Towards the computer-aided diagnosis of dementia based on the geometric and network connectivity of structural MRI data

Towards the Computer-aided Diagnosis of Dementia based on the Geometric and Network Connectivity of Structural MRI Data

Garry M. Smith[1,2], Zhivko V. Stoyanov[1,2], Danica V. Greetham[3], Peter Grindrod[4], J. Doug Saddy[2], and for the Alzheimers Disease Neuroimaging Initiative*

[1] School of Systems Engineering, University of Reading, Reading, RG6 6AY, UK,
[2] Centre for Integrative Neuroscience and Neurodynamics, University of Reading, RG6 6AH, UK
[3] Centre for the Mathematics of Human Behaviour, Department of Mathematics and Statistics, University of Reading, RG6 6AX, UK
[4] Mathematical Institute, University of Oxford, Oxford, OX2 6GG, UK

{g.m.smith, z.v.stoyanov, j.d.saddy, d.v.greetham}@reading.ac.uk
peter.grindrod@maths.ox.ac.uk

Abstract. We present an intuitive geometric approach for analysing the structure and fragility of T1-weighted structural MRI scans of human brains. Apart from computing characteristics like the surface area and volume of regions of the brain that consist of highly active voxels, we also employ Network Theory in order to test how close these regions are to breaking apart. This analysis is used in an attempt to automatically classify subjects into three categories: Alzheimer's disease, mild cognitive impairment and healthy controls, for the CADDementia Challenge.

Keywords: MRI, dementia, mild cognitive impairment, voxel, automatic, diagnosis, graph Laplacian, Network Theory

1 Introduction

The UK government reports[5] that there are currently 800,000 dementia sufferers in the UK alone, with the disease costing the economy £23 billion per year. By 2040 the government estimates the costs associated with the disease

* Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

[5] https://www.gov.uk/government/policies/improving-care-for-people-with-dementia

will triple as the number of dementia sufferers increases to 1.6 million people. The government is responding through a number of initiatives including: The National Dementia Strategy[6] published in 2009; The Prime Minister's Dementia Challenge[7] launched in 2012; and by increasing the annual funding of dementia research to $\approx$ £66 million by 2015. As well as gaining a better understanding of the disease, the government aims to increase diagnosis rates so that they are among the best in Europe. Clearly the risk of neurological diseases, such as Alzheimer's disease (AD), to public health is of international importance and as such cross-cutting, interdisciplinary research combining ideas from across fields has the potential to contribute to the future well-being of global health.

The CADDementia Challenge[8] was established by the Biomedical Imaging Group Rotterdam, Erasmus MC, Rotterdam, in order to provide a standardised evaluation framework for the Computer-Aided Diagnosis of Dementia, based on T1-weighted structural Magnetic Resonance Image (MRI) data. The primary aim of the CADDementia Challenge is to objectively validate the different image-based diagnosis/classification methods that are emerging from research centres, such that suitably robust techniques may be identified as candidates for clinical use. The CADDementia competition requires participants to[8]: i. use a common dataset for training algorithms, as well as ii. a previously unseen multi-centre test dataset (to avoid over-training), and iii. to perform a multi-class diagnosis of Alzheimer's disease, Mild Cognitive Impairment (MCI) and controls.

We present an intuitive geometric algorithm for analysing the structure and fragility of MRI data. Apart from computing characteristics like the surface area and volume of regions of the brain that consist of highly active voxels, we also employ Network Theory[2] in order to test how close these regions are to breaking apart. This analysis is used in an attempt to classify structural MRI scans into three categories: CN (controls), MCI and AD.

The algorithm presented in this paper was executed over MRI data from the CADDementia and ADNI datasets using up to 120 computer CPU cores simultaneously to perform the analysis. However, with minor modifications to the job submission and data handling mechanisms currently employed, the same algorithm could be scaled up to massive numbers of CPU cores, provisioned on demand, through private and/or public Cloud providers, thereby potentially, allowing health authorities to offer screening as part of routine health checks to a larger proportion of the population and at greater frequencies.

