it is a mathematical technique that uses orthogonal transformations to convert a dataset with a set of possible correlating features to a set of uncorrelated features. The uncorrelated features are called the principal components. In its simplest form, the data is transformed to a new coordinate system such that the greatest variance of any projection of the data is on the first principal component, the feature with the second-greatest variance is projected on the second coordinate, and so forth. Because the data is transformed in such a way the number of resulting features are always equal or less than in the initial dataset. Good introductions to the mathematics that drive PCA can be found in [Jolliffe, 2002; Smith, 2002].

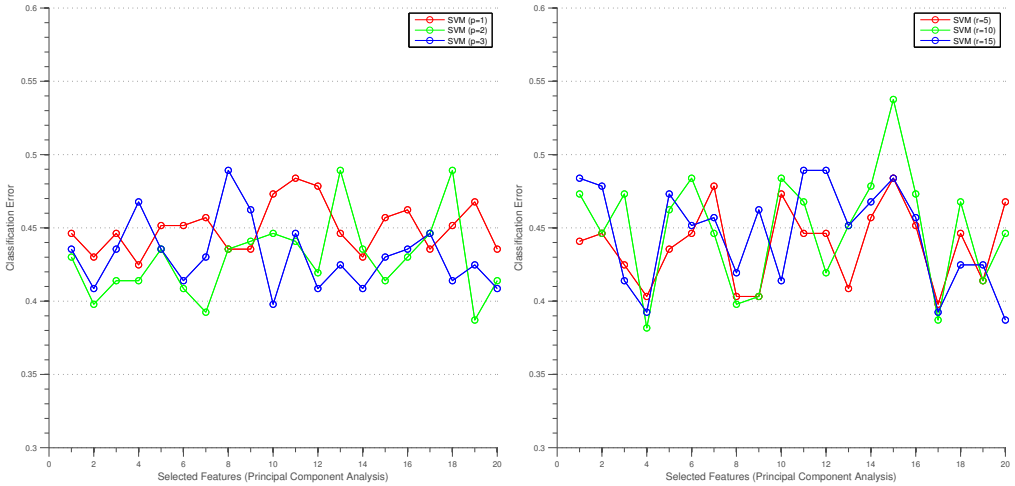
The dataset can be viewed as 186 points in a 116 dimensional space, one point for each participant and one dimension for each region in the atlas. Using PCA we want to find a projection to a lower dimension while retaining a large enough percentage of the original variance in the data that still allows to discriminate between the classes. That is, we want to select the first k out of the n principal components and train a classifier on it, such that we remove $n-k$ features with the lowest variance that are contributing the least to the discrimination. We trained and tested classifiers up to the first 20 principal components, that is for each set of components in $\{\{p_1\},\{p_1,p_2\},\ldots,\{p_1,\ldots,p_{20}\}\}$. The number of 20 was selected because this all allowed us to reduce the 116 dimensional space to 'only' 20 dimensions while still retaining 70% of the variance of the data in the original space.

Fig. 8.2 shows the four panels with classifiers trained and tested on the dataset where feature reduction is done using PCA. In the first Fig. 8.2a we see some improvement when using PCA over the initial classification using the entire dataset. The lowest errors attained are 0.3817, 0.4086, 0.4032 for the QDC, PC and LOG classifiers, respectively. For the QDC and PC there is no visible relationship between the number of principal components and the error. As for the LOG, it seems that the error is going up when more



(a) Quadratic, Parzen and Logistic Classifiers

(b) Linear, Nearest Mean and Fisher's Linear Classifiers



(c) SVM with polynomial kernels

(d) SVM with radial basis kernels

Figure 8.2: Classification using Principal Component Analysis

features are selected. There is definitely an overall improvement since independent of the number of features selected, the error does not go over 0.5.

Fig. 8.2b shows that the performance of the NN classifier is not served by using PCA. This is in contrast to the LD which seems also to be unaffected by the number of features, but the error varies much more. The error of the Fisher's classifier seems to go up as more principal components are selected.

Fig. 8.2c and Fig. 8.2d show the classification errors of the SVM with a polynomial and radial basis kernel, respectively. The selection of more principal components does neither improve or make the classification error much worse.

We can now determine whether feature selection using PCA is advantageous and if so, how much does it improve the classification? Table 8.3 shows the average and 'best' errors obtained when PCA is used, compared to classification using all features in the dataset. First we compare the first and second column: the error for classifying the dataset when the threshold is set at 0.2 against the minimum error for each of the classifiers over all principal components used. Here we find that the best errors using PCA are about 0.05 lower than when classifying using all features. When comparing the former against the mean error we find that also the mean error is in almost all cases better or the performance is about equal, but never much worse.

The lowest error obtained is still very high: 0.3817. This might seem very high but it makes more sense when considering Fig. 8.3. It shows how the data looks when we reduce the dataset to two dimensions. The red and blue dots represents the typical developing children and children with ADHD, respectively. The green line is the NM classifier and the black lines are the support vector machine with a radial basis ($r = 5$). The data points are mixed and very hard to separate, whether linearly or using a polynomial or radial kernel without overtraining the classifier.

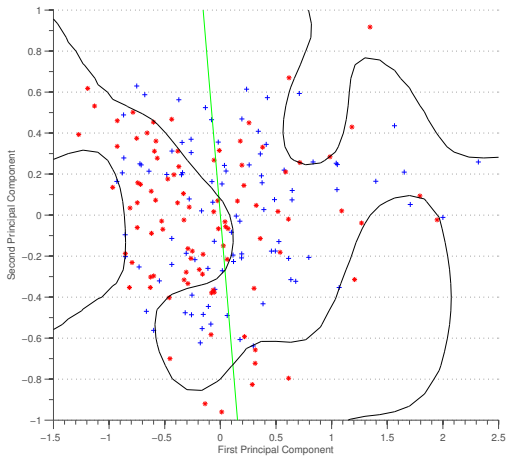


Figure 8.3: Scatterplot of the first and second principal components

8.1.2 Feature Selection

Feature selection, also known as feature reduction or subset selection, is a technique for selecting a subset of features from a dataset. It is widely used in areas such as artificial intelligence, statistics and bio-informatics. By selecting the 'best' k out of n features, it removes the most irrelevant features that do not provide much discriminatory power for classification. As such, feature selection, like PCA, enhances the discriminatory capabilities of a classifier and it improves both the learning process. In contrast to PCA, it gives the user a better understanding in the model, because it allows for easier interpretation of the model.

Various algorithms are available for the evaluation and selection of subsections of the data. There are two major categories in which the algorithms could fall: subset selection and feature ranking. The former performs a (possible exhaustive) search for possible features to be added to the optimal subset. The latter ranks each of the features using a specific metric and remove all the features from the data set that did not receive a score over the baseline.

We inspected the results of feature selection with the *nearest-neighbor* criterion for the following algorithms: best-k, forward-selection and backward-selection. Like PCA, we chose the number of features in the range of one up to 20. We obtained the best results using the individual feature selection and will discuss the results below.

Fig. 8.4a and Fig. 8.4b show a nearly consistent, slightly better than chance, performance for each of the classifiers. Each of the classifiers seem to perform better when they are only using one feature than when more features are added to the subset, albeit there is not much difference.

We can make the same observation for the first two selected features when using support vector machines, see Fig. 8.4c and Fig. 8.4d. The performance of the support vector machines with a first or second order polynomial kernel are similar to the linear or quadratic classifiers as shown in Fig. 8.4a and 8.4b. When a third order polynomial kernel is chosen in combination with more than three features, the classifier performs even worse than assigning each of the children to a class by random chance. Using a radial basis kernel does not improve the performance either, when more than two features are selected.

