Structural Analysis of Brain Networks

Gabe Alvarez
galvare2@stanford.edu

December 12, 2016

1 Introduction

In recent years, advances in technology such as the fMRI have opened the door to a deeper understanding of the brains of both humans and other animals. The fMRI allows us to map out the activity of the brain in a detailed resolution, which has led to new modes of understanding the structure of the brain.

At its lowest level, the brain can be thought of as a network of connected neurons, with neurons represented as network nodes and connections represented as network edges. On higher levels, the brain can be viewed as a network as well, with large-scale regions of the brain as nodes and levels of communication between those regions as weighted edges. The brains of complex animals such as humans are so complicated that our current technology is incapable of generating neuron-level data for the whole brain, but region-level data still proves useful and complements our region-level understanding of the brains of such animals.

Network analysis provides us with a framework for understanding the structure of the brain with greater clarity. Simple algorithms have provided us with insight into how efficiently brain networks are laid out, specifically how they tend to exhibit small-world behavior, meaning high clustering coefficients and low average path lengths, as well as how they exhibit scale-free behavior, meaning structural patterns are mirrored when looking at the network on different scales.

A more sophisticated mode of understanding these networks comes from understanding their community structure. This means partitioning the network into several sub-networks which have high levels of connectedness within each sub-network and have low levels of connectedness between different sub-networks. This problem has many approaches, depending on the optimization used and the nature of the desired communities. One novel approach to this problem is to partition the network into communities using motifs, or specific arrangements of a few nodes, attempting not to split an occurrence of a motif across communities.

The structural layout of the brain has been shown to correspond with its real-world functioning, showing that understanding brains using network analysis will lead us to greater understanding of the differences between individuals and may give us insight into known brain pathologies. Some studies have linked cognitive performance with the global network structure of the brain, such as "Efficiency of Functional Brain Networks and Intellectual Performance" by Van den Heuvel et al. This paper found a correlation between IQ and certain metrics of network efficiency. Other papers, such as "Functional neural network analysis in frontotemporal dementia and Alzheimer's disease using EEG and graph theory," present evidence that specific brain pathologies are linked with large-scale disruptions in the brain's connectivity network. It is possible that similar useful conclusions could be drawn from analyzing the community structure of brains.

This paper consists of analyzing the basic properties, community structure, and motif structure of brain networks, cross-referencing these results with the names and functions of the regions contained in the communities found with these algorithms.

2 Related Work

This paper uses data sets generated from fMRI data on live subjects. Methods for collecting such data are summarized in "Exploring the brain network: A review on resting-state fMRI functional connectivity," a paper which provides a solid review of this field of research. As well as discussing the data collection methods, this paper also provides details on some basic properties of the networks generated using these methods. For example, Heuvel et al. describe how the human brain's network structure is optimized for efficiency, pointing to evidence that the brain follows a small-world organization, allowing for both high levels of local connectivity and short paths between any two regions of the brain. They also point out that the degree distribution follows an exponential truncated power law, which assigns much more probability to nodes having many connections, meaning that in real brain networks, there are a few nodes that serve as hubs, connecting to many different brain regions. The power law distribution also means that the brain exhibits scale-free properties, implying that analyzing the connections between individual neurons may yield similar results as analyzing the connections between small clusters of neurons or larger brain regions. The authors also mention that networks laid out according to power-law distributions are known to be robust to randomized failures of nodes, meaning the brain is designed to be able to handle the biological limitations of neurons that don't always perform perfectly.

Further analysis of similar properties is presented in "Studying the human brain anatomical network via diffusion-weighted MRI and Graph Theory" by Iturria-Medina et al. The authors of this paper list several global metrics to characterize the network, including betweenness centrality, clustering coefficient, average least path length, and degree distribution. There is also more detailed analysis, such as the concepts of local and global efficiency, which they define as capturing how easily information can flow through the network. Global efficiency is proportional to the sum of the inverses of the shortest path lengths, and local efficiency is proportional to this sum restricted to the neighbors of a given node. Next, there is vulnerability, which measures how much global efficiency decreases when a given node is removed. This metric is important in studying brains, as their biological hardware always has the possibility of malfunction. Finally, the paper briefly touches on the concept of motifs as an area of future exploration. The authors acknowledge that they are just scraping the surface, and better results could be yielded from more complicated analysis - they mention hierarchical features and functional motif composition as future directions.