2 The geometric and network structure of MRI data

Mathematically, MRI data can be considered as a collection of small cuboids (voxels) and for each of them we have a non-negative value which represents tissue properties. A central point in our approach is that in AD part of the neural

[6] https://www.gov.uk/government/publications/living-well-with-dementia-a-national-dementia-strategy

[7] http://dementiachallenge.dh.gov.uk/

[8] http://grand-challenge.org/site/caddementia/home/

mass degenerates progressively, therefore it is reasonable to assume that negative changes in the T1 signal gradients would be a feature of neural degeneration (this would be reverse for T2). We use these changes in voxel values as markers of degeneration in order to trace a path of similarity over long distances in the brain.

A $3D$ T1-weighted MRI image consists of $n_1 \times n_2 \times n_3$ voxels and $f(i,j,k) \geq 0$ is the level of T1-weighted signal recorded in voxel (i,j,k). Then we define

$$M := \max\{f(i,j,k) \mid 1 \leq i \leq n_1, 1 \leq j \leq n_2, 1 \leq k \leq n_3\}.$$

Further, for each voxel, instead of the recorded signal, $f(i,j,k)$, we consider a normalised signal,

$$g(i,j,k) := \frac{f(i,j,k)}{M}, \tag{1}$$

where we now have $0 \leq g(i,j,k) \leq 1$ for all voxels; we normalise the signal according to (1) for each brain. We can think of this as a way of introducing a similar scaling across all brains. The idea is to focus on the part of the brain (i.e. on those voxels) for which the signal is above a certain threshold, θ. Mathematically, this set is denoted by

$$A_\theta := \{(i,j,k) \mid g(i,j,k) \geq \theta, 1 \leq i \leq n_1, 1 \leq j \leq n_2, 1 \leq k \leq n_3\}.$$

We work with $\theta = 0.6, 0.61, 0.62, \ldots, 0.8$. This follows from both computational and physiological reasons: computing eigenvalues for a whole brain matrix is computationally intractable, but more importantly starting at a 0.6 threshold allows us to generate connectivity networks based on primarily white matter values (see Figures 1 & 2).

For each brain, we consider the 3D set A_θ and compute its surface area, S_θ, and its volume, V_θ; we also compute a measure of the fragility of its structure, f_θ, i.e how close A_θ is to "breaking" apart into smaller components.

Apart from being a geometrical 3D object, we can think of A_θ as a network, denoted by N_θ, in which two voxels are connected if they share a face or an edge (but not a corner).

The advantage of interpreting A_θ, as a graph, or a network N_θ, is that we can apply certain techniques from Spectral Graph Theory. Each graph/network can be represented with a matrix. Computing eigenvalues of such a matrix gives us a spectrum - an array of values that describes some structural characteristics of the given graph. The most widely used matrices assigned to graphs are adjacency, Laplacian or normalised Laplacian matrices. For a review of using spectra of graphs in computational biology see [1]. A comprehensive study of a normalised Laplacian spectrum with detailed definitions and many examples of its applications is given in [2]. A useful property of the eigenvalues of the normalised Laplacian matrix is that they are all real and are between 0 and 2 for any number of vertices and edges. Furthermore, the number of zero eigenvalues correspond to the number of connected components of the corresponding graph [2]. The smallest positive eigenvalue of the Laplacian matrix is called *algebraic connectivity* [4] and is an indicator of the robustness of the graph [7] to