The best error obtained using a logistic linear classifier was 0.4086 using 13 features. It is again not very impressive, which suggest that even feature selection is not a very good method to reduce the dimensionality of the data. This is confirmed when a scatterplot was made of the first and second best features, shown in Fig. 8.5. It shows the data and two classifiers (green for the nearest mean classifier and black for the support vector machine with a radial basis kernel ($r = 5$)). It is obvious why the classifiers have this much trouble finding a linear or polynomial separation of the data points.

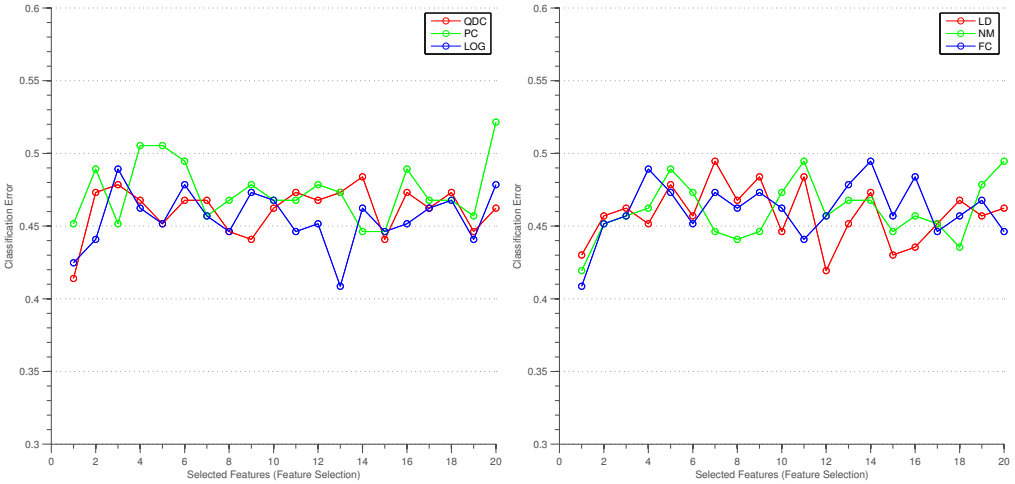
8.1.3 Fisher Linear Mapping

Whereas the Principal Component Analysis is applied to the covariance of the complete data set, the Fisher's Linear Mapping (FLM) is used to optimize the linear class separability. That is, PCA gives the user the opportunity to choose k out of the n principal components (with $k \in [1,\ldots,n]$). The goal of the FLM is to find a mapping, not necessarily orthogonal, for a dataset to a k-dimensional subspace, where k is less than the number of classes. Hence, in this case it maps the data to a 1-dimensional space.

The mapping is found by maximizing the between scatter over the within scatter. Formally defined as:

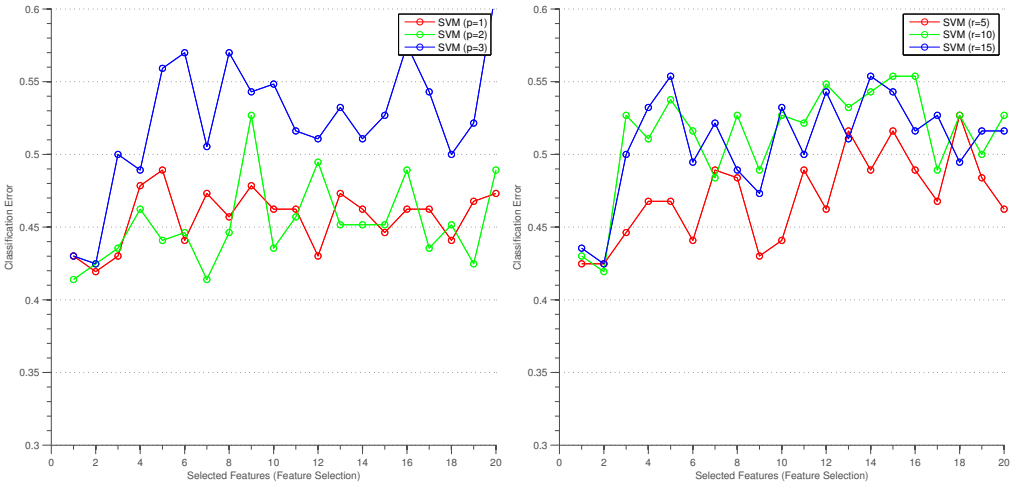
$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}, \tag{8.1}$$

where S_B is the 'between class scatter matrix' and S_W is the 'within class scatter matrix'.



(a) Quadratic, Parzen and Logistic Classifiers

(b) Linear, Nearest Mean and Fisher's Linear Classifiers



(c) SVM with polynomial kernels

(d) SVM with radial basis kernels

Figure 8.4: Classification using Feature Selection

The class scatter matrices are defined as follows:

$$S_B = \sum_c (\mu_c - \bar{x})(\mu_c - \bar{x})^T. \tag{8.2}$$

$$S_W = \sum_c \sum_{i \in c} (x_i - \mu_c)(x_i - \mu_c)^T. \tag{8.3}$$

In two-class problems this corresponds to computing the scatter of one class relative to the scatter of the other class and hence can be simply computed as:

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T.$$

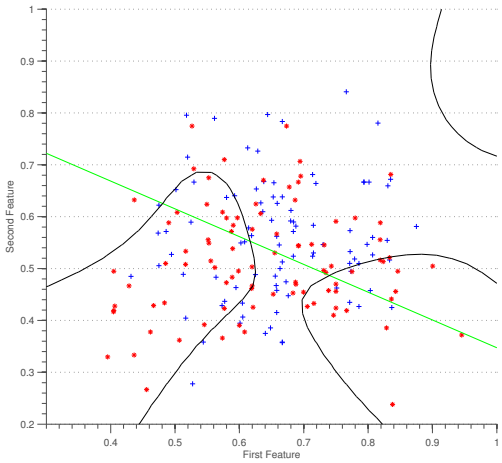


Figure 8.5: The two 'best' features with Feature Selection

So why does this work? The formula defines a good solution as one that separates the class means as far as possible, relative to the sum of the variances of the data in each class. Besides the fact that the object is supervised in comparison to PCA, it is also invariant with respect to scalings. Due to this invariance in scaling, we can choose $\mathbf{w}$ such that we get $\mathbf{w}^T S_W \mathbf{w} = 1$. Therefore, we can transform our problem in the following optimization problem:

$$\begin{aligned} \underset{w}{\text{minimize}} \quad & -\frac{1}{2}\mathbf{w}^T S_B \mathbf{w}, \\ \text{subject to} \quad & \mathbf{w}^T S_W \mathbf{w} = 1. \end{aligned} \tag{8.4}$$

Fig. 8.6 shows the result of mapping the dataset to a single dimension using Fisher's linear discriminant analysis. We can see that the data is very well separated, except for some data points in the middle. The classification error can be expected to improve significantly, but it is obvious that there will still exist an error. Table 8.4 shows the classification error for all classifiers. It is not unexpected that many classifiers attain the same low error, because the aforementioned figure showed that some participants were mapped to the wrong side of the line. The best attained error is as low as 0.0753, which means that a remarkable 93% of the 183 children can be classified correctly whether they are normal or have ADHD based on the cluster coefficient of the different anatomical regions in their brain. Finally, Table 8.4 also shows that, because the data is mapped to a one dimensional space, simple linear classifiers outperform the complex classifiers.