My project aims to look at the higher-order structure of brain networks, rather than looking at purely local or purely global phenomena, and central to this is community detection. A popular community detection algorithm is the CNM algorithm, outlined in "Finding community structure in very large networks" by Aaron Clauset, M. E. J. Newman, and Cristopher Moore, which provides an efficient and scalable method of partitioning a network into communities.

A recent insight comes from "Higher-order organization of complex networks" by Benson et al., which outlines an algorithm to partition a graph into communities based on motifs. The algorithm in Benson et al. attempts to partition the network so that most occurrences of the motif are each contained in a single community, and few occurrences are split across communities.

Partitioning brain regions into communities is more useful with knowledge of what those regions actually do. The data sets I use for my work have labels which correspond to known brain regions, which can be looked up in published tables such as at brainmaps.org/index.php?p=abbrevs-macaca. These labels can then be cross-referenced with known information about these regions, and one can determine whether the community partitioning makes sense in light of this information.

3 Data Sets

My main data comes from the "Brain Connectivity Toolbox" at https://sites.google.com/site/bctnet/datasets. This data includes region-level connectivity net-

works for human, macaque, and cat brains. All the datasets are undirected and weighted, with nodes representing brain regions and edge weights indicating the strength of functional connections between the regions. The human dataset has 638 nodes and 18,625 edges, the macaque network has 71 nodes and 746 edges, and the cat dataset has 52 nodes and 820 edges. Each dataset comes with labels for all regions, allowing them to be looked up and understood in the context of neuroscientific knowledge about the region.

4 Methods and Algorithms

4.1 Standard Community Detection

Important algorithms used in my work are those which analyze higher-order properties of the networks using community detection. One commonly-used community detection algorithm is the Clauset-Newman-Moore (CNM) algorithm, which is useful because it combines both reasonable performance and results, unlike other algorithms such as the Girvan-Newman algorithm, which has trouble scaling to large networks. One can find the details of the CNM algorithm in "Finding community structure in very large networks." My project uses the CNM algorithm as a baseline reference for comparison with other algorithms which use different partitioning criteria.

4.2 Motif Sampling - Naive Method

The main algorithm I make use of is the motif-based community detection algorithm presented in "Higher-order organization of complex networks" by Benson et al., which allows for community partitioning based on the occurrence of motifs. Communities are partitioned with respect to a single motif, with the partitioning optimized so that many instances of the motif occur within a single community and few are cut across different communities.
In order to find meaningful results from this algorithm, one must run it using motifs which are significant in the brain network, meaning they occur more commonly than one would expect in a random network with the same degree distribution. I devised a simple algorithm for finding the most common motifs of size n in a network, by simply sampling n nodes at a time and counting up how many times each motif occurs in these sets of n nodes. One can then compare the result of this algorithm with a randomly-generated network of the same size. Those motifs which come up more often in the brain network than the randomly-generated network are likely significant to the structure of the brain network, and hence deserve closer examination. The benefit of this algorithm is that it does not require an exhaustive search of all nodes in the network, and it works for any value of n.

During testing, I learned that this method has its shortcomings. As an estimator of actual motif frequencies in the network, it is unbiased, but its variance is quite high. This is especially true for motifs which occur very infrequently in the network, notably many motifs of size 5.

4.3 Motif Sampling - 3-Path Method

More effective estimators of motif frequencies in a network can be constructed by taking into account structural information about the motifs themselves. This leads to estimation methods which have lower variance but which are less general than the naive method presented above. I implemented one such method, which is described in "Path Sampling: A Fast and Provable Method for Estimating 4-Vertex Subgraph Counts," by Jha, Seshadri, and Pinar. This paper presents a lower-variance method for estimating counts of motifs of size 4 in a network, but does not extend to motifs of other sizes.

The key insights behind this algorithm have to do with awareness of the structure of the motifs of size 4. All connected motifs of size 4 except the 3-star (see figure 4 in Results) have at least one path of length 3 in their structure. This means that instead of sampling sets of nodes, one can sample 3-paths randomly in the network and estimate total motif counts from the frequencies of each motif occurring along these 3-paths. One can finish by using linear relationships between subgraph counts to extrapolate the count of 3-star motifs from these results.

This algorithm works well for computing counts for motifs which do not contain a 4-cycle, but converges slowly for motifs which do. A small tweak of the algorithm makes it robust to these 4-cycle motifs. At each step, we randomly sample instead a *centered* 3-path, which is a 3-path in which the two ends are connected to form a 4-cycle, and where the labeling order of the nodes fits certain conditions. Again, we use these samples to count the number of times they appear in each type of motif and calculate the estimated motif counts from these samples.