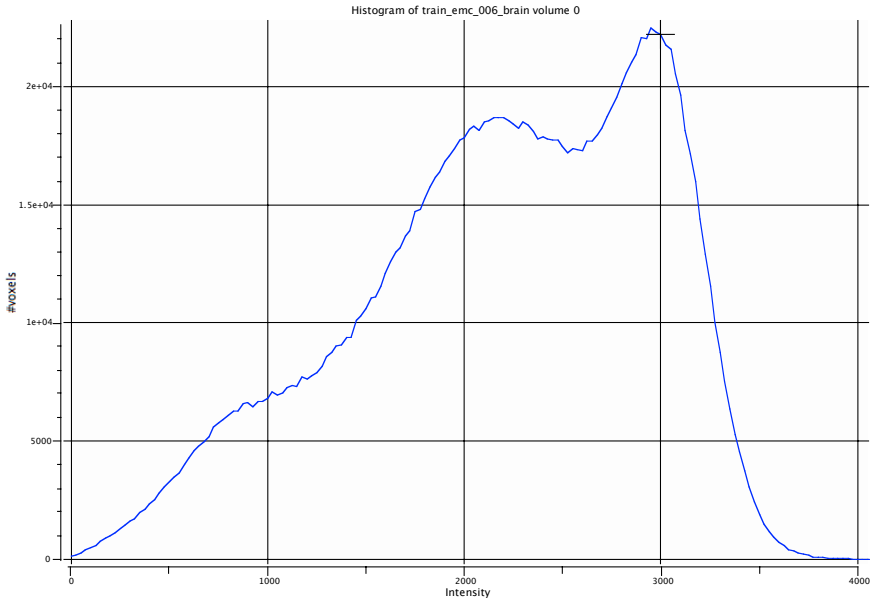

Fig. 1. A histogram showing the distribution of the intensities across a brain; the two peaks roughly indicate the range of intensities for white and grey matter. We can see that by choosing $\theta \geq 0.6$ we predominantly select white matter.

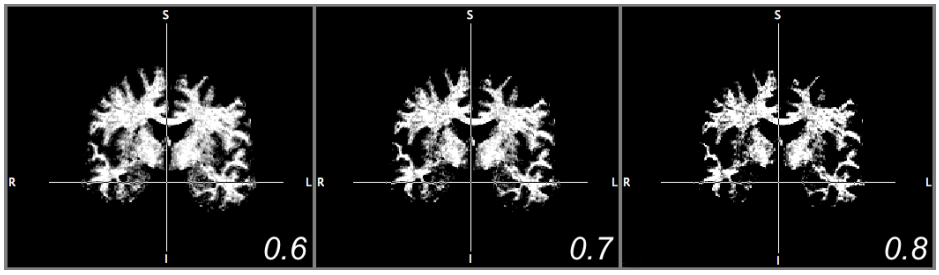

Fig. 2. Coronal view of a single brain depicting the tissue that is selected as θ increases from 0.6 to 0.8. Here, for illustrative purposes, the fraction θ is over the 95^{th} percentile value, whereas in reality we work with a fraction θ of the absolute maximum. The main disadvantage of the latter approach is that it is sensitive to noise, but one of the advantages is that the size of the brain that is left after thresholding becomes computationally manageable.

vertex and edge failures and to betweenness in networks, which can help with identifying communities [8]. The eigenvector corresponding to algebraic connectivity, but also to the second smallest normalised Laplacian eigenvalue is used for spectral clustering [10].

Now, if A_θ is split into m disjoint parts, this will correspond to N_θ consisting of m connected components, which in turn corresponds to m eigenvalues equal to zero in the normalised Laplacian spectrum of N_θ. The eigenvalues close to zero (around the second smallest normalised Laplacian eigenvalue) give us an indication of the fragility of A_θ. The larger the number of eigenvalues that are close to zero, the more fragile (i.e. sensitive to breaking apart) A_θ is. Hence, given a threshold, θ, we denote by f_θ the number of eigenvalues that are close to zero in that particular N_θ, and call this *fragility*. Here, by 'close to zero' we mean those eigenvalues that are less than 0.001, a number which we determined experimentally.

Additionally, we have to compute the surface area, S_θ, and the volume, V_θ; since the values of n_1, n_2 and n_3 (defined at the beginning of this section) can be different (for example, between the three centres EMC, UP and VUMC in the training and test sets) we assumed in our computations that the edges of a single voxel are equal to $\frac{1}{n_1}$, $\frac{1}{n_2}$ and $\frac{1}{n_3}$, respectively. From here one can compute the area of each face of a voxel, as well as its volume (the latter is equal to $\frac{1}{n_1 n_2 n_3}$). Once re-scaled the surface area and volume of A_θ can be computed.

We calibrated the algorithm against the CADDementia training set by combining S_θ (surface area), V_θ (volume) and f_θ (fragility), with the age of the subject and used these four features (numbers) as *predictors* for the stage of neural degeneration (CN, MCI or AD). We firstly used gender to split the subjects apart into two groups.