8.1.4 Conclusion

In this section we explored and investigated the classification of children with and without ADHD based on the cluster coefficient computed between the anatomical regions from fMRI data. When all the cluster coefficients are used, the best error obtained was 0.41. Using the unsupervised principal components technique (for one up to 20 principal components) the error was slightly lowered to 0.3817. After that we explored feature

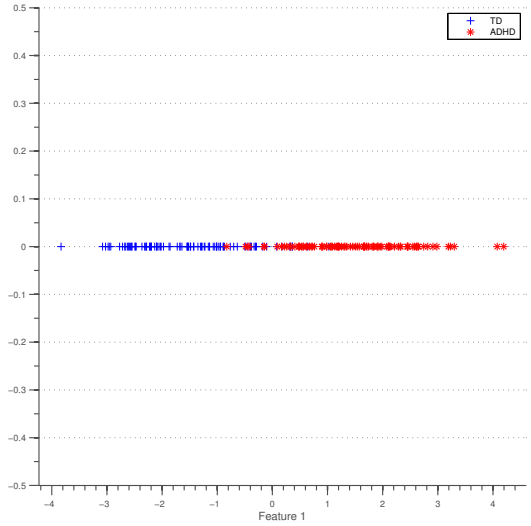


Figure 8.6: The dataset mapped to one dimension

TABLE 8.4: Errors obtained using Fisher's mapping ($\theta = 0.2$)

Classifier	Error	Classifier	Error
QDC	0.0753	SVM (p=2)	0.0753
PC	0.0753	SVM (p=3)	0.0806
LogL	0.0753	SVM (p=4)	0.0914
LD	0.0753	SVM (r=5)	0.0860
NM	0.0753	SVM (r=10)	0.0753
FL	0.0753	SVM (r=15)	0.0753

selection using different algorithms: best k out of n, forward selection and backward selection. The first mentioned yielded the best results, albeit a mere 'best' error of 0.4086. Finally, we used Fisher's Linear Discriminant Analysis to find a supervised, optimal solution to map the 116 dimensions to a one dimensional space where the between class scatter is maximized while the within class scatter is minimized. This yielded a nearly separable mapping and dramatically improved the correct classification rate: about 93% of the children in balanced dataset were diagnosed correctly using the anatomical cluster coefficients as features.

8.2 Male TD / ADHD Classification

Since Fisher's Linear Mapping significantly outperforms the other feature reduction algorithms, Principal Component Analysis and Feature Selection, we only perform classification using FLM in this and the sequent section. Again, we set the correlation threshold to 0.2, to make the results comparable.

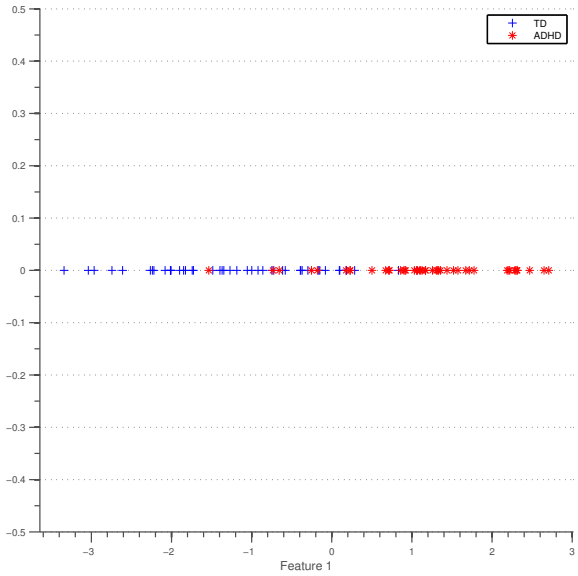


Figure 8.7: The male TD/ADHD dataset mapped to one dimension

Fig. 8.7 shows the male dataset mapped to one dimension using FLM. Visual inspection tells us that the data is slightly more distorted and will be harder to separate. Thus, we can expect the error to be higher. This is confirmed with the results shown in Table 8.5. The best classification results in an error of 0.11, which means that slightly more than 88% of the data was classified correctly. The 'worst' classifier is still correct in 84% of the cases. The higher classification error is likely to be caused by the lower number of subjects in our dataset.

TABLE 8.5: Male TD/ADHD errors obtained using Fisher's mapping ($\theta = 0.2$)

Classifier	**Error**	**Classifier**	**Error**
QDC	0.1333	SVM (p=2)	0.1222
PC	0.1444	SVM (p=3)	0.1111
LogL	0.1222	SVM (p=4)	0.1111
LD	0.1333	SVM (r=5)	0.1333
NM	0.1333	SVM (r=10)	0.1556
FL	0.1333	SVM (r=15)	0.1333

8.3 Female TD / ADHD Classification

For the small balanced dataset consisting of 25 females in each group we also tested the classifiers using FLM. Fig. 8.8 shows even more distortion in the mapping to a one dimensional space. This results in a higher classification error as can be seen in Table 8.6.

The best classifier assigns 78% to the right classes whereas the worst classifier only manages to classify about two-thirds of the samples to the correct class.

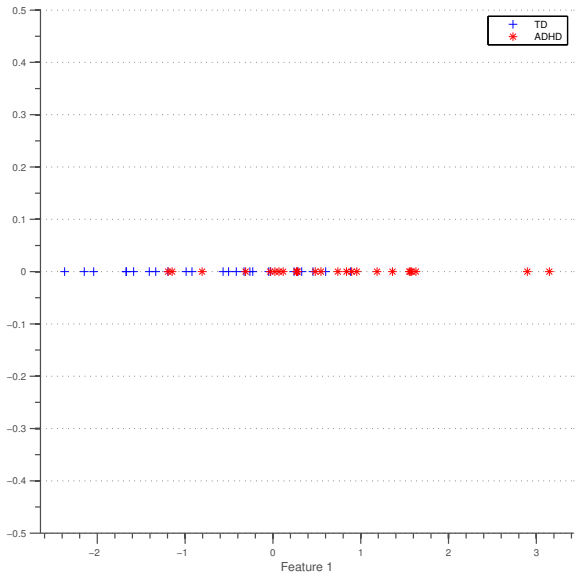


Figure 8.8: The female TD/ADHD dataset mapped to one dimension

TABLE 8.6: Female TD/ADHD errors obtained using Fisher's mapping ($\theta = 0.2$)

Classifier	**Error**	**Classifier**	**Error**
QDC	0.2600	SVM (p=2)	0.2200
PC	0.2600	SVM (p=3)	0.3200
LogL	0.2400	SVM (p=4)	0.3400
LD	0.2400	SVM (r=5)	0.2600
NM	0.2200	SVM (r=10)	0.2800
FL	0.2600	SVM (r=15)	0.2600

8.4 Conclusions

In this chapter we put several classifiers to the test for classifying participants with and without ADHD. As features we used the cluster coefficients for each of the 116 regions as defined in the AAL template. Using all features didn't yield any good performance and it was clear that feature reduction had to be used. We found that both PCA as well as individual feature selection were not the right methods in this case, for they did not improve the performance of the classifier significantly. Fisher's Linear Mapping, on the other hand, did improve the classification error a lot. We found classification errors as low as 7% when classifying between children with and without ADHD. As the dataset was

much smaller when classifying males and females with and without ADHD we saw the error go up to 0.11 and 0.22, respectively.

CHAPTER 9

ADHD CLASSIFICATION USING K-PRMEANS CLUSTERS

This chapter describes the results from classification of ADHD using functional k-prmeans clusters. The disadvantage of functional parcellation, in contrast to anatomical as described in the previous chapter, is that it does not allow for physiological interpretation. We explored the idea of clustering the fMRI scans into k regions using the stable-k-prmeans algorithm, compute the cluster coefficients using a threshold of 0.2 (such that it allows for comparison with the previous described experiments), and then try to classify the data into TD and ADHD children.