4.4 Motif-Based Spectral Clustering

The above sampling methods point out motifs of interest in the networks. Using the most significant motifs, I run the Benson et al. algorithm for motif-based clustering. This algorithm partitions the network as follows.

Begin by computing the co-occurrence matrix W_M, an nxn matrix where n is the number of nodes in the network. Each entry W_{ij} in this matrix is the number of times the two nodes occur together in the chosen motif in the network.

From this matrix, compute its normalized Laplacian and compute the spectral ordering of this Laplacian. To do this, let D be the diagonal matrix $D_{ii} = \sum_j (W_M)_{ij}$, the matrix with row-sums of W_M on the diagonal. Let $D^{-1/2}$ be the same matrix but with each entry raised to the $-1/2$ power. The Laplacian matrix L_M can now be computed as $L_M = D^{-1/2}(D - W_M)D^{-1/2}$. To find the spectral ordering, we compute the eigenvalues of L_M and let z be the eigenvector corresponding to the second-smallest eigenvalue. From here, the spectral ordering s is the by-value ordering of $D^{-1/2}z$, i.e. each element s_i is the index of the i-th smallest value of $D^{-1/2}z$.

Finally, find the prefix set of s with the smallest motif conductance, formally, $S := argmin_r \phi_M(S_r)$, where $S_r = s_1, , s_r$. Motif conductance measures the extent to which occurrences of the given motif occur across different clusters, as opposed to occurring within one cluster. Specifically, it is the *cut* divided by the minimum *volume* of the clusters, where cut is the number of edges within the motif that cross between clusters, and volume is the total number of edges within the motif in a given cluster.

The Benson et al. algorithm is a community detection algorithm which takes structured information about motifs into account. I compare the results of the motif algorithm with the results of the CNM community detection algorithm described above.

5 Results and Analysis

5.1 Basic Properties

The first aspect of this project is getting to know the different data sets by assessing some basic measures of their structure.

Clustering Coefficient

Network	Clustering Coefficient
Macaque	0.496
Cat	0.553
Human	0.384

Across these three datasets, the simpler, less intelligent animals had higher clustering coefficients. All were roughly in the same ballpark.

Degree Distributions

A chart of the degree distributions of each network is shown in figure 1. Note that the networks are of different sizes, which is why the human dataset, which is the largest, has more nodes of high degree. All of them seem to follow a "long-tailed" pattern, although the small size of the data sets makes this hard to see for sure.

Path Length

Figure 2 is a chart of the distribution of shortest path length. Having short path lengths is a measure of the efficiency of information flow through the network. Interestingly, it seems that the cat network has the shortest paths, followed by human, then macaque.

5.2 CNM Community Detection

Running CNM community detection on the Macaque dataset led to the results in the following table, which

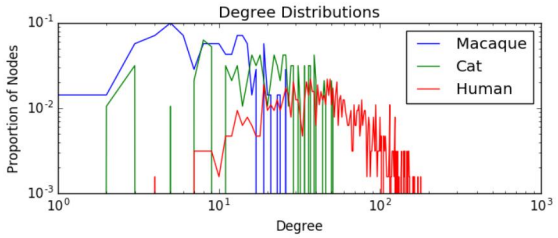

Figure 1: Degree Distributions

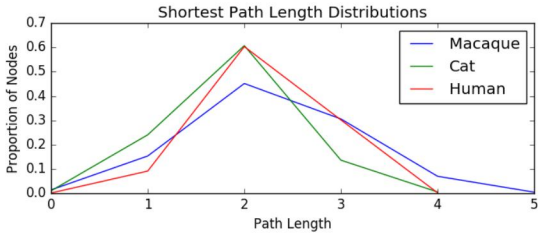

Figure 2: Shortest Path Length Distributions

shows its outputted community sizes and modularity values:

Network	Community Sizes	Modularity
Macaque	17, 19, 34	0.389
Cat	13, 15, 24	0.294
Human	103, 201, 334	0.364

For the cat and macaque networks, I had access to labels for each region, allowing me to analyze the neuroscience significance of the communities for each. The labels are abbreviations whose meanings can be found at brainmaps.org/index.php?p=abbrevs-cat for the cat dataset and brainmaps.org/index.php?p=abbrevs-macaca for the macaque dataset.