For illustration purposes, let us consider the group of 13 females (out of 30 subjects) from the training set. For a fixed threshold, θ, we use multinomial logistic regression, which is discussed in detail in [5], [9] and [3]. Specifically, MATLAB is used to compute the multinomial logistic regression (the function `mnrfit`) with *predictors* $X_\theta = [\text{age}, S_\theta, V_\theta, f_\theta]$ and the *responses*, Y, are the labels for the diagnoses of the subjects (0 = CN (control); 1 = MCI; and 2 = AD). As an output of `mnrfit` we get a matrix of coefficient estimates, B_θ. This derives B_θ, we then remove the labels, Y (diagnoses); using only B_θ and (the same) predictors, X_θ, we compute the probabilities for each subject being diagnosed with CN, MCI or AD. The latter probabilities are computed using the MATLAB function `mnrval`. The output for $\theta = 0.66$ is given in Table 1 (columns 4, 5 and 6). Amongst the probabilities p_{CN}, p_{MCI} and p_{AD} we choose the highest, and this determines the class to which a subject is assigned (the third column of Table 1).

For each θ, as in Table 1, we compare our predictions with the set of diagnoses and choose the values of θ for which we get best agreement. In all the tests we performed on the CADDementia training set as well as with a large subset of the ADNI dataset, we consistently found, across both male and female groups, that $\theta = 0.63, 0.64, 0.65, 0.66$ and $\theta = 0.71, 0.72$ gave the best fit between the diagnoses and our predictions. However, in our submission for the CADDementia challenge we chose $\theta = 0.71$ for females and $\theta = 0.63$ for males because these were the most stable values we were getting, that is, small changes in θ did not lead to significant changes in our predictions.

Table 1. The output from the MATLAB function `mnrval` applied to the group of female subjects on the training CAD dataset, $\theta = 0.66$. The second column, that is, the diagnosis of each subject, is only given as a reference here. The input parameters to the the function `mnrval` are only the matrix B_θ and the (matrix of) predictors, X_θ. The function `mnrval` outputs the probabilities p_{CN}, p_{MCI} and p_{AD}.

subject ID	diagn.	predict.	p_{CN}	p_{MCI}	p_{AD}
train_emc_002	2	1	0	0.78	0.22
train_emc_003	0	0	0.99	0.005	0.003
train_emc_008	0	0	0.89	0.0008	0.1
train_emc_009	2	2	0	0	1
train_emc_011	1	1	0	0.87	0.13
train_up_001	2	2	0	0.004	0.995
train_up_002	1	1	0	0.64	0.36
train_vumc_004	2	2	0	0.03	0.97
train_vumc_005	0	0	0.98	0.01	0.01
train_vumc_008	2	2	0	0	1
train_vumc_010	1	1	0.004	0.92	0.08
train_vumc_012	1	1	0.05	0.74	0.21
train_vumc_013	2	2	0.1	0.001	0.90

Given our value for θ we compute the matrix B_θ on female subjects in the CADDementia training set. Further, we can find the corresponding predictors, $[\text{age}, S_\theta, V_\theta, f_\theta]$, for the female subjects from the CADDementia test data set. Therefore, we can use B_θ with those new predictors as input parameters to the function `mnrval` and derive the corresponding probabilities, p_{CN}, p_{MCI} and p_{AD}.

3 Materials

The algorithm described in §2 was tested against the CADDementia training set as well as suitable data from the ADNI database as described next.

3.1 CADDementia Data

The CADDementia dataset[9] comprises of 384 T1-weighted 3T MRI scans in gzipped Nifti format of subjects with AD, MCI and healthy controls, that were captured from multiple different centres. The original data (6.2 Gbytes[10]) as well as a non-uniformity corrected version (16 Gbytes[10]) is provided and all data is reported as being clinically-representative. A training subset comprising of 30 scans that are reported to be equally distributed over the originating centres, as well as the corresponding diagnostic labels are provided. The demographic metadata comprises of age and gender. Note that we used the non-uniformity-corrected version of the dataset.

[9] http://grand-challenge.org/site/caddementia/download_data

[10] `gzip` compressed.

3.2 ADNI Data

The Alzheimers Disease Neuroimaging Initiative (ADNI) database[11] was launched in 2003 by the National Institute on Ageing (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organisations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimers disease (AD). The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date, the ADNI Website reports that these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD.

Only T1-weighted structural 3T MRI data was selected for use in our investigation as contained in the following ADNI collections:

- `AD-{bl,m06,m12,m24}-3.0T`
- `ADNI1:{Baseline,Annual 2Yr,Complete{1,2,3}Yr}`
- `MCI-{bl,m06,m12,m18,m24,m36}-3.0T`
- `Normal-m{06,12,24,36}-3.0T`

resulting in multiple scans being retrieved for each of 189 subjects (33 Gbytes non-compressed), with age ranging from 58 to 93 years old.