We start by clustering the fMRI scan of each participant in $k = \{190, 191, \ldots, 210\}$ regions. We repeated the clustering 100 times for each participant and looked for the 'best' clustering in terms of lowest conductivity inverse proportionally to the frequency of the result. That is, the more frequent a result with a low conductivity is found, the better. We found that clustering the brain into 194 regions yielded, on average, the best result. There is no rationale for performing clustering in for 190 to 210 regions, besides that is about twice as big as the AAL anatomical template. Similarly to the previous chapter we used 'PRTools' for classification with the same classifiers (Table 8.1).

9.1 TD / ADHD Classification

In the previous chapter we discussed Fisher's Linear Mapping. This the dimensionality reduction algorithm that provided the best results over Principal Component Analysis and Feature Selection.

For classification of the typical developing and ADHD populations we constructed a dataset consisting of 194 cluster coefficients. That is, one feature for each of the clusters we obtained using the stable-k-prmeans algorithm.

Table 9.1 shows an overview of the classification errors. The best classifier is a Support Vector Machine with a third order polynomial kernel. This classifier is over 90% accurate. We again see that many classifiers obtain the same error, which is because the tested classifiers are linear and the data is mapped to a one-dimensional space. The decision boundaries of the classifiers are overlapping and fail to classify the same items correctly. When comparing these results to the errors obtained using anatomical parcellation, we see that the order of the errors is similar.

TABLE 9.1: Errors obtained using Fisher's mapping ($\theta = 0.2$) and k-prmeans for the TD and ADHD populations

Classifier	Error	Classifier	Error
QDC	0.1075	SVM (p=2)	0.1075
PC	0.1129	SVM (p=3)	0.0968
LogL	0.1129	SVM (p=4)	0.1075
LD	0.1075	SVM (r=5)	0.1344
NM	0.1075	SVM (r=10)	0.1075
FL	0.1075	SVM (r=15)	0.1075

9.2 Male TD / ADHD Classification

We again split the male subjects in two balanced groups (typically developing and ADHD) of 45 individuals each. We construct a dataset with 194 features for the cluster coefficient of each functional region. The classification results are summarized in Table 9.2. In contrast to the between group classification, the Support Vector Machine with a third order polynomial kernel is now performing the worst on this dataset. The error has gone up by 0.09 on average and from the total of 90 subjects we classify 'only' 82% correct.

TABLE 9.2: Errors obtained using Fisher's mapping ($\theta = 0.2$) and k-prmeans for the male populations

Classifier	Error	Classifier	Error
QDC	0.1889	SVM (p=2)	0.1889
PC	0.1778	SVM (p=3)	0.2111
LogL	0.1889	SVM (p=4)	0.1778
LD	0.1889	SVM (r=5)	0.1889
NM	0.1889	SVM (r=10)	0.1889
FL	0.1889	SVM (r=15)	0.1778

9.3 Female TD / ADHD Classification

The dataset of the within group female population consists of only 25 subjects per group. This is about half of the male within group comparison and can therefore expect the error to go up. Table 9.3 shows a big increase in error. For this dataset, several classifiers are tied for the 'best' error of 0.38, which implies that a mere 62% is classified correctly. The Support Vector Machine with a fourth order polynomial kernel performs even worse than by chance.

TABLE 9.3: Errors obtained using Fisher's mapping ($\theta = 0.2$) and k-prmeans for the female populations

Classifier	**Error**	**Classifier**	**Error**
QDC	0.3800	SVM (p=2)	0.4000
PC	0.4800	SVM (p=3)	0.4800
LogL	0.3800	SVM (p=4)	0.5200
LD	0.3800	SVM (r=5)	0.3800
NM	0.4000	SVM (r=10)	0.3800
FL	0.3800	SVM (r=15)	0.4400

9.4 Conclusions

Apparently, when using a large sample size such for the TD and ADHD populations functional parcellation using k-prmeans shows to be a promising alternative. Even though we have different features as well as a different number it only performs slightly worse than classification using the cluster coefficients derived from anatomical parcellation. Because the error goes up significantly (to the level of performing at chance), it is clear that a large dataset is required for this method to produce good results.

Part IV

Conclusions

CHAPTER 10

CONCLUSIONS AND FUTURE WORK

At the introduction of this thesis we embarked on two problems. First, we noted that recent research has shown a significant increase over the past few years in children diagnosed with ADHD. To date, every diagnosis is still performed per physicians' interpretation of a child's behavior. Second, the class of networks known as small-world networks are occurring in many (natural) phenomena. What makes them so interesting is the short average path length between each set of nodes and the high degree to which nodes in the graph tend to cluster. Cluster algorithms relying on a Euclidean distance frequently fail to produce a good partitioning because of the short path length.

These two, seemingly unrelated, problems were unified under the following research question:
Are there significant differences in small-world properties in a graph representation of the brain between normal children and children diagnosed with attention-deficit / hyperactivity disorder?

In order to provide a more accurate and reliable method to establish whether a child has ADHD, we attacked the problem from two perspectives. We proposed a method to extract a graph from functional Magnetic Resonance Imaging data. Then a graph can be derived using an anatomical parcellation template such as AAL or be based on functional connectivity using a cluster algorithm. We investigated (section 7) the anatomical differences in the small-world properties between a typical developing and ADHD population. For the cluster coefficient (section 7.3.2), we found significant differences in several regions. Each of the regions corresponded to literature which also found differences between populations in those regions (section 7.4). This gave a lot of support to consider the cluster coefficient as a indicative metric for identifying ADHD. To test the hypothesis that we can diagnose whether a subject has ADHD, we tested several classifiers and dimensionality reduction methods (chapter 8). Success rates of about 94% when classifying a total balanced population of 196 children using Fisher's Linear Mapping (section 8.1.3).

The second method is based on functional parcellation of fMRI data. Repeated clustering of each scan was performed and classification using the cluster coefficient resulted in errors of the same order as anatomical parcellation, albeit slightly higher.

Two different approaches to processing fMRI data and extracting the small-world properties both resulted in performance over 90%. A clear indication that there might be intrinsic differences in the neuronal structure for typically developing children and children with ADHD.

10.1 Future Work

Unfortunately, there was only a limited amount of time that could be devoted to this project and there are still many interesting open ideas to explore in the future. This project constitutes of three parts: a cluster algorithm that uses the distance between personalized PageRank vectors, anatomical analysis of the small-world properties of ADHD and disease classification. Each of the ideas improves one piece of the puzzle.

10.1.1 K-prmeans

The k-prmeans has only been tested on a limited number of networks. Networks which are of the small-world class and for which we would expect this algorithm to work. It would be very interesting to put the algorithm to the test on a variety of problems, such as meshes and other esoteric or less common graph models.

10.1.2 Anatomical Analysis

In this thesis we investigated the differences in regions for the anatomical parcellation between children with and without ADHD. The cluster coefficient turned out the be a powerful indication for this particular brain disease. However, there are several other neuronal abnormalities that might show similar differences in one or more of the small-world properties, such as Alzheimer, Parkinson's or dyslexia. A small sample of fMRI scans from one of these diseases would enable preliminary investigation and for a large population it is possible to determine whether there are significant differences between properties which might then be related to literature.

10.1.3 Classification

Classification was only performed between typically developing children and children with ADHD. There are, as discussed in the introduction, three different types of ADHD. With these distinctions in mind, there are two interesting options to explore. First, the physiological implications. It is possible to find differences in small-world properties in regions between children with different types of ADHD and if so, can these differences be related to the literature? Second, the classifiers are based on a nominal scale (a child is classified as either typical developing or diagnosed with ADHD). Interesting would it be to investigate the discriminative qualities of an ordinal classifier.

BIBLIOGRAPHY

Diagnostic and Statistical Manual of Mental Disorders DSM-IV-TR Fourth Edition (Text Revision). Amer Psychiatric Pub, 4th edition, July 2000. ISBN 0890420254.