Here, I simply report the contents of each cluster. The significance of these results in terms of my project is how they compare with the results of motif-based partitioning, which I discuss in that section.

Macaque network:

Community One: ['V1', 'V2', 'V3', 'V3a', 'V4', 'V4t', 'MT', 'PO', 'PIP', 'VP', 'VOT', 'MSTd', 'MSTl', 'FST', 'LIP', 'VIP', 'DP', 'PITd', 'FEF']

Community Two: ['TF', 'PITv', 'CITd', 'CITv', 'AITv', 'TH', 'STPp', 'A7a', 'AITd', 'A46', 'STPa', 'TGV', 'ER', 'A35', 'A25', 'A12', 'A11', 'A13', 'A23', 'PAAL', 'TS2', 'TS3', 'A10', 'TGD', 'TS1', 'A24', 'A9', 'A32', 'PAAR', 'KA', 'PAL', 'PROA', 'REIT', 'PAAC']

Community Three: ['A7b', 'TPT', 'A5', 'S2', 'A6', 'SMA', 'A45', 'ID', 'G', 'A3a', 'A1', 'A2', 'A3b', 'A4', 'R1', 'A3', 'IG']

Cat network:

Community One: ['17', '18', '19', 'PLLS', 'PMLS', 'AMLS', 'VLS', '21a', '21b', 'ALLS', '7', '5Bl', 'DLS']

Community Two: ['20a', '20b', 'EPp', 'PS', '6m', 'PFCL', 'Ia', 'Ig', '35', 'Cga', 'CGp', '36', 'RS', 'AI', 'AII', 'AAF', 'P', 'VPc', 'Tem', 'PFCMd', 'Enr', 'PFCMil', 'Sb', 'pSb']

Community Three: ['6l', '5Al', '5Bm', 'AES', '4g', '4', '5Am', 'SSAi', 'SSAo', '3a', '3b', 'SIV', '1', '2', 'SII']

5.3 Motif Sampling - Naive

For each network, I ran the naive motif detection algorithm described above 5 times for 200,000 samples each. I considered as possible motifs all connected subgraphs of size 3-5. Visualizations of these subgraphs are presented below. Running the algorithm 5 times for each network allowed me to compute the variance across these runs to determine the validity of the estimator.

As a comparison baseline, I also ran the naive algorithm on randomly-generated preferential attachment networks with the same number of nodes and

a similar number of edges for each real network. For these estimates, I ran the algorithm 5 times for 200,000 samples each, where each run used a different randomly generated network, and I again computed the average and variance for each network.

I then computed a comparison by computing the ratio of occurrences of the motif in the real brain network compared to the random preferential attachment networks imitating it, called the Motif Occurrence Ratio. For the motifs of size 5, I applied a smoothing in which I added 1 occurrence of each motif for every network, allowing for comparisons of networks in which some motifs did not appear at all. This smoothing was not added for motifs of size 4 and 3 because every motif occurred at least once in every network.

I report here the Motif Occurrence Ratio for each motif. I also report some information about the standard deviations of the estimators, to get a sense of how much of the results are due to random noise.

Motifs of size 3

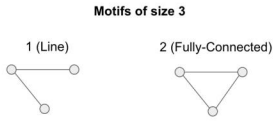

Figure 3: Motifs of size 3

There are two connected motifs of size 3 (see figure 3).

Motif Occurrence Ratios, Naive, Size 3

Motif	Macaque	Cat	Human
1	2.02	1.74	2.16
2	0.805	0.915	0.707

The "motif frequency" is the chance of any randomly-drawn set of nodes containing edges that form the given motif. The motif frequencies for motifs of size 3 were on the order of 10^{-2} with a std of 10^{-4} for macaque, 10^{-1} with a std of 10^{-4} for cat, and 10^{-2} with a std of 10^{-4} for human.

Across the three networks, the line motif consistently tended to occur more often than in random networks and the fully-connected motif occurred less often.

Motifs of size 4:

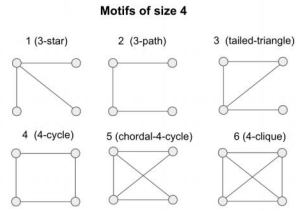

Figure 4: Motifs of size 4

There are 6 connected motifs of size 4 (see figure 4). In all three networks, the top two motifs were "fully-connected" and "one-missing", respectively. This result suggests that when considering groups of 4 regions, brain networks are consistently more likely than random networks to have extremely densely connected groups of 4 neurons.