4 Compute platform

The computers used for this study (see Table 2) comprised of workstations from our Analysis Laboratory (AL) as well as our small Infiniband-connected compute cluster (IB), all of which combine to report a total of 120 Intel CPU cores, to a HTCondor[12] batch queue. The Gentoo Linux[13] operating system, with kernel `3.14.4-gentoo x86_64` was used across all machines. The notation '`S-R`' in the *Disk* column indicates that the system disk is SSD and the scratch disk is rotational. The *HT* column indicates if hyper-threading[14] was enabled.

5 Workflow and execution times

Conceptually the algorithm and all processing presented in this paper is entirely automated, however in terms of our current workflow there are presently a number of steps between the stages, that we invoke manually (e.g. instruct a script to execute).

[11] http://adni.loni.usc.edu
[12] HTCondor version 7.8.8, see http://research.cs.wisc.edu/htcondor
[13] http://www.gentoo.org
[14] http://en.wikipedia.org/wiki/Hyper-threading

Table 2. The computational resources used during this study

#	Grp	Nodes(s)	Qty	Model	GHz	# CPU	# Cores	HT	Mem (GB)	GPGPU	Disk	Network
1	AL	ncpc{50-62}	13	i7-4770	3.40	1	8	✓	32	C2050	S-S	GigE
2	AL	ncpc5	1	i7 920	2.67	1	4	×	24	C2050	S-S	GigE
3	IB	ncpc139	1	X3460	2.80	1	4	×	16	-	S-R	QDR IB
4	IB	ncpc14{1,2}	2	i7 950	3.07	1	4	×	24	-	S-R	QDR IB
6	IB	ncpc14{5-9}	5	X5570	2.93	2	8	×	24	-	S-R	QDR IB
7	IB	ncpc150	1	X5690	3.46	2	12	×	48	3×C2050	S-S	QDR IB
10	SRV	ncsrv{1,2,3,4}	4	-	-	-	-	-	-	-	R-R	QDR IB

The workflow is described thus:

1. **Compile Matlab code to a standalone binary** - The Matlab code described in §2 is compiled to a standalone binary using the Matlab Compiler in preparation for license-free parallel execution[15] across the compute nodes.
2. **Download the CADDementia and ADNI Data** - Given the CADDementia data consisted of relatively few files, we elected to download manually, although this could potentially have been scripted. We obtained the ADNI data via[16].
3. **Unpack data** - This operation is scripted and can either be performed sequentially from a single machine or in parallel using, e.g. our HTCondor installation.
4. **Brain extraction** - Brains are extracted from all the scans using the following FMRIB FSL[17] `roi` and `bet` commands in a data parallel way on our cluster: `standard_space_roi <in_file> <intermediate_file> -b; bet <intermediate_file> <output_file> -f 0.15`
5. **Process the data** - The compiled Matlab code is executed in a data parallel way (each data file is submitted as an independent task to the batch queue) via our computational resources. Results are written back to a single location on our network file system.
6. **Final classification** - The results from the previous step (together with some of the demographics) are used as *predictors* (discussed in §2) in order to do the final classification of the subjects (into the following classes: CN, MCI and AD). The classification is done in MATLAB, using the functions `mnrfit` and `mnrval`.

The preprocessing task in workflow step #4 (brain extraction) takes less than 90 seconds for a single brain scan, and ≈ 15 minutes for all 354 CADDementia test brains to be extracted in parallel across the cluster (sequentially this same operation could take up to ≈ 8 hours to complete on a single computer).

[15] http://www.mathworks.co.uk/products/compiler/mcr
[16] https://ida.loni.usc.edu/login.jsp
[17] FMRIB Software Library, http://fsl.fmrib.ox.ac.uk/fsl/fslwiki