S. Achard and E. Bullmore. Efficiency and cost of economical brain functional networks. *PLoS Comput Biol*, 3(2):e17, Jan 2007. doi: 10.1371/journal.pcbi.0030017.sg002.

F. Amico, J. Stauber, N. Koutsouleris, and T. Frodl. Anterior cingulate cortex gray matter abnormalities in adults with attention deficit hyperactivity disorder: a voxel-based morphometry study. *Psychiatry Res*, 191:31–35, Jan 2011.

R. Andersen, F. Chung, and K. Lang. Local graph partitioning using pagerank vectors. In *Foundations of Computer Science, 2006. FOCS '06. 47th Annual IEEE Symposium on*, pages 475 –486, oct. 2006. doi: 10.1109/FOCS.2006.44.

D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. *SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, Jan 2007. URL `http://portal.acm.org/ft_gateway.cfm?id=1283494&type=pdf&coll=DL&dl=GUIDE&CFID=10734338&CFTOKEN=12753815`.

No authors listed. Clinical practice guideline: diagnosis and evaluation of the child with attention-deficit/hyperactivity disorder. American Academy of Pediatrics. *Pediatrics*, 105:1158–1170, May 2000.

No authors listed. Increasing prevalence of parent-reported attention-deficit/hyperactivity disorder among children — United States, 2003 and 2007. *MMWR Morb. Mortal. Wkly. Rep.*, 59:1439–1443, Nov 2010.

E. H. Aylward, A. L. Reiss, M. J. Reader, H. S. Singer, J. E. Brown, and M. B. Denckla. Basal ganglia volumes in children with attention-deficit hyperactivity disorder. *J. Child Neurol.*, 11:112–115, Mar 1996.

A. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999. doi: 10.1126/science.286.5439.509.

M. J. Batty, E. B. Liddle, A. Pitiot, R. Toro, M. J. Groom, G. Scerif, M. Liotti, P. F. Liddle, T. Paus, and C. Hollis. Cortical gray matter in attention-deficit/hyperactivity disorder: a structural magnetic resonance imaging study. *J Am Acad Child Adolesc Psychiatry*, 49:229–238, Mar 2010.

P. C. Berquin, J. N. Giedd, L. K. Jacobsen, S. D. Hamburger, A. L. Krain, J. L. Rapoport, and F. X. Castellanos. Cerebellum in attention-deficit hyperactivity disorder: a morphometric MRI study. *Neurology*, 50:1087–1093, Apr 1998.

M. Bianchini, M. Gori, and F. Scarselli. Inside pagerank. *ACM Transactions on Internet Technology (TOIT)*, 5(1):92–128, 2005.

S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang. Complex networks: structure and dynamics. *Physics Reports*, 424:175–308, 2006. doi: 10.1016/j.physrep.2005.10.009.

S. Brin, R. Motwani, and T. Winograd. What can you do with a web in your pocket. *Data Engineering Bulletin*, 21:37–47, 1998.

K. Bryan and T. Leise. The $25,000,000,000 eigenvector. *The linear algebra behind Google*, 2006. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.133.6739&rep=rep1&type=pdf.

G. Bush, J. A. Frazier, S. L. Rauch, L. J. Seidman, P. J. Whalen, M. A. Jenike, B. R. Rosen, and J. Biederman. Anterior cingulate cortex dysfunction in attention-deficit/hyperactivity disorder revealed by fMRI and the Counting Stroop. *Biol. Psychiatry*, 45:1542–1552, Jun 1999.

Q. Cao, Y. Zang, C. Zhu, X. Cao, L. Sun, X. Zhou, and Y. Wang. Alerting deficits in children with attention deficit/hyperactivity disorder: Event-related fmri evidence. *Brain Research*, 1219:159 – 168, 2008. ISSN 0006-8993. doi: DOI:10.1016/j.brainres.2008.04.028. URL http://www.sciencedirect.com/science/article/pii/S0006899308008822.

Q. J. Cao, Y. F. Zang, and Y. F. Wang. [Brain functions in attention deficit hyperactivity disorder combined and inattentive subtypes: A resting-state functional magnetic resonance imaging study]. *Beijing Da Xue Xue Bao*, 39:261–265, Jun 2007.

X. Cao, Q. Cao, X. Long, L. Sun, M. Sui, C. Zhu, X. Zuo, Y. Zang, and Y. Wang. Abnormal resting-state functional connectivity patterns of the putamen in medication-naâĹŽÃŸve children with attention deficit hyperactivity disorder. *Brain Res.*, 1303:195–206, Dec 2009.

S. Carmona, O. Vilarroya, A. Bielsa, V. Tremols, J. C. Soliva, M. Rovira, J. Tomas, C. Raheb, J. D. Gispert, S. Batlle, and A. Bulbena. Global and regional gray matter reductions in ADHD: a voxel-based morphometric study. *Neurosci. Lett.*, 389:88–93, Dec 2005.

F. X. Castellanos, J. N. Giedd, W. L. Marsh, S. D. Hamburger, A. C. Vaituzis, D. P. Dickstein, S. E. Sarfatti, Y. C. Vauss, J. W. Snell, N. Lange, D. Kaysen, A. L. Krain, G. F. Ritchie, J. C. Rajapakse, and J. L. Rapoport. Quantitative brain magnetic resonance imaging in attention-deficit hyperactivity disorder. *Arch. Gen. Psychiatry*, 53:607–616, Jul 1996.

F. X. Castellanos, J. N. Giedd, P. C. Berquin, J. M. Walter, W. Sharp, T. Tran, A. C. Vaituzis, J. D. Blumenthal, J. Nelson, T. M. Bastain, A. Zijdenbos, A. C. Evans, and J. L. Rapoport. Quantitative brain magnetic resonance imaging in girls with attention-deficit/hyperactivity disorder. *Arch. Gen. Psychiatry*, 58:289–295, Mar 2001.

F. X. Castellanos, P. P. Lee, W. Sharp, N. O. Jeffries, D. K. Greenstein, L. S. Clasen, J. D. Blumenthal, R. S. James, C. L. Ebens, J. M. Walter, A. Zijdenbos, A. C. Evans, J. N. Giedd, and J. L. Rapoport. Developmental trajectories of brain volume abnormalities in children and adolescents with attention-deficit/hyperactivity disorder. *JAMA*, 288:1740–1748, Oct 2002.

S. Chakrabarti, M. van den Berg, and B. Dom. Focused crawling: a new approach to topic-specific web resource discovery. *Comput Netw*, 31(11-16):1623–1640, Jan 1999.

F. Chung. A local graph partitioning algorithm using heat kernel pagerank. In *Proceedings of the 6th International Workshop on Algorithms and Models for the Web-Graph*, WAW '09, pages 62–75, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 978-3-540-95994-6. URL http://dx.doi.org/10.1007/978-3-540-95995-3_6.

F. Chung and W. Zhao. Pagerank and random walks on graphs, 2008.

R. Cohen and S. Havlin. Scale-free networks are ultrasmall. *Phys Rev Lett*, 90:058701, 2003. doi: 10.1103/PhysRevLett.90.058701.

J. J. Collins and C. C. Chow. It's a small world. *Nature*, 393:409–410, Jun 1998.

S. Dasgupta. The hardness of k -means clustering. Technical Report CS2008-0916, University of California - San Diego Computer Science Department, January 2008.

D. W. Davis and P. G. Williams. Attention deficit/hyperactivity disorder in preschool-age children: issues and concerns. *Clin Pediatr (Phila)*, 50:144–152, Feb 2011.

M. de Kunder. The size of the world wide web. http://www.worldwidewebsize.com/, 2011. [Online; Accessed: 04/19/2011].

B. E. Depue, G. C. Burgess, L. C. Bidwell, E. G. Willcutt, and M. T. Banich. Behavioral performance predicts grey matter reductions in the right inferior frontal gyrus in young adults with combined type ADHD. *Psychiatry Res*, 182:231–237, Jun 2010.