Motif Occurrence Ratios, Naive, Size 4

Motif	Macaque	Cat	Human
1	0.452	0.583	0.326
2	0.692	0.923	0.563
3	1.15	1.17	1.03
4	0.75	0.789	0.444
5	2.02	1.5	1.9
6	4.02	2.87	5.57

For more information about the variances for motifs of size 4, see the later comparison with the path-based sampling method.

Motifs of size 5:

There are 23 connected motifs of size 5 (5).

Motif Occurrence Ratios, Naive, Size 5

Motif	Macaque	Cat	Human
17	3.46	1.84	2.83
22	5.82	2.77	6
23	3.25	3.88	2.25

The motif frequencies for motifs of size 5 were on the order of 10^{-3} with a std of 10^{-5} for macaque,

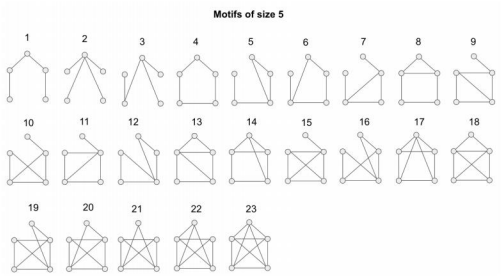

Figure 5: Motifs of size 5

10^{-2} with a std of 10^{-4} for cat, and 10^{-4} with a std of 10^{-5} for human.
The results for motifs of size 5 were more varied than smaller sizes, likely due to their low occurrence frequencies. The most connected motifs tended to have the highest occurrence rates here as well, similar to motifs of size 4. I reported above only these top-occurring motifs for brevity.

5.4 Motif Sampling - Path Method

I also computed the most commonly-occurring motifs in each network using the more sophisticated 3-path sampling method. As mentioned above, this method works only for motifs of size 4. As with the naive method, I ran this algorithm 5 times with 200,000 samples each time, on both the real brain networks and randomly-generated networks simulating them.

Motif Occ. Ratios, Path-Sampling, Size 4

Motif	Macaque	Cat	Human
1	0.522	0.578	0.333
2	0.793	0.927	0.562
3	1.29	1.19	1.04
4	0.939	0.788	0.494
5	2.11	1.5	1.95
6	4.75	2.8	4.96

To determine whether these results are more valid than the naive method, I compare the variances of the results with the two methods for the brain networks. The following tables compare estimator variances for each motif count on each network:

Sample Stds for Naive Method

Motif	Macaque	Cat	Human
1	0.000125	0.000321	6.17e-05
2	0.000236	0.000375	0.000166
3	7.19e-05	0.000342	5.73e-05
4	6.3e-05	0.000252	2.37e-05
5	9.14e-05	0.000557	5.63e-05
6	6.17e-05	6.8e-05	1.96e-05

Sample Stds for Path Sampling

Motif	Macaque	Cat	Human
1	3.36e-05	0.000262	1.08e-05
2	8.54e-05	0.00026	2.32e-05
3	4.48e-05	0.000276	5.53e-06
4	7.8e-06	3.8e-05	1.03e-06
5	6.88e-06	8.27e-05	1.86e-06
6	1.49e-06	1.14e-05	2.37e-07

One can see that the path sampling method performs significantly better as an estimator of occurrence counts for motifs of size 4, as its standard deviation is lower by around an order of magnitude for each motif.

5.5 Motif-Based Community Detection

I ran the motif partitioning algorithm for the top motif of each size for each network, as determined by the above methodology. For each network, I report the community sizes and motif conductance of the partitioning.

Network	Motif Size	Cmty Sizes	Conductance
Macaque	3	28, 41	0.257
Macaque	4	14, 55	0.250
Cat	3	25, 26	0.346
Cat	4	15, 36	0.250
Cat	5	14, 37	0.272
Human	3	199, 439	0.290

For brevity, I am unable to discuss all of the partitions above, so I focus on two which are of note. First, here are the clusters found for the macaque partitioning, based on the chordal-4-cycle motif of size 4:

Cluster one: ['V1', 'V2', 'V3', 'VP', 'V3a', 'V4', 'VOT', 'V4t', ' MT', 'STd', 'STl', 'FST', 'PIP', 'VIP']
Cluster two: ['ITv', 'ITd', 'ITv', 'ITd', 'ITv', 'TPp', 'TPa', ' TF', ' TH', ' PO', 'LIP', ' DP', 'A7a', 'FEF', 'A46', 'TGV', ' ER', 'A3b', ' A1', ' A2', ' A5', ' R1', ' S2', 'A7b', ' IG', 'A35', ' A4', ' A6', 'SMA', ' A3', 'A23', 'A24', ' A9', 'A32', 'A25', 'A10', 'A45', 'A12', 'A13', 'AAL', 'PAL', 'ROA', 'TGD', 'TS1', 'TS2', 'TS3']
Cluster one seems to contain all the visual regions, various sensory processing and perceptual-motor coordination regions (the posterior intraparietal area, ventral intraparietal area, and the temporal sulcus), with a few temporal regions such as middle temporal area and FST.
Cluster two contains all the auditory cortex regions and a scattering of all other areas. It seems that the algorithm placed into cluster one a few highly-related regions and placed all the other "miscellaneous" regions into cluster two. Importantly, brain regions that fit very similar purposes were put into the same clusters - for example, all the visual cortex areas were put into cluster one and all the auditory cortex areas were put into cluster two. This shows that the motif partitioning took into account regions which were highly connected with one another, which makes sense as the motif in question was a subgraph with very dense connections.

Next, here are the clusters found in the cat brain based on motif #23 of size 5:
Cluster one: ['5Am', '5Al', '5Bm', 'SSAi', 'SSAo', '3a', '3b', '1', '2', 'SII', 'SIV', '4g', '4', '6l']
Cluster two: ['17', '18', '19', 'PLLS', 'PMLS', 'AMLS', 'ALLS', 'VLS', 'DLS', '21a', '21b', '20a', '20b', '7', 'AES', 'PS', 'AI', 'AII', 'AAF', 'P', 'VPc', 'EPp', 'Tem', '6m', '5Bl', 'PFCMil', 'PFCMd', 'PFCL', 'Ia', 'Ig', 'Cga', 'CGp', 'RS', '35', '36', 'pSb', 'Enr']

Here, the partitioning was done based on an even more dense motif. We can see again here that motifs were clustered based on very close relationships between regions - all of the areas in regions 1-5 were assigned to cluster one, and all of the higher regions were assigned to cluster two.

For both of the above partitions, the small cluster is similar to one of the communities detected using CNM. However, the small cluster here is smaller than its corresponding cluster under CNM. This suggests that partitioning based on very dense motifs like we are doing here is similar to standard community detection under CNM, but with a higher threshold of being closely connected to make it into the small cluster, as the motif only counts for sets of nodes which are connected very closely.

As another means of comparison, I also implemented and ran the standard spectral community detection algorithm on these same networks, which attempts to minimize the number of overall cuts of edges, rather than of motif occurrences. This algorithm attempts to minimize conductance and hence should produce similar results to CNM, and this is exactly what I found. The algorithm is designed to output two communities. In all the networks, it output a small, densely-connected community which was nearly identical to the small, densely-connected community returned by CNM. The other community returned by the spectral algorithm was nearly identical to the union of all the other communities returned by CNM.

References

Madhav Jha, C. Seshadhri, Ali Pinar. "Path Sampling: A Fast and Provable Method for Estimating 4-Vertex Subgraph Counts." WWW 15. May 18 - 22, 2015. Pages 495-505.

Van den Heuvel, Martijn P. et al. "Exploring the brain network: A review on resting-state fMRI functional connectivity" European Neuropsychopharmacology. Volume 20, Issue 8, August 2010, Pages 519534.

Iturria-Medina, Yasser et al. "Studying the human brain anatomical network via diffusion-weighted MRI and Graph Theory" NeuroImage. Volume 40, Issue 3, 15 April 2008, Pages 10641076.

Van den Heuvel, Martijn P. et al. "Efficiency of Functional Brain Networks and Intellectual Performance." Journal of Neuroscience. June 2009, 29 (23) 7619-7624

De Haan, Willem et al. "Functional neural network analysis in frontotemporal dementia and Alzheimer's disease using EEG and graph theory." BMC Neuroscience. 2009. 10:101

Fortunato, Santo. "Community Detection in Graphs." Physics Reports. Volume 486, Issues 35, February 2010, Pages 75174.

Benson, Austin R. et al. "Higher-order organization of complex networks." Science. 08 Jul 2016: Vol. 353, Issue 6295, pp. 163-166.

Milo, R. et al. "Network Motifs: Simple Building Blocks of Complex Networks." Science. 25 Oct 2002: Vol. 298, Issue 5594, pp. 824-827