The length of the computation (workflow step #5) of a single MRI scan depends, mostly, on the amount of voxels that were left in the set A_θ (defined in §2) and also, in the computation of the spectrum of the normalised Laplacian matrix, on how fragile or connected A_θ is as a 3D structure. Therefore, MRI scans in which the set A_θ was rather large took longer to compute (the length being dependant on the complexity of the MATLAB eigenvalue solver, `eigs`). For example, workflow step #5 takes anywhere between 6 and 24 minutes for the majority of the brain scans processed. However, twenty six of the 354 CADDementia test scans took considerably longer with, for example, `stripped_test_vumc_116.nii`, requiring ≈ 60 hours of actual compute time and a memory footprint of 4.1Gbytes. Moreover, despite having an otherwise empty batch queue there were more data items to be processed from the CADDementia test dataset than the available processors (exposed by the batch queue) across our local compute resources, hence the overall end-to-end wall clock time to getting a result for the scan was in fact $87\frac{1}{2}$ hours, after taking into account the time the job spent languishing in the queue. The final classification (workflow step #6) takes less than a minute on a single machine.

The `condor_submit` files for workflow steps #3,4,5 are automatically generated by a BASH[18] script that dynamically identifies compute tasks based on the previously downloaded data.

6 Conclusions

The results that we observed while testing our method with the CADDementia training set (consistently, less than 20% incorrect predictions) and the ADNI dataset (consistently, less than 35% incorrect predictions) appear promising. The approach we have presented here is intuitive and easy to implement. We believe it is a potential step towards employing Network Theory in the analysis and classification of neural diseases, and as such it can be extended to include more sophisticated techniques from Network Theory.

This technique is agnostic to underlying tissue properties as well as to the nature of the signal. We have previously applied a similar approach to resting state fMRI data [6]. For the purposes of this competition we have intentionally biased the algorithm in favour of white matter by stepping up the threshold values. Alternatively, we could use a step down method to capture properties of grey matter. It is also possible to target specific tissues using a gating procedure.

A further advantage, which we consider important, is that the workflow can be fully automated and potentially, this can be implemented in such a way that, for example, a computer Web Service running in the cloud would ingest MRI data and return a diagnosis within a short time over the network; issues of trust, privacy, authorisation and security would need to be looked at to ensure compliance with the legislation for medical data protection.

The work we have described takes a data parallel approach to processing many scans at the same time, and for the most part, results are obtained within

[18] http://en.wikipedia.org/wiki/Bash_%28Unix_shell%29

a matter of minutes. However as we have seen, MRI scans in which the set A_θ is rather large can take substantial amounts of time to compute, and thus are candidates for a finer granularity of parallelism, which we shall investigate in future work. In addition to increasing the speed of the diagnosis we can also improve accuracy by applying our approach to the parts of the brain whose structural changes are known to be highly correlated with the presence or absence of AD and/or MCI.

7 Acknowledgements

We thank Tom Johnstone and Shan Shen from CINN, for their helpful comments and suggestions. We gratefully acknowledge the support of the University of Reading Research Endowment Trust Fund (research fellowship for GS). The computational aspects of this work were based on a foundation previously laid down during the EPSRC NeuroCloud project under grant EP/I016856/1. Data collection and sharing for this project was partly funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award W81XWH-12-2-0012).

References

1. Anirban Banerjee and Joergen Jost. Graph spectra as a systematic tool in computational biology. *Discrete Applied Mathematics*, 157(10):2425 – 2431, 2009. Networks in Computational Biology.
2. Fan R. K. Chung. *Spectral Graph Theory*. CBMS Regional Conference Series in Mathematics. AMS, 1997.
3. A. J. Dobson and A. G. Barnett. *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC. Taylor and Francis Group, 2008.
4. M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23:298–305, 1973.
5. P. Grindrod. *Mathematical Underpinnings of Analytics*. Oxford University Press, 2014.
6. Peter Grindrod, Zhivko V Stoyanov, Garry M Smith, and J Doug Saddy. Primary evolving networks and the comparative analysis of robust and fragile structures. *Journal of Complex Networks*, September 2013.
7. A. Jamakovic and S. Uhlig. On the relationship between the algebraic connectivity and graph's robustness to node and link failures. In *Next Generation Internet Networks, 3rd EuroNGI Conference on*, pages 96–102, May 2007.
8. Javier Martn Hernndez, Zongwei Li, and Piet Van Mieghem. Weighted betweenness and algebraic connectivity. *Journal of Complex Networks*, 2014.
9. P. McCullagh and J. A. Nelder. *Generalized Linear Models*. New York: Chapman and Hall, 1990.
10. J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE. Trans. on Pattern Analysis and Machine Intelligence*, 22:888–905, 2000.