P. Dibbets, E. A. Evers, P. P. Hurks, K. Bakker, and J. Jolles. Differential brain activation patterns in adult attention-deficit hyperactivity disorder (ADHD) associated with task switching. *Neuropsychology*, 24:413–423, Jul 2010.

R. P. W. Duin, P. Juszczak, P. Paclík, E. Pekalska, D. de Ridder, D. M. J. Tax, and S. Verzakov. PR-Tools4.1, a matlab toolbox for pattern recognition, 2007. http://prtools.org.

S. Durston, N. T. Tottenham, K. M. Thomas, M. C. Davidson, I. Eigsti, Y. Yang, A. M. Ulug, and B. J. Casey. Differential patterns of striatal activation in young children with and without adhd. *Biological Psychiatry*, 53(10):871 – 878, 2003. ISSN 0006-3223. doi: DOI:10.1016/S0006-3223(02)01904-2. URL http://www.sciencedirect.com/science/article/pii/S0006322302019042.

S. Durston, H. E. Hulshoff Pol, H. G. Schnack, J. K. Buitelaar, M. P. Steenhuis, R. B. Minderaa, R. S. Kahn, and H. van Engeland. Magnetic resonance imaging of boys with attention-deficit/hyperactivity disorder and their unaffected siblings. *J Am Acad Child Adolesc Psychiatry*, 43:332–340, Mar 2004.

P. Erdös and A. Rényi. On random graphs, i. *Publicationes Mathematicae (Debrecen)*, 6: 290–297, 1959. URL http://www.renyi.hu/~{}p_erdos/Erdos.html#1959-11.

P. Erdös and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5:17–61, 1960.

P. Erdös and A. Rényi. On the strength of connectedness of a random graph. *Acta Mathematica Hungarica*, 12:261–267, 1961. ISSN 0236-5294. URL http://dx.doi.org/10.1007/BF02066689. 10.1007/BF02066689.

E. Garfield. New international professional society signals the maturing of scientometrics and informetrics. *The Scientist*, Jan 1995. URL http://www.the-scientist.com/images/yr1995/august/issi_950821.html.

D. F. Gleich. *Models and Algorithms for PageRank Sensitivity*. PhD thesis, Stanford University, September 2009. URL http://www.stanford.edu/group/SOL/dissertations/pagerank-sensitivity-thesis-online.pdf.

P. Grindrod. Range-dependent random graphs and their application to modeling large small-world Proteome datasets. *Phys Rev E Stat Nonlin Soft Matter Phys*, 66:066702, Dec 2002.

R. Guimera and L. A. N. Amaral. Modeling the world-wide airport network. *The European Physical Journal B - Condensed Matter and Complex Systems*, 38:381–385, 2004. ISSN 1434-6028. URL http://dx.doi.org/10.1140/epjb/e2004-00131-0. 10.1140/epjb/e2004-00131-0.

T. H. Haveliwala. Topic-sensitive pagerank, 2002.

T. H. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, 15:2003, 2003.

D. J. Higham. Unravelling small world networks. *Journal of Computational and Applied Mathematics*, 158(1):61 – 74, 2003. ISSN 0377-0427. doi: DOI:10.1016/S0377-0427(03)00471-0. URL http://www.sciencedirect.com/science/article/pii/S0377042703004710. Selection of papers from the Conference on Computational and Mathematical Methods for Science and Engineering, Alicante University, Spain, 20-25 September 2002.

C. C. Hilgetag, G. A. P. C. Burns, M. A. O. O'Neill, J. W. Scannell, and M. P. Young. Anatomical connectivity defines the organization of clusters of cortical areas in the macaque monkey and the cat. *Phil Trans R Soc Lond B*, 355(1393):91–110, 2000. doi: 10.1098/rstb.2000.0551.

D. E. Hill, R. A. Yeo, R. A. Campbell, B. Hart, J. Vigil, and W. Brooks. Magnetic resonance imaging correlates of attention-deficit/hyperactivity disorder in children. *Neuropsychology*, 17:496–506, Jul 2003.

G. Jeh and J. Widom. Scaling personalized web search. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 271–279, New York, NY, USA, 2003. ACM. ISBN 1-58113-680-3. doi: http://doi.acm.org/10.1145/775152.775191. URL http://doi.acm.org/10.1145/775152.775191.

H. Jeong, B. Tombor, R. Albert, Z. N. Oltvar, and A. L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407:651–654, 2000. doi: 10.1038/35036627.

I. T. Jolliffe. *Principal component analysis*. Springer, New York, 2002.

M. Kaiser and C. C. Hilgetag. Edge vulnerability in neural and metabolic networks. *Biol Cybern*, 90:311–317, 2004. doi: 10.1007/s00422-004-0479-1.

F. Karinthy. Chain-links. *'Everything is different'*, 1929.

F. R. Karsz, A. Vance, V. A. Anderson, P. G. Brann, S. J. Wood, C. Pantelis, and W. J. Brewer. Olfactory impairments in child attention-deficit/hyperactivity disorder. *J Clin Psychiatry*, 69:1462–1468, Sep 2008.

J. M. Kleinberg. Navigation in a small world. *Nature*, 406:845, Aug 2000.

A. N. Langville and C. D. Meyer. Deeper inside pagerank. *Internet Mathematics*, 1(3): 335–380, 2004.

S. Lawrence and C. L. Giles. Accessibility of information on the web. *Nature*, 400(6740): 107–109, Jan 1999.

LivingInternet. Internet statistics. http://www.livinginternet.com/i/im_stats.htm. [Online; Accessed: 08/14/2011].

S. P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

R. Duncan Luce and Albert D. Perry. A method of matrix analysis of group structure. *Psychometrika. A Journal Devoted to the Development of Psychology as a Quantitative Rational Science*, 14:95–116, 1949.

M. Mahajan, P. Nimbhorkar, and K. Varadarajan. The planar k-means problem is np-hard. In *Proceedings of the 3rd International Workshop on Algorithms and Computation*, WALCOM '09, pages 274–285, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 978-3-642-00201-4. URL http://dx.doi.org/10.1007/978-3-642-00202-1_24.

D. S. Mandell, W. W. Thompson, E. S. Weintraub, F. Destefano, and M. B. Blank. Trends in diagnosis rates for autism and ADHD at hospital discharge in the context of other psychiatric diagnoses. *Psychiatr Serv*, 56:56–62, Jan 2005.

F. A. Middleton and P. L. Strick. Cerebellar projections to the prefrontal cortex of the primate. *J. Neurosci.*, 21:700–712, Jan 2001.

S. Milgram. The small world problem. *Psychology today*, 2:60–67, 1967.

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

M. E. J. Newman. The structure and function of complex networks. *Siam Rev*, 45(2): 167–256, Jan 2003b.

F. A. Nielsen, M. S. Christensen, K. H. Madsen, T. E. Lund, and L. K. Hansen. fMRI neuroinformatics. *IEEE Eng Med Biol Mag*, 25:112–119, 2006.

S. Ogawa, T. M. Lee, A. R. Kay, and D. W. Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc. Natl. Acad. Sci. U.S.A.*, 87:9868–9872, Dec 1990a.

S. Ogawa, T. M. Lee, A. S. Nayak, and P. Glynn. Oxygenation-sensitive contrast in magnetic resonance image of rodent brain at high magnetic fields. *Magn Reson Med*, 14: 68–78, Apr 1990b.

S. Overmeyer, E. T. Bullmore, J. Suckling, A. Simmons, S. C. Williams, P. J. Santosh, and E. Taylor. Distributed grey and white matter deficits in hyperkinetic disorder: MRI evidence for anatomical abnormality in an attentional network. *Psychol Med*, 31:1425–1435, Nov 2001.

L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1998.

M. Penrose. *Geometric Random Graphs*. Oxford University Press, 2003.

E. Perlov, A. Philipsen, L. Tebartz van Elst, D. Ebert, J. Henning, S. Maier, E. Bubl, and B. Hesslinger. Hippocampus and amygdala morphology in adults with attention-deficit hyperactivity disorder. *J Psychiatry Neurosci*, 33:509–515, Nov 2008.

D. J. Peterson, M. Ryan, S. L. Rimrodt, L. E. Cutting, M. B. Denckla, W. E. Kaufmann, and E. M. Mahone. Increased Regional Fractional Anisotropy in Highly Screened Attention-Deficit Hyperactivity Disorder (ADHD). *J Child Neurol*, May 2011.

K. J. Plessen, R. Bansal, H. Zhu, R. Whiteman, J. Amat, G. A. Quackenbush, L. Martin, K. Durkin, C. Blair, J. Royal, K. Hugdahl, and B. S. Peterson. Hippocampus and amygdala morphology in attention-deficit/hyperactivity disorder. *Arch. Gen. Psychiatry*, 63: 795–807, Jul 2006.

N. Przulj, D. G. Corneil, and I. Jurisica. Modeling interactome: scale-free or geometric?. *Bioinformatics*, 20(18):3508–3515, 2004. URL http://dblp.uni-trier.de/db/journals/bioinformatics/bioinformatics20.html#PrzuljCJ04.

Discovering Psychology. The human brain. http://www.learner.org/discoveringpsychology/brain/index.html. [Online; Accessed: 08/16/2011].

M. Qiu, Z. Ye, Q. Li, G. Liu, B. Xie, and J. Wang. Changes of brain structure and function in adhd children. *Brain Topography*, pages 1–10, 2010. ISSN 0896-0267. URL http://dx.doi.org/10.1007/s10548-010-0168-4. 10.1007/s10548-010-0168-4.

M. E. Raichle. Circulatory and metabolic correlates of brain function in normal humans. *Comprehensive Physiology 2011, Supplement 5: Handbook of Physiology, The Nervous System, Higher Functions of the Brain*, pages 643–674, 1987.

A. Rapoport. Contribution to the theory of random and biased nets. *Bulletin of mathematical biology*, Jan 1957. URL http://www.springerlink.com/index/MN2U24251N05N45J.pdf.

ReelSEO. 25 jawdropping youtube video facts, figures and statistics. http://www.reelseo.com/youtube-statistics/, April 2011. [Online; Accessed: 04/20/2011].

J. Rosack. Brain Scans Reveal Physiology of ADHD. *Psychiatr News*, 39(1):26–27, January 2004. URL http://pn.psychiatryonline.org/cgi/content/abstract/39/1/26.

C. S. Roy and C. S. Sherrington. On the Regulation of the Blood-supply of the Brain. *J. Physiol. (Lond.)*, 11:85–158, Jan 1890.

K. Rubia, A. Cubillo, A. B. Smith, J. Woolley, I. Heyman, and M. J. Brammer. Disorder-specific dysfunction in right inferior prefrontal cortex during two inhibition tasks in boys with attention-deficit hyperactivity disorder compared to boys with obsessive-compulsive disorder. *Hum Brain Mapp*, 31:287–299, Feb 2010.

D. Sasayama, A. Hayashida, H. Yamasue, Y. Harada, T. Kaneko, K. Kasai, S. Washizuka, and N. Amano. Neuroanatomical correlates of attention-deficit-hyperactivity disorder accounting for comorbid oppositional defiant disorder and conduct disorder. *Psychiatry Clin. Neurosci.*, 64:394–402, Aug 2010.

M. Schecklmann, E. Schenk, A. Maisch, S. Kreiker, C. Jacob, A. Warnke, M. Gerlach, A. J. Fallgatter, and M. Romanos. Altered frontal and temporal brain function during olfactory stimulation in adult attention-deficit/hyperactivity disorder. *Neuropsychobiology*, 63:66–76, 2011.

L. J. Seidman, E. M. Valera, and N. Makris. Structural brain imaging of attention-deficit/hyperactivity disorder. *Biol. Psychiatry*, 57:1263–1272, Jun 2005.

L. J. Seidman, J. Biederman, L. Liang, E. M. Valera, M. C. Monuteaux, A. Brown, J. Kaiser, T. Spencer, S. V. Faraone, and N. Makris. Gray matter alterations in adults with attention-deficit/hyperactivity disorder identified by voxel based morphometry. *Biol. Psychiatry*, 69:857–866, May 2011.

P. Shaw, M. Gilliam, M. Liverpool, C. Weddle, M. Malek, W. Sharp, D. Greenstein, A. Evans, J. L. Rapoport, and J. Giedd. Cortical development in typically developing children with symptoms of hyperactivity and impulsivity: support for a dimensional view of attention deficit hyperactivity disorder. *Am J Psychiatry*, 168:143–151, Feb 2011.

Lindsay I Smith. A tutorial on principal components analysis. Technical report, Cornell University, USA, February 26 2002.

O. Sporns and J.D. Zwi. The small-world of the cerebral cortex. *Neuroinformatics*, 2(2): 145–162, 2004. doi: 10.1385/NI:2:2:145.

C. Stam and J. Reijneveld. Graph theoretical analysis of complex networks in the brain. *Nonlinear Biomedical Physics*, 1(1):3, 2007. ISSN 1753-4631. doi: 10.1186/1753-4631-1-3. URL http://www.nonlinearbiomedphys.com/content/1/1/3.

K. Supekar, M. Musen, and V. Menon. Development of large-scale functional brain networks in children. *PLoS Biol.*, 7:e1000157, Jul 2009.

S. J. Suskauer, D. J. Simmonds, B. S. Caffo, M. B. Denckla, J. J. Pekar, and S. H. Mostofsky. fMRI of intrasubject variability in ADHD: anomalous premotor activity with prefrontal compensation. *J Am Acad Child Adolesc Psychiatry*, 47:1141–1150, Oct 2008a.

S. J. Suskauer, D. J. Simmonds, S. Fotedar, J. G. Blankner, J. J. Pekar, M. B. Denckla, and S. H. Mostofsky. Functional magnetic resonance imaging evidence for abnormalities in response selection in attention deficit hyperactivity disorder: differences in activation associated with response inhibition but not habitual motor response. *J Cogn Neurosci*, 20:478–493, Mar 2008b.

L. Tamm, V. Menon, and A. L. Reiss. Parietal attentional system aberrations during target detection in adolescents with attention deficit hyperactivity disorder: event-related fMRI evidence. *Am J Psychiatry*, 163:1033–1043, Jun 2006.

J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):pp. 425–443, 1969. ISSN 00380431. URL http://www.jstor.org/stable/2786545.

Twitter. #numbers. http://blog.twitter.com/2011/03/numbers.html, March 2011. [Online; Accessed: 04/20/2011].

N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, 15:273–289, Jan 2002.

V. M. Verkhlyutov, G. V. Gapienko, V. L. Ushakov, G. V. Portnova, I. A. Verkhlyutova, N. V. Anisimov, and Y. A. Pirogov. MRI morphometry of the cerebral ventricles in patients with attention deficit hyperactivity disorder. *Neurosci. Behav. Physiol.*, 40:295–303, Mar 2010.

D. J. Watts and S. H. Strogatz. Collective dynamics of "small-world" networks. *Nature*, 393:440–442, 1998. doi: 10.1038/30918.

I. Xenarios, D. W. Rice, L. Salwinski, M. K. Baron, E. M. Marcotte, and D. Eisenberg. DIP: the database of interacting proteins. *Nucleic Acids Res.*, 28:289–291, Jan 2000.

H. Yang, Q. Z. Wu, L. T. Guo, Q. Q. Li, X. Y. Long, X. Q. Huang, R. C. Chan, and Q. Y. Gong. Abnormal spontaneous brain activity in medication-naâĹŽÃŸve ADHD children: A resting state fMRI study. *Neurosci Lett*, Jul 2011.

STATISTICAL TABLES

TABLE A.1: *p*-values for between group t-test for the CPL

1	**2**	**3**	**4**	**5**	**6**
0.023	0.064	0.587	0.365	0.859	0.987
0.014	0.048	0.566	0.330	0.824	0.937
0.017	0.055	0.557	0.254	0.758	0.988
0.009	0.040	0.571	0.165	0.659	0.770
0.014	0.062	0.603	0.157	0.617	0.643
0.013	0.063	0.645	0.149	0.617	0.736
0.016	0.073	0.679	0.193	0.721	0.901
0.013	0.069	0.604	0.141	0.628	0.753
0.013	0.068	0.551	0.140	0.586	0.694
0.008	0.047	0.623	0.098	0.509	0.557
0.014	0.062	0.680	0.128	0.567	0.627
0.015	0.072	0.664	0.123	0.588	0.683
0.020	0.090	0.784	0.118	0.581	0.734
0.026	0.127	0.673	0.166	0.634	0.962
0.040	0.134	0.749	0.126	0.583	0.849
0.052	0.158	0.589	0.159	0.519	0.806
0.025	0.084	0.480	0.121	0.482	0.925
0.036	0.140	0.273	0.204	0.518	0.693
0.060	0.179	0.231	0.164	0.453	0.700
0.159	0.383	0.223	0.218	0.460	0.828
0.225	0.466	0.151	0.230	0.425	0.852
0.391	0.734	0.117	0.344	0.429	0.620
0.565	0.668	0.259	0.425	0.531	0.900
0.990	0.816	0.258	0.372	0.538	0.953
0.880	0.999	0.131	0.654	0.564	0.790
0.962	0.807	0.385	0.376	0.483	0.999
0.279	0.523	0.397	0.668	0.528	0.698
0.344	0.429	0.298	0.463	0.813	0.901
0.320	0.400	0.706	0.194	0.834	0.994
0.298	0.590	0.925	0.365	0.802	0.740
0.444	0.697	0.576	0.336	0.948	0.987
0.130	0.218	0.781	0.107	0.735	0.847
0.086	0.143	0.699	0.466	0.885	0.662
0.025	0.157	0.419	0.757	0.669	0.563
0.015	0.201	0.701	0.576	0.936	0.578
0.025	0.252	0.889	0.710	0.920	0.335

Columns:

1 TD / ADHD
2 Male / Female
3 Male TD / ADHD
4 Female TD / ADHD
5 Male TD / Female TD
6 Male ADHD / Female ADHD

TABLE A.2: *p*-values for between group t-test for the clustering coefficient

1	**2**	**3**	**4**	**5**	**6**
0.019	0.222	0.364	0.399	0.986	0.555
0.022	0.213	0.337	0.379	0.981	0.549
0.022	0.197	0.288	0.370	0.972	0.567
0.026	0.204	0.261	0.351	0.987	0.511
0.032	0.195	0.231	0.316	0.997	0.475
0.034	0.210	0.214	0.315	0.996	0.472
0.028	0.182	0.155	0.402	0.902	0.491
0.023	0.193	0.135	0.455	0.847	0.456
0.033	0.234	0.144	0.426	0.866	0.455
0.042	0.240	0.148	0.419	0.884	0.434
0.032	0.252	0.148	0.374	0.868	0.478
0.036	0.276	0.180	0.363	0.848	0.606
0.029	0.332	0.183	0.364	0.785	0.680
0.025	0.279	0.161	0.353	0.800	0.678
0.022	0.331	0.121	0.402	0.711	0.743
0.029	0.295	0.087	0.354	0.712	0.576
0.024	0.277	0.069	0.298	0.733	0.473
0.019	0.298	0.112	0.261	0.838	0.375
0.021	0.333	0.082	0.204	0.797	0.306
0.026	0.411	0.049	0.282	0.660	0.343
0.044	0.421	0.045	0.362	0.577	0.362
0.047	0.252	0.048	0.248	0.681	0.277
0.072	0.344	0.029	0.163	0.725	0.274
0.054	0.333	0.016	0.209	0.614	0.348
0.062	0.221	0.004	0.283	0.476	0.442
0.055	0.254	0.007	0.212	0.505	0.345
0.105	0.419	0.013	0.199	0.638	0.308
0.117	0.491	0.014	0.166	0.640	0.318
0.183	0.488	0.017	0.241	0.488	0.418
0.115	0.729	0.016	0.132	0.504	0.425
0.149	0.922	0.008	0.314	0.242	0.553
0.417	0.630	0.013	0.131	0.399	0.353
0.984	0.634	0.009	0.307	0.243	0.717
0.981	0.776	0.022	0.209	0.290	0.818
0.910	0.654	0.070	0.250	0.351	0.889
0.677	0.722	0.142	0.077	0.574	0.753

Columns:

1 TD / ADHD
2 Male / Female
3 Male TD / ADHD
4 Female TD / ADHD
5 Male TD / Female TD
6 Male ADHD / Female ADHD

TABLE A.3: Regions found to be significantly different for TD children

Threshold (θ)	**Region**	**Region**	**Region**
1 (0.15)	0	0	0
2 (0.16)	0	0	0
3 (0.17)	0	0	0
4 (0.18)	82	0	0
5 (0.19)	0	0	0
6 (0.20)	0	0	0
7 (0.21)	0	0	0
8 (0.22)	0	0	0
9 (0.23)	57	86	0
10 (0.24)	0	0	0
11 (0.25)	0	0	0
12 (0.26)	57	0	0
13 (0.27)	57	58	86
14 (0.28)	86	0	0
15 (0.29)	2	57	58
16 (0.30)	1	101	0
17 (0.31)	1	88	0
18 (0.32)	1	2	88
19 (0.33)	1	88	107
20 (0.34)	1	0	0

TABLE A.4: Regions found to be significantly different for children with ADHD

Threshold (θ)	**Region**	**Region**
1 (0.15)	0	0
2 (0.16)	0	0
3 (0.17)	0	0
4 (0.18)	0	0
5 (0.19)	0	0
6 (0.20)	0	0
7 (0.21)	0	0
8 (0.22)	0	0
9 (0.23)	0	0
10 (0.24)	0	0
11 (0.25)	0	0
12 (0.26)	0	0
13 (0.27)	0	0
14 (0.28)	0	0
15 (0.29)	0	0
16 (0.30)	17	0
17 (0.31)	17	0
18 (0.32)	17	0
19 (0.33)	17	4
20 (0.32)	17	0

TABLE A.5: Errors obtained for each of the classifiers ($\theta = 0.2$)

Classifier	**Error** ($\theta = 0.2$)	**Min PCA**	**Mean PCA**
QDC	0.5000	0.3817	0.4207
PC	0.4624	0.4086	0.4344
LogL	0.4462	0.4032	0.4468
LD	0.4624	0.3978	0.4527
NM	0.4140	0.3925	0.4191
FL	0.4839	0.4194	0.4522
SVM (p=2)	0.4731	0.4247	0.4500
SVM (p=3)	0.4301	0.3871	0.4290
SVM (p=4)	0.4409	0.3978	0.4312
SVM (r=5)	0.4785	0.3978	0.4387
SVM (r=10)	0.4462	0.3817	0.4497
SVM (r=15)	0.4785	0.3871	0.